Probability and Statistics with Examples using R

Siva Athreya, Deepayan Sarkar, and Steve Tanner

November 19, 2024

Version: – November 19, 2024

CONTENTS

PREFACE

We believe that many foundational ideas of Probability and Statistics are best understood when their natural connection is emphasised. We feel that the interested student should learn the mathematical rigour of Probability, the motivating examples and techniques from Statistics, and an instructive technology to perform computations relating to both in an inclusive manner. These formed our main motivations for writing this book. We have chosen to use the R software environment to demonstrate an available computational tool.

The book is intended to be an undergraduate text for a course on Probability Theory. We had in mind courses such as the one year (two semester) Probability course at many universities in India such as the Indian Statistical Institute or Chennai Mathematical Institue, or a one semester (or two quarter) Probability course as is commonly offered as an upper division, post-calculus elective at many North American universities. The Statistics material and the package R are introduced so as to emphasise motivations and applications of the probabilistic material. We assume that our readers are well-versed in calculus, have a basic understanding of the theory of sets and functions, combinatorics, and proof techniques, and have at least a passing awareness of the distinction between countable and uncountable infinities. We do not assume any particular experience of Linear Algebra or Real Analysis.

In Chapter 1 of this book we begin with an introduction to Outcomes, Sample Space, Events and the axiomatic definition of Probability. Then we discuss the concepts of conditional probability, independence and Bayes' Theorem. We conclude this chapter with a basic introduction to R. R is a Free Open Source software environment that runs on all major software platforms, and instructions to download and install it are available at https://www.r-project.org/.

We begin Chapter 2 by applying the notion of independence to repeated trials (Bernoulli Trials) and discuss the Binomial and Geometric distributions. We introduce the Poisson distribution as a limiting approximation of the Binomial. We conclude this section with a discussion on Sampling with and without replacement. The Hypergeometric Distribution is thus introduced here and we prove its approximation to the Binomial. Throughout this chapter and later in the book we provide the R code for calculating the probabilities associated with common distributions.

In Chapters 3 and 4 we introduce discrete random variables (functions on a sample space whose range is countable) and related concepts. In Chapter 3, we define the probability

mass function, distribution function, and independence for random variables. We introduce the Multinomial distribution and show the memoryless property of the Geometric random variable. The chapter concludes by providing a method to compute the distributions of functions of one and several random variables, defining the concept of joint distribution along the way. In Chapter 4, we define Expectation, Variance, Covariance, Conditional Expectation and Conditional Variance for discrete random variables. Results involving these quantities for standard distributions are presented (and proved) as well. We also state and prove the Markov and Chebyshev inequalities along with the notion of standardising random variables to mean zero and variance one.

Working with uncountable spaces and understanding the probability density function of an absolutely continuous random variable are challenging without assuming a background in Real Analysis but we make a modest attempt towards this in Chapter 5. We begin with a description of uncountable sample spaces. After having described events in a temporary manner in Chapter 1 we provide a precise definition here but comfort the reader that we shall avoid the most general events and at most consider countable union/intersection of intervals. This allows us to be fairly rigorous with random variables having piecewise continuous probability density functions using results from basic calculus. After this we imitate the program conducted in Chapter 3. Standard distributions such as Uniform, Exponential, and Normal are discussed. While computing densities of sums and ratios of independent random variables we introduce the Gamma distribution and use it to derive the Beta distribution as an example of ratio of dependent Gamma random variables.

In Chapter 6 we define Variance, Covariance, Conditional Expectation and Conditional Variance for continuous random variables and summarise their properties. Moment generating functions for random variables are defined. At this point, to respect the minimal background assumption on our reader we state a few important results without proof such as the fact that the moment generating functions characterise distribution of a random variable. The chapter ends with a section on Bivariate Normal random variables. Here we have done all computations in this section without using Linear Algebra but the notational efficiency of using Linear Algebra is explained via exercises for the interested reader.

With the foundational ideas of Probability laid out we proceed in Chapter 7 with Sampling and Descriptive Statistics. The empirical distribution, the sample mean, variance and proportion are defined along with their properties. Simulation is used to develop intuition regarding sampling variability, and plots such as Histograms, Hanging Rootograms, and Q-Q Plots are introduced and illustrated using R.

Limit Theorems for Sampling Distributions discussed in Chapter 7 are the objective of Chapter 8. We begin with a brief description of multivariate joint densities and Order statistics. The t-distribution and χ^2 (chi-square) distributions are introduced in this chapter.

The sample mean and variance from a normal population are discussed in relation to t and χ^2. We prove the Weak Law of Large numbers and the Central Limit Theorem for random variables possessing a moment generating function. We do state a more general version of the Central Limit Theorem and also the Strong Law of Large numbers, providing a proof of the latter in the Appendix. Along with R code we discuss the continuity correction and applications of the Central Limit Theorem via examples. We then discuss the delta method and its application to variance stabilising transformations. The chapter concludes with a derivation of the Central Limit Theorem for the median.

We end the book with two chapters focused solely on results and techniques from statistics. In Chapter 9 we discuss Estimation and Confidence Intervals. We briefly describe Method of Moments Estimators and Maximum Likelihood Estimators. We then introduce Confidence Intervals using the Pivotal Quantity approach. For cases when natural pivotal quantities do not exist, we illustrate the use of the Central Limit Theorem to obtain approximate confidence intervals. Finally we derive confidence intervals based on the sample median and compare its performance with intervals based on the sample mean via simulations.

In Chapter 10 we explore a non-traditional approach to Hypothesis Testing based on p-values rather than pre-determined significance levels. We first formulate the multinomial goodness of fit and independence problems in the familiar parametric set up. After that we describe an intuitive approach to derive suitable test statistics for various hypothesis testing examples. We then proceed to outline a likelihood ratio based approach to derive test statistics systematically. We conclude this chapter with a discussion of the goodness of fit and independence tests.

R code for most of the computations done are given in the book itself, and the reader should be able to reproduce and extend them easily. Code for figures are not given in the book, but are available at a website accompanying the book.

The Appendix includes some relevant mathematical details not covered in the main matter of the book. The topics included are the Jacobian method for computing distribution of transformations of random variables and the Strong Law of Large Numbers.

Siva Athreya, *Deepayan Sarkar*, and *Steve Tanner*
November 19, 2024

1

BASIC CONCEPTS

1.1 DEFINITIONS AND PROPERTIES

Most of the problems in probability and statistics involve determining how likely it is that certain things will occur. Before we can talk about what is likely or unlikely, we need to know what is possible. In other words, we need some framework in which to discuss what sorts of things have the potential to occur. To that end, we begin by introducing the basic concepts of "sample space", "experiment", "outcome", and "event". We also define what we mean by a "probability" and provide some examples to demonstrate the consequences of the definition.

1.1.1 *Definitions*

Definition 1.1.1. (Sample Space) *A sample space S is a set. The elements of the set S will be called "outcomes" and should be viewed as a listing of all possibilities that might occur. We will call the process of actually selecting one of these outcomes an "experiment".*

For its familiarity and simplicity we will frequently use the example of rolling a die. In that case our sample space would be $S = \{1, 2, 3, 4, 5, 6\}$, a complete listing of all of the outcomes on the die. Performing an experiment in this case would mean rolling the die and recording the number that it shows. However, sample space outcomes need not be numeric. If we are flipping a coin (another simple and familiar example) experiments would result in one of two outcomes and the appropriate sample space would be $S = \{Heads, Tails\}$.

For a more interesting example, if we are discussing which country will win the next World Cup, outcomes might include Brazil, Spain, Canada, and Thailand. Here the set S might be all the world's countries. An experiment in this case requires waiting for the next World Cup and identifying the country which wins the championship game. Though we have not yet explained how probability relates to a sample space, soccer fans amongst our readers may regard this example as a foreshadowing that not all outcomes of a sample space will necessarily have the same associated probabilities.

Definition 1.1.2. (Temporary Definition of Event) *Given a sample space S, an "event" is any subset $E \subset S$.*

This definition will allow us to talk about how likely it is that a range of possible outcomes might occur. Continuing our examples above we might want to talk about the probability that a die rolls a number larger than two. This would involve the event $\{3, 4, 5, 6\}$ as a subset of $\{1, 2, 3, 4, 5, 6\}$. In the soccer example we might ask whether the World Cup will be won by a South American country. This subset of our list of all the world's nations would contain Brazil as an element, but not Spain.

It is worth noting that the definition of "event" includes both S, the sample space itself, and $\varnothing$, the empty set, as legitimate examples. As we introduce more complicated examples we will see that it is not always necessary (or even possible) to regard every single subset of a sample space as a legitimate event, but since the reasons for that may be distracting at this point we will use the above as a temporary definition of "event" and refine the definition when it becomes necessary.

To each event, we want to assign a chance (or "probability") which will be a number between 0 and 1. So if the probability of an event E is 0.72, we interpret that as saying, "When an experiment is performed, it has a 72% chance of resulting in an outcome contained in the event E". Probabilities will satisfy two axioms stated and explained below. This formal definition is due to Andrey Kolmogorov (1903-1987).

Definition 1.1.3. (Probability Space Axioms) *Let S be a sample space and let $\mathcal{F}$ be the collection of all events.*
*A "**probability**" is a function $P : \mathcal{F} \to [0, 1]$ such that*

(1) $P(S) = 1$; and

(2) If $E_1, E_2, \ldots$ are a countable collection of disjoint events (that is, $E_i \cap E_j = \varnothing$ if $i \neq j$), then

$$P(\bigcup_{j=1}^{\infty} E_j) = \sum_{j=1}^{\infty} P(E_j). \tag{1.1.1}$$

The first axiom is relatively straight forward. It simply reiterates that S did, indeed, include all possibilities, and therefore there is a 100% chance that an experiment will result in some outcome included in S. The second axiom is not as complicated as it looks. It simply says that probabilities add when combining a countable number of disjoint events. It is implicit that the series on right hand side of the equation (1.1.1) converges. Further

(1.1.1) also holds when combining finite number of disjoint events (see Theorem 1.1.4 below).

Returning to our World Cup example, suppose A is a list of all North American countries and E is a list of all European countries. If it happens that $P(A) = 0.05$ and $P(E) = 0.57$ then $P(A \cup E) = 0.62$. In other words, if there is a 5% chance the next World Cup will be won by a North American nation and a 57% chance that it will be won by a European nation, then there is a 62% chance that it will be won by a nation from either Europe or North America. The disjointness of these events is obvious as (if we discount island territories) there is no country that is in both North America and Europe.

The requirement of axiom two that the collection of events be countable is important. We shall see shortly that, as a consequence of axiom two, disjoint additivity also applies to any finite collection of events. It does not apply to uncountably infinite collections of events, though that fact will not be relevant until later in the text when we discuss continuous probability spaces.

1.1.2 *Basic Properties*

There are some immediate consequences of these probability axioms which we will state and prove before returning to some simple examples.

Theorem 1.1.4. *Let P be a probability on a sample space S. Then,*

(1) $P(\varnothing) = 0$;

(2) If $E_1, E_2, \ldots E_n$ are a finite collection of disjoint events, then

$$P\left(\bigcup_{j=1}^{n} E_j\right) = \sum_{j=1}^{n} P(E_j);$$

(3) If E and F are events with $E \subset F$, then $P(E) \leq P(F)$;

(4) If E and F are events with $E \subset F$, then $P(F \setminus E) = P(F) - P(E)$;

(5) Let E^c be the complement of event E. Then $P(E^c) = 1 - P(E)$; and

(6) If E and F are events then $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

Proof. (1) - The empty set is disjoint from itself, so $\varnothing, \varnothing, \ldots$ is a countable disjoint collection of events. From the second axiom, $P\left(\bigcup_{j=1}^{\infty} E_j\right) = \sum_{j=1}^{\infty} P(E_j)$. When this is

applied to the collection of empty sets we have $P(\varnothing) = \sum_{j=1}^{\infty} P(\varnothing)$. If $P(\varnothing)$ had any non-zero value, the right hand side of this equation would be a divergent series while the left hand side would be a number. Therefore, $P(\varnothing) = 0$.

Proof of (2) - To use axiom two we need to make this a countable collection of events. We may do so while preserving disjointness by including copies of the empty set. Define $E_j = \varnothing$ for $j > n$. Then $E_1, E_2, \ldots, E_n, \varnothing, \varnothing, \ldots$ is a countable collection of disjoint sets and therefore $P\left(\bigcup_{j=1}^{\infty} E_j\right) = \sum_{j=1}^{\infty} P(E_j)$. However, the empty sets add nothing to the union and so $\bigcup_{j=1}^{\infty} E_j = \bigcup_{j=1}^{n} E_j$. Likewise since we have shown $P(\varnothing) = 0$ these sets also add nothing to the sum, so $\sum_{j=1}^{\infty} P(E_j) = \sum_{j=1}^{n} P(E_j)$.

Combining these gives the result:

$$P\left(\bigcup_{j=1}^{n} E_j\right) = P\left(\bigcup_{j=1}^{\infty} E_j\right) = \sum_{j=1}^{\infty} P(E_j) = \sum_{j=1}^{n} P(E_j).$$

Proof of (3) - If $E \subset F$, then E and $F \setminus E$ are disjoint events with a union equal to F. Using (2) above gives $P(F) = P(E \cup (F \setminus E)) = P(E) + P(F \setminus E)$.
Since probabilities are assumed to be positive, it follows that $P(F) \geq P(E)$.

Proof of (4) - As with the proof of (3) above, E and $F \setminus E$ are disjoint events with $E \cup (F \setminus E) = F$. Therefore $P(F) = P(E) + P(F \setminus E)$ from which we get the result.

Proof of (5) - This is simple a special case of (4) where $F = S$.

Proof of (6) - We may disassemble $E \cup F$ disjointly as $E \cup F = E \cup (F \setminus E)$. Then from (2) we have $P(E \cup F) = P(E) + P(F \setminus E)$.
Next, since $F \setminus E \subset F$ and since $F \setminus (F \setminus E) = E \cap F$ we can use (4) to write $P(E \cup F) = P(E) + P(F) - P(E \cap F)$. ■

Example 1.1.5. A coin flip can come up either "heads" or "tails", so $S = \{heads, tails\}$. A coin is considered "fair" if each of these outcomes is equally likely. Which axioms or properties above can be used to reach the (obvious) conclusion that both outcomes have a 50% chance of occurring?

Each outcome can also be regarded as an event. So $E = \{heads\}$ and $F = \{tails\}$ are two disjoint events. If the coin is fair, each of these events is equally likely, so $P(E) = P(F) = p$ for some value of p. However, using the second axiom, $1 = P(S) = P(E \cup F) = P(E) + P(F) = 2p$. Therefore, $p = 0.5$, or in other words each of the two possibilities has a 50% chance of occurring on any flip of a fair coin. ■

In the examples above we have explicitly described the sample space S, but in many cases this is neither necessary nor desirable. We may still use the probability space axioms and their consequences when we know the probabilities of certain events even if the sample space is not explicitly described.

Example 1.1.6. A certain sea-side town has a small fishing industry. The quantity of fish caught by the town in a given year is variable, but we know there is a 35% chance that the town's fleet will catch over 400 tons of fish, but only a 10% chance that they will catch over 500 tons of fish. How likely is it they will catch between 400 and 500 tons of fish?

The answer to this may be obvious without resorting to sets, but we use it as a first example to illustrate the proper use of events. Note, though, that we will not explicitly describe the sample space S.

There are two relevant events described in the problem above. We have F representing "the town's fleet will catch over 400 tons of fish" and E representing "the town's fleet will catch over 500 tons of fish". We are given that $P(E) = 0.1$ while $P(F) = 0.35$.

Of course $E \subset F$ since if over 500 tons of fish are caught, the actual tonnage will be over 400 as well. The event that the town's fleet will catch between 400 and 500 tons of fish is $F \setminus E$ since E hasn't occurred, but F has. So using property (4) from above we have $P(F \setminus E) = P(F) - P(E) = 0.35 - 0.1 = 0.25$. In other words there is a 25% chance that between 400 and 500 tons of fish will be caught. ■

Example 1.1.7. Suppose we know there is a 60% chance that it will rain tomorrow and a 70% chance the high temperature will be above $30°C$. Suppose we also know that there is a 40% chance that the high temperature will be above $30°C$ and it will rain. How likely is it tomorrow will be a dry day that does not go above $30°C$?

The answer to this question may not be so obvious, but our first step is still to view the pieces of information in terms of events and probabilities. We have one event E which represents "It will rain tomorrow" and another F which represents "The high will be above $30°C$ tomorrow". Our given probabilities tell us $P(E) = 0.6$, $P(F) = 0.7$, and $P(E \cap F) = 0.4$. We are trying to determine $P(E^c \cap F^c)$. We can do so using properties (5) and (6) proven above, together with the set-algebraic fact that $E^c \cap F^c = (E \cup F)^c$.

From (5) we know $P(E \cup F) = P(E) + P(F) - P(E \cap F) = 0.7 + 0.6 - 0.4 = 0.9$. (This is the probability that it either will rain or be above 30 ^0C).

Then from (6) and the set-algebraic fact, $P(E^c \cap F^c) = P((E \cup F)^c) = 1 - P(E \cup F) = 1 - 0.9 = 0.1$. So there is a 10% chance tomorrow will be a dry day that does not reach $30°C$. ■

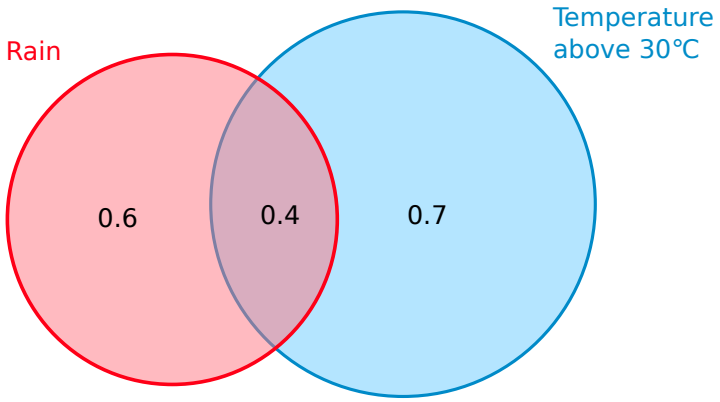


Figure 1.1: A Venn diagram that describes the probabilities from Example 1.1.7.

EXERCISES

Ex. 1.1.1. Consider the sample space $\Omega = \{a, b, c, d, e\}$. Given that $\{a, b, e\}$, and $\{b, c\}$ are both **events**, what other subsets of Ω must be events due to the requirement that the collection of events is closed under taking unions, intersections, and compliments?

Ex. 1.1.2. There are two positions - Cashier and Waiter - open at the local restaurant. There are two male applicants (David and Rajesh) two female applicants (Veronica and Megha). The Cashier position is chosen by selecting one of the four applicants at random. The Waiter position is then chosen by selecting at random one of the three remaining applicants.

(a) List the elements of the sample space S.

(b) List the elements of the event A that the position of Cashier is filled by a female applicant.

(c) List the elements of the event B that exactly one of the two positions is filled by a female applicant.

(d) List the elements of the event C that neither position was filled by a female applicant.

(e) Sketch a Venn diagram to show the relationship among the events A, B, C and S.

Ex. 1.1.3. A jar contains a large collection of red, green, and white marbles. Marbles are drawn from the jar one at a time. The color of the marble is recorded and it is put back in

the jar before the next draw. Let R_n denote the event that the n-th draw is a red marble and let G_n denote the event that the n-th draw is a green marble. For example, $R_1 \cap G_2$ would denote the event that the first marble was red and the second was green. In terms of these events (and appropriate set-theoretic symbols – union, intersection, and complement) find expressions for the events in parts (a), (b), and (c) below.

(a) The first marble drawn is white. (We might call this W_1, but show that it can be written in terms of the R_n and G_n sets described above).

(b) The first marble drawn is green and the second marble drawn is not white.

(c) The first and second draws are different colors.

(d) Let $E = R_1 \cup G_2$ and let $F = R_1^c \cap R_2$. Are E and F disjoint? Why or why not?

Ex. 1.1.4. Suppose there are only thirteen teams with a non-zero chance of winning the next World Cup. Suppose those teams are Spain (with a 14% chance), the Netherlands (with a 11% chance), Germany (with a 11% chance), Italy (with a 10% chance), Brazil (with a 10% chance), England (with a 9% chance), Argentina (with a 9% chance), Russia (with a 7% chance), France (with an 6% chance), Turkey (with a 4% chance), Paraguay (with a 4% chance), Croatia (with a 4% chance) and Portugal (with a 1% chance).

(a) What is the probability that the next World Cup will be won by a South American country?

(b) What is the probability that the next World Cup will be won by a country that is not from South America? (Think of two ways to do this problem - one directly and one using part (5) of Theorem 1.1.4. Which do you prefer and why?)

Ex. 1.1.5. If A and B are disjoint events and $P(A) = 0.3$ and $P(B) = 0.6$, find $P(A \cup B)$, $P(A^c)$ and $P(A^c \cap B)$.

Ex. 1.1.6. Suppose E and F are events in a sample space S. Suppose that $P(E) = 0.7$ and $P(F) = 0.5$.

(a) What is the largest possible value of $P(E \cap F)$? Explain.

(b) What is the smallest possible value of $P(E \cap F)$? Explain.

Ex. 1.1.7. A biologist is modeling the size of a frog population in a series of ponds. She is concerned with both the number of egg masses laid by the frogs during breeding season and the annual precipitation into the ponds. She knows that in a given year there is an 86% chance that there will be over 150 egg masses deposited by the frogs (event E) and that there is a 64% chance that the annual precipitation will be over 17 inches (event F).

(a) In terms of E and F, what is the event "there will be over 150 egg masses and an annual precipitation of over 17 inches"?

(b) In terms of E and F, what is the event "there will be 150 or fewer egg masses and the annual precipitation will be over 17 inches"?

(c) Suppose the probability of the event from (a) is 59%. What is the probability of the event from (b)?

Ex. 1.1.8. In part (6) of Theorem 1.1.4 we showed that

$$P(E \cup F) = P(E) + P(F) - P(E \cap F).$$

Versions of this rule for three or more sets are explored below.

(a) Prove that $P(A \cup B \cup C)$ is equal to

$$P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

for any events A, B, and C.

(b) Use part (a) to answer the following question. Suppose that in a certain United States city 49.3% of the population is male, 11.6% of the population is sixty-five years of age or older, and 13.8% of the population is Hispanic. Further, suppose 5.1% is both male and at least sixty-five, 1.8% is both male and Hispanic, and 5.9% is Hispanic and at least sixty-five. Finally, suppose that 0.7% of the population consists of Hispanic men that are at least sixty-five years old. What percentage of people in this city consists of non-Hispanic women younger than sixty-five years old?

(c) Find a four-set version of the equation. That is, write $P(A \cup B \cup C \cup D)$ in terms of probabilities of intersections of A, B, C, and D.

(d) Find an n-set version of the equation.

Ex. 1.1.9. A and B are two events. P(A)=0.4, P(B)=0.3, P(A∪B)=0.6. Find the following probabilities:

(a) $P(A \cap B)$;

(b) P(Only A happens); and

(c) P(Exactly one of A or B happens).

Ex. 1.1.10. In the next subsection we begin to look at probability spaces where each of the outcomes are equally likely. This problem will help develop some early intuition for such problems.

(a) Suppose we roll a die and so $S = \{1, 2, 3, 4, 5, 6\}$. Each outcome separately $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}$ is an event. Suppose each of these events is equally likely. What must the probability of each event be? What axioms or properties are you using to come to your conclusion?

(b) With the same assumptions as in part (a), how would you determine the probability of an event like $E = \{1, 3, 4, 6\}$? What axioms or properties are you using to come to your conclusion?

(c) If $S = \{1, 2, 3, ..., n\}$ and each single-outcome event is equally likely, what would be the probability of each of these events?

(d) Suppose $E \subset S$ is an event in the sample space from part (c). Explain how you could determine $P(E)$.

Ex. 1.1.11. Suppose A and B are subsets of a sample space Ω.

(a) Show that $(A - B) \cup B = A$ when $B \subset A$.

(b) Show by example that the equality doesn't always hold if B is not a subset of A.

Ex. 1.1.12. Let A and B be events.

(a) Suppose $P(A) = P(B) = 0$. Prove that $P(A \cup B) = 0$.

(b) Suppose $P(A) = P(B) = 1$. Prove that $P(A \cap B) = 1$.

Ex. 1.1.13. Let A_n be a sequence of events.

(a) Suppose $A_n \subseteq A_{n+1}$ for all $n \geq 1$. Show that

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} P(A_n)$$

(b) Suppose $A_n \supseteq A_{n+1}$ for all $n \geq 1$. Show that

$$P\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} P(A_n)$$

1.2 EQUALLY LIKELY OUTCOMES

When a sample space S consists of only a countable collection of outcomes, describing the probability of each individual outcome is sufficient to describe the probability of all events. This is because if $A \subset S$ we may simply compute

$$P(A) = P\Big(\bigcup_{\omega \in A} \{\omega\}\Big) = \sum_{\omega \in A} P(\{\omega\}).$$

This assignment of probabilities to each outcome is called a "distribution" since it describes how probability is distributed amongst the possibilities. Perhaps the simplest example arises when there are a finite collection of equally likely outcomes. Think of examples such as flipping a fair coin ("heads" and "tails" are equally likely to occur), rolling a fair die (1, 2, 3, 4, 5, and 6 are equally likely), or drawing a set of numbers for a lottery (many possibilities, but in a typical lottery, each outcome is as likely as any other). Such distributions are common enough that it is useful to have shorthand notations for them. In the case of a sample space $S = \{\omega_1, \ldots, \omega_n\}$ where each outcome is equally likely, the probability is referred to as a "uniform distribution" and is denoted by $\text{Uniform}(\{\omega_1, \ldots, \omega_n\})$. In such situations, computing probabilities simply reduces to computing the number of outcomes in a given event and consequently becomes a combinatorial problem.

Theorem 1.2.1. Uniform$(\{\omega_1, \ldots, \omega_n\})$: *Let $S = \{\omega_1, \omega_2, \ldots, \omega_n\}$ be a non-empty, finite set. If $E \subset S$ is any subset of S, let $P(E) = \frac{|E|}{|S|}$ (where $|E|$ represents the number of elements of E). Then P defines a probability on S and P assigns equal probability to each individual outcome in S.*

Proof. As $E \subset S$, we know that $|E| \leq |S|$ and so $0 \leq P(E) \leq 1$. So, we must prove that P satisfies the two probability axioms.

The first axiom is satisfied because $P(S) = \frac{|S|}{|S|} = 1$. To verify the second axiom, suppose $E_1, E_2, \ldots$ is a countable collection of disjoint events. As S is finite, only finitely many of these E_j can be non-empty, so we may list the non-empty events as $E_1, E_2, \ldots, E_n$. For $j > n$ we know $E_j = \varnothing$ and so $P(E_j) = 0$ by definition. As the events are disjoint, to find the number of elements in their union we simply add the elements of each event separately. That is, $|E_1 \cup E_2 \cup \cdots \cup E_n| = |E_1| + |E_2| + \cdots + |E_n|$, and so

$$P\left(\bigcup_{j=1}^{\infty} E_j\right) = \frac{\left|\bigcup_{j=1}^{\infty} E_j\right|}{|S|} = \frac{\sum_{j=1}^{n} |E_j|}{|S|} = \sum_{j=1}^{n} \frac{|E_j|}{|S|} = \sum_{j=1}^{n} P(E_j) = \sum_{j=1}^{\infty} P(E_j).$$

Finally, let $\omega \in S$ be any single outcome and let $E = \{\omega\}$. Then $P(E) = \frac{1}{|S|}$, so every outcome in S is equally likely. ■

Example 1.2.2. A deck of twenty cards labeled $1, 2, 3, \ldots, 20$ is shuffled and a card selected at random. What is the probability that the number on the card is a multiple of six?

The description of the scenario suggests that each of the twenty cards is as likely to be chosen as any other. In this case $S = \{1, 2, 3, ..., 20\}$ while $E = \{6, 12, 18\}$. Therefore, $P(E) = \frac{|E|}{|S|} = \frac{3}{20} = 0.15$. There is a 15% chance that the card will be a multiple of six. ■

Example 1.2.3. Two dice are rolled. How likely is it that their sum will equal eight?

Since we are looking at a sum of dice, it might be tempting to regard the sample space as $S = \{2, 3, 4, ..., 11, 12\}$, the collection of possible sums. While this is a possible approach (and one we will return to later), it is not the case that all of these outcomes are equally likely. Instead we can view an experiment as tossing a first die and a second die and recording the pair of numbers that occur on each of the dice. Each of these pairs is as likely as any other to occur. So

$$S = \left\{ \begin{array}{cccccc} (1,1), & (1,2), & (1,3), & (1,4), & (1,5), & (1,6) \\ (2,1), & (2,2), & (2,3), & (2,4), & (2,5), & (2,6) \\ (3,1), & (3,2), & (3,3), & (3,4), & (3,5), & (3,6) \\ (4,1), & (4,2), & (4,3), & (4,4), & (4,5), & (4,6) \\ (5,1), & (5,2), & (5,3), & (5,4), & (5,5), & (5,6) \\ (6,1), & (6,2), & (6,3), & (6,4), & (6,5), & (6,6) \end{array} \right\}$$

and $|S| = 6 \times 6 = 36$. The event that the sum of the dice is an eight is

$$E = \{(2,6), (3,5), (4,4), (5,3), (6,2)\}.$$

Therefore, $P(E) = \frac{|E|}{|S|} = \frac{5}{36}$. ■

Example 1.2.4. A seven letter code is selected at random with every code as likely to be selected as any other code (so AQRVTAS and CRXAOLZ would be two possibilities). How likely is it that such a code has at least one letter used more than once? (This would happen with the first code above with a repeated A - but not with the second).

As with the examples above, the solution amounts to counting numbers of outcomes. However, unlike the examples above the numbers involved here are quite large and we will need to use some combinatorics to find the solution. The sample space S consists of all seven-letter codes from AAAAAAA to ZZZZZZZ. Each of the seven spots in the code could be any of twenty-six letters, so $|S| = 26^7 = 8,031,810,176$. If E is the event for which there is at least one letter used more than once, it is easier to count E^c, the event

where no letter is repeated. Since in this case each new letter rules out a possibility for the next letter there are $26 \times 25 \times 24 \times 23 \times 22 \times 21 \times 20 = 3,315,312,000$ such possibilities.

This lets us compute $P(E^c) = \frac{3,315,312,000}{8,031,810,176}$ from which we find $P(E) = 1 - P(E^c) = \frac{4,716,498,176}{8,031,810,176} \approx 0.587$. That is, there is about a 58.7% chance that such a code will have a repeated letter. ■

Example 1.2.5. A group of twelve people includes Grant and Dilip. A group of three people is to be randomly selected from the twelve. How likely is it that this three-person group will include Grant, but not Dilip?

Here, S is the collection of all three-person groups, each of which is as likely to be selected as any other. The number of ways of selecting a three-person group from a pool of twelve is $|S| = \binom{12}{3} = 220$. The event E consists of those three-person groups that include Grant, but not Dilip. Such groups must include two people other than Grant and there are ten people remaining from which to select the two, so $|E| = \binom{10}{2} = 45$. Therefore, $P(E) = \frac{45}{220} = \frac{9}{44}$. ■

EXERCISES

Ex. 1.2.1. A day is selected at random from a given week with each day as likely to be selected as any other.

(a) What is the sample space S? What is the size of S?

(b) Let E be the event that the selected day is a Saturday or a Sunday. What is the probability of E.

Ex. 1.2.2. A box contains 500 envelopes, of which 50 contain Rs 100 in cash, 100 contain Rs 50 in cash and 350 contain Rs 10. An envelope can be purchased at Rs 25 from the owner, who will pick an envelope at random and give it to you. Write down the sample space for the net money gained by you. If each envelope is as likely to be selected as any other envelope, what is the probability that the first envelope purchased contains less than Rs 100?

Ex. 1.2.3. Three dice are tossed.

(a) Describe (in words) the sample space S and give an example of an object in S.

(b) What is the size of S?

(c) Let E be the event that the first two dice both come up "1". What is the size of E? What is the probability of E?

(d) Let G be the event that the three dice show three different numbers. What is the size of G? What is the probability of G?

(e) Let F be the event that the third die is larger than the sum of the first two. What is the size of F? What is the probability of F?

Ex. 1.2.4. Suppose that each of three women at a party throws her hat into the center of the room. The hats are first mixed up and then each one randomly selects a hat. Describe the probability space for the possible selection of hats. If all of these selections are equally likely, what is the probability that none of the three women selects her own hat?

Ex. 1.2.5. A group of ten people includes Sona and Adam. A group of five people is to be randomly selected from the ten. How likely is it that this group of five people will include neither Sona nor Adam?

Ex. 1.2.6. There are eight students with two females and six males. They are split into two groups A and B, of four each.

(a) In how many different ways can this be done?

(b) What is the probability that two females end up in group A?

(c) What is the probability that there is one female in each group?

Ex. 1.2.7. Sheela has lost her key to her room. The security officer gives her 50 keys and tells her that one of them will open her room. She decides to try each key successively and notes down the number of the attempt at which the room opens. Describe the sample space for this experiment. Do you think it is realistic that each of these outcomes is equally likely? Why or why not?

Ex. 1.2.8. Suppose that n balls, of which k are red, are arranged at random in a line. What is the probability that all the red balls are next to each other?

Ex. 1.2.9. Consider a deck of 50 cards. Each card has one of 5 colors (black, blue, green, red, and yellow), and is printed with a number (1,2,3,4,5,6,7,8,9, or 10) so that each of the 50 color/number combinations is represented exactly once. A **hand** is produced by dealing out five different cards from the deck. The order in which the cards were dealt does not matter.

(a) How many different hands are there?

(b) How many hands consist of cards of identical color? What is the probability of being dealt such a hand?

(c) What is the probability of being dealt a hand that contains exactly three cards with one number, and two cards with a different number?

(d) What is the probability of being dealt a hand that contains two cards with one number, two cards with a different number, and one card of a third number?

Ex. 1.2.10. Suppose you are in charge of quality control for a light bulb manufacturing company. Suppose that in the process of producing 100 light bulbs, either all 100 bulbs will work properly, or through some manufacturing error twenty of the 100 will not work. Suppose your quality control procedure is to randomly select ten bulbs from a 100 bulb batch and test them to see if they work properly. How likely is this procedure to detect if a batch has bad bulbs in it?

Ex. 1.2.11. A fair die is rolled five times. What is the probability of getting at least two 5's and at least two 6's among the five rolls.

Ex. 1.2.12. **(The "Birthday Problem")** For a group of N people, if their birthdays were listed one-by-one, there are 365^N different ways that such a list might read (if we ignore February 29 as a possibility). Suppose each of those possible lists is as likely as any other.

(a) For a group of two people, let E be the event that they have the same birthday. What is the size of E? What is the probability of E?

(b) For a group of three people, let F be the event that at least two of the three have the same birthday. What is the size of F? What is the probability of F? (Hint: It is easier to find the size of F^c than it is to find the size of F).

(c) For a group of four people, how likely is it that at least two of the four have the same birthday?

(d) How large a group of people would you need to have before it becomes more likely than not that at least two of them share a birthday?

Ex. 1.2.13. A coin is tossed 100 times.

(a) How likely is it that the 100 tosses will produce exactly fifty heads and fifty tails?

(b) How likely is it that the number of heads will be between 50 and 55 (inclusive)?

Ex. 1.2.14. Suppose I have a coin that I claim is "fair" (equally likely to come up heads or tails) and that my friend claims is weighted towards heads. Suppose I flip the coin twenty times and find that it comes up heads on sixteen of those twenty flips. While this seems to favor my friend's hypothesis, it is still possible that I am correct about the coin and that

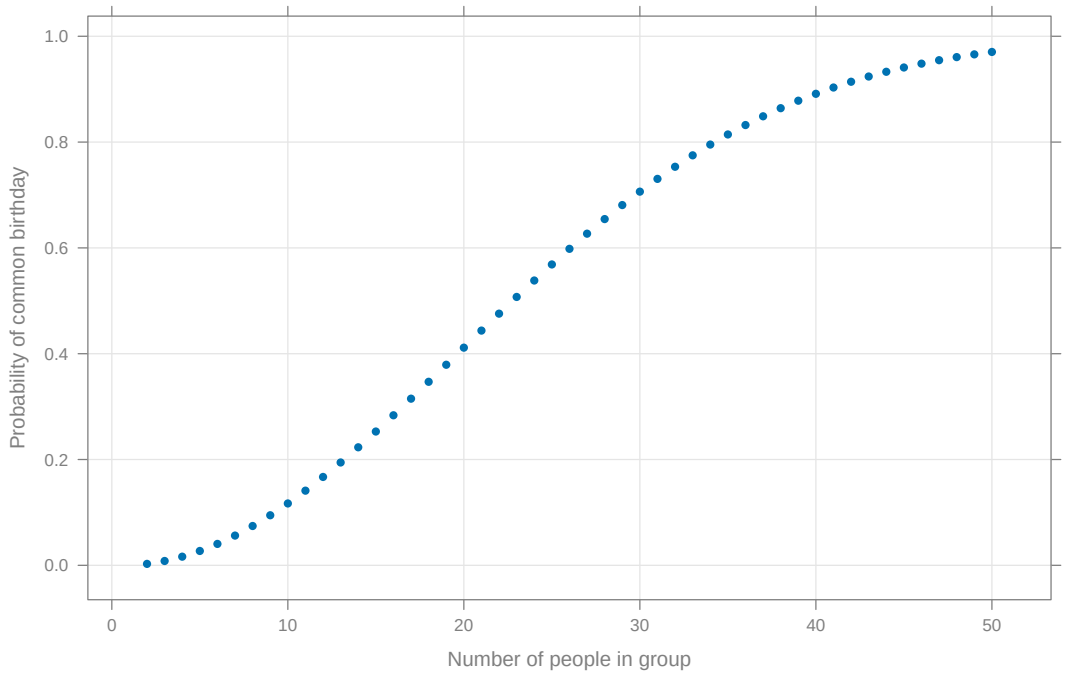


Figure 1.2: The birthday problem discussed in Exercise 1.2.12

just by chance the coin happened to come up heads more often than tails on this series of flips. Let S be the sample space of all possible sequences of flips. The size of S is then 2^{20}, and if I am correct about the coin being "fair", each of these outcomes are equally likely.

(a) Let E be the event that exactly sixteen of the flips come up heads. What is the size of E? What is the probability of E?

(b) Let F be the event that at least sixteen of the flips come up heads. What is the size of F? What is the probability of F?

Note that the probability of F is the chance of getting a result as extreme as the one I observed if I happen to be correct about the coin being fair. The larger $P(F)$ is, the more reasonable seems my assumption about the coin being fair. The smaller $P(F)$ is, the more that assumption looks doubtful. This is the basic idea behind the statistical concept of "hypothesis testing" which we will revisit in Chapter 9.

Ex. 1.2.15. Suppose that r indistinguishable balls are placed in n distinguishable boxes so that each distinguishable arrangement is equally likely. Find the probability that no box will be empty.

Ex. 1.2.16. Suppose that 10 potato sticks are broken into two - one long and one short piece. The 20 pieces are now arranged into 10 random pairs chosen uniformly.

(a) Find the probability that each of pairs consists of two pieces that were originally part of the same potato stick.

(b) Find the probability that each pair consists of a long piece and a short piece.

Ex. 1.2.17. Let S be a non-empty, countable (finite or infinite) set such that for each $\omega \in S, 0 \leq p_\omega \leq 1$. Let $\mathcal{F}$ be the collection of all events. Suppose $P : \mathcal{F} \to [0,1]$ is given by

$$P(E) = \sum_{\omega \in E} p_\omega,$$

for any event E.

(a) Show that P satisfies Axiom 2 in Definition 1.1.3.

(b) Further, conclude that if $P(S) = 1$ then P defines a probability on S.

1.3 CONDITIONAL PROBABILITY AND BAYES' THEOREM

In the previous section we introduced an axiomatic definition of "probability" and discussed the concept of an "event". Now we look at ways in which the knowledge that one event has occurred may be used as information to inform and alter the probability of another event.

Example 1.3.1. Consider the experiment of tossing a fair coin three times with sample space $S = \{hhh, hht, hth, htt, thh, tht, tth, ttt\}$. Let A be the event that there are two or more heads. As all outcomes are equally likely,

$$P(A) = \frac{|A|}{|S|} = \frac{|\{hhh, hht, hth, thh\}|}{8} = \frac{1}{2}.$$

Let B be the event that there is a head in the first toss. As above,

$$P(B) = \frac{|B|}{|S|} = \frac{|\{hhh, hht, hth, htt\}|}{8} = \frac{1}{2}.$$

Now suppose we are asked to find the probability of at least two or more heads among the three tosses, but we are also given the additional information that the first toss was a head. In other words, we are asked to find the probability of A, given the information that event B has definitely occurred. As the additional information guarantees that B is now a list of all possible outcomes, it makes intuitive sense to view the event B as a new sample

space and then identify the subset $A \cap B = \{hhh, hht, hth\}$ of B consisting of outcomes for which there are at least two heads. We conclude that the probability of at least two or more heads in three tosses given that the first toss was a head is

$$\frac{|A \cap B|}{|B|} = \frac{3}{4}.$$

■

This is a legitimate way to view the problem and it leads to the correct solution. However, this method has one very serious drawback—it requires us to change both our sample space and our probability function in order to carry out the computation. It would be preferable to have a method that allows us to work within the original framework of the sample space S and to talk about the "conditional probability" of A given that the result of the experiment will be an outcome in B. This is denoted as $P(A \mid B)$ and is read as "the (conditional) probability of A given B."

Suppose S is a finite set of equally likely outcomes from a given experiment. Then for any two non-empty events A and B, the conditional probability of A given B is given by

$$\frac{|A \cap B|}{|B|} = \frac{\frac{|A \cap B|}{|S|}}{\frac{|B|}{|S|}} = \frac{P(A \cap B)}{P(B)}.$$

This leads us to a formal definition of conditional probability for general sample spaces.

Definition 1.3.2. (Conditional Probability) *Let S be a sample space with probability P. Let A and B be two events with $P(B) > 0$. Then the conditional probability of A given B written as $P(A \mid B)$ and is defined by*

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

This definition makes it possible to compute a conditional probability in terms of the original (unconditioned) probability function.

Example 1.3.3. A pair of dice are thrown. If it is known that one die shows a 4, what is the probability the other die shows a 6?

Let B be the event that one of the dice shows a 4. So,

$$B = \Big\{(4,1), (4,2), (4,3), (4,4), (4,5), (4,6), (1,4), (2,4), (3,4), (5,4), (6,4)\Big\}.$$

Let A be the event that one of the dice is a 6. So,

$$A = \Big\{(6,1), (6,2), (6,3), (6,4), (6,5), (6,6), (1,6), (2,6), (3,6), (4,6), (5,6)\Big\}.$$

then

$$A \cap B = \big\{(4,6),(6,4)\}.$$

Hence

$$\begin{aligned} P(\text{one die shows 6} \mid \text{one die shows 4}) &= P(A \mid B) = \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(\text{one die shows 6 and the other shows 4})}{P(\text{one die shows 4})} \\ &= \frac{2/36}{11/36} = \frac{2}{11}. \end{aligned}$$

■

In many applications, the conditional probabilities are implicitly defined within the context of the problem. In such cases, it is useful to have a method for computing non-conditional probabilities from the given conditional ones. Two such methods are given by the next results and the subsequent examples.

Example 1.3.4. An economic model predicts that if interest rates rise, then there is a 60% chance that unemployment will increase, but that if interest rates do not rise, then there is only a 30% chance that unemployment will increase. If the economist believes there is a 40% chance that interest rates will rise, what should she calculate is the probability that unemployment will increase?

Let B be the event that interest rates rise and A be the event that unemployment increases. We know the values

$$P(B) = 0.4,\ P(B^c) = 0.6,\ P(A \mid B) = 0.6,\ \text{and}\ P(A \mid B^c) = 0.3.$$

Using the axioms of probability and definition of conditional probability we have

$$\begin{aligned} P(A) &= P((A \cap B) \cup (A \cap B^c)) \\ &= P((A \cap B)) + P(A \cap B^c)) \\ &= P(A \mid B)P(B) + P(A \mid B^c)P(B^c) \\ &= 0.6 \times 0.4 + 0.3 \times 0.6 = 0.42. \end{aligned}$$

So there is a 42% chance that unemployment will increase. ■

Theorem 1.3.5. *Let A be an event and let $\{B_i : 1 \leq i \leq n\}$ be a disjoint collection of events for which $P(B_i) > 0$ for all i and such that $A \subset \bigcup_{i=1}^{n} B_i$. Suppose $P(B_i)$ and $P(A \mid B_i)$ are known. Then $P(A)$ may be computed as*

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i)P(B_i).$$

Proof. The events $(A \cap B_i)$ and $(A \cap B_j)$ are disjoint if $i \neq j$ and

$$\bigcup_{i=1}^{n}(A \cap B_i) = A \cap \left(\bigcup_{i=1}^{n} B_i\right) = A.$$

So,

$$\begin{aligned} P(A) &= P\left(\bigcup_{i=1}^{n}(A \cap B_i)\right) \\ &= \sum_{i=1}^{n} P(A \cap B_i) \\ &= \sum_{i=1}^{n} P(A \mid B_i)\, P(B_i). \end{aligned}$$

■

A nearly identical proof holds when there are only countably many B_i (see Exercise 1.3.11).

Example 1.3.6. Suppose we have coloured balls distributed in three boxes in quantities as given by the table below:

	Box 1	Box 2	Box 3
Red	4	3	3
Green	3	3	4
Blue	5	2	3

A box is selected at random. From that box a ball is selected at random. How likely is it that a red ball is drawn?

Let B_1, B_2, and B_3 be the events that Box 1, 2, or 3 is selected, respectively. Note that these events are disjoint and cover all possibilities in the sample space. Let R be the event that the selected ball is red. Then by Theorem 1.3.5,

$$\begin{aligned} P(R) &= P(R \mid B_1)P(B_1) + P(R \mid B_2)P(B_2) + P(R \mid B_3)P(B_3) \\ &= \frac{4}{12} \cdot \frac{1}{3} + \frac{3}{8} \cdot \frac{1}{3} + \frac{3}{10} \cdot \frac{1}{3} = \frac{121}{360}. \end{aligned}$$

■

Example 1.3.7. **(Polya's Urn Scheme)** Suppose there is an urn that contains r red balls and b black balls. A ball is drawn at random and its colour noted. It is replaced with $c > 0$ balls of the same colour. The procedure is then repeated. For $j = 1, 2, \ldots$, let R_j and B_j be the events that the j-th ball drawn is red and black respectively. Clearly $P(R_1) = \frac{r}{b+r}$ and $P(B_1) = \frac{b}{b+r}$. When the first ball is replaced, c new balls will be added to the urn, so that when the second ball is drawn there will be $r + b + c$ balls available. From this it can easily be checked that $P(R_2 \mid R_1) = \frac{r+c}{b+r+c}$ and $P(R_2 \mid B_1) = \frac{r}{b+r+c}$. Noting that R_1 and B_1 are disjoint and together represent the entire sample space, $P(R_2)$ can be computed as

$$\begin{aligned} P(R_2) &= P(R_1)P(R_2 \mid R_1) + P(B_1)P(R_2 \mid B_1) \\ &= \frac{r}{b+r} \cdot \frac{r+c}{b+r+c} + \frac{b}{b+r} \cdot \frac{r}{b+r+c} \\ &= \frac{r(r+b+c)}{(r+b+c)(b+r)} \frac{r}{b+r} = P(R_1). \end{aligned}$$

One can show that $P(R_j) = \frac{r}{b+r}$ for all $j \geq 1$. ■

The urn schemes were originally developed by George Polya (1887–1985). Various modifications to Polya's urn scheme are discussed in the exercises.

Above we have described how conditioning on an event B may be viewed as modifying the original probability based on the additional information provided by knowing that B has occurred. Frequently in applications, we gain information more than once in the process of an experiment. The following theorem shows how to deal with such a situation.

Theorem 1.3.8. *For an integer $n \geq 2$, let $A_1, A_2, \ldots, A_n$ be a collection of events for which $\bigcap_{j=1}^{n-1} A_j$ has positive probability. Then,*

$$P\left(\bigcap_{j=1}^{n} A_j\right) = P(A_1) \cdot \prod_{j=2}^{n} P\left(A_j \,\middle|\, \bigcap_{k=1}^{j-1} A_k\right).$$

The proof of this theorem is left as Exercise 1.3.14, but we will provide a framework in which to make sense of the equality. Usually the events $A_1, \ldots, A_n$ are viewed as a sequence in time for which we know the probability of a given event provided that all of the others before it have already occurred. Then we can calculate $P(A_1 \cap A_2 \cap \cdots \cap A_n)$ by taking the product of the values $P(A_1), P(A_2 \mid A_1), P(A_3 \mid A_1 \cap A_2), \ldots, P(A_n \mid A_1 \cap \cdots \cap A_{n-1})$.

Example 1.3.9. A probability class has fifteen students - four seniors, eight juniors, and three sophomores. Three different students are selected at random to present homework

problems. What is the probability the selection will be a junior, a sophomore, and a junior again, in that order?

Let A_1 be the event that the first selection is a junior. Let A_2 be the event that the second selection is a sophomore, and let A_3 be the event that the third selection is a junior. The problem asks for $P(A_1 \cap A_2 \cap A_3)$ which we can calculate using Theorem 1.3.8.

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \\ &= \frac{8}{15} \cdot \frac{3}{14} \cdot \frac{7}{13} = \frac{4}{65}. \end{aligned}$$

■

1.3.1 *Bayes' Theorem*

It is often the case that we know the conditional probability of A given B, but want to know the conditional probability of B given A instead. It is possible to calculate one quantity from the other using a formula known as Bayes' theorem. We introduce this with a motivating example.

Example 1.3.10. We return to Example 1.3.6. In that example we had three boxes containing balls given by the table below.

	Box 1	Box 2	Box 3
Red	4	3	3
Green	3	3	4
Blue	5	2	3

A box is selected at random. From the box a ball is selected at random. When we looked at conditional probabilities we saw how to determine the probability of an event such as {the ball drawn is red}. Now suppose we know the ball is red and want to determine the probability of the event {the ball was drawn from box 3}. That is, if R is the event that a red ball is chosen and if B_1, B_2, and B_3 are the events that boxes 1, 2, and 3 are selected, we want to determine the conditional probability $P(B_3 \mid R)$. The difficulty is that while the conditional probabilities $P(R \mid B_1), P(R \mid B_2)$, and $P(R \mid B_3)$ are easy to determine, calculating the conditional probability with the order of the events reversed is not immediately obvious.

Using the definition of conditional probability we have that

$$P\Big(B_3 \mid R\Big) = \frac{P(B_3 \cap R)}{P(R)},$$

so we can rewrite $P(B_3 \cap R) = P(R \mid B_3)P(B_3) = \frac{3}{10} \times \frac{1}{3} = \frac{1}{10}$. On the other hand, we can decompose the event R over which box was chosen. This is exactly what we did to solve Example 1.3.6 where we found that $P(R) = \frac{121}{360}$. Hence,

$$P(B_3 \mid R) = \frac{P(B_3 \cap R)}{P(R)} = \frac{1/10}{121/360} = \frac{36}{121} \approx 0.298.$$

So if we know that a red ball was drawn, there is slightly less than a 30% chance that it came from Box 3. ■

In the above example the description of the experiment allowed us to determine $P(B_1)$, $P(B_2)$, $P(B_3)$, $P(R \mid B_1)$, $P(R \mid B_2)$, and $P(R \mid B_3)$. We were then able to use the definition of conditional probability to find $P(B_3 \mid R)$. Such a computation can be done in general.

Theorem 1.3.11. (Bayes' Theorem) *Suppose A is an event, $\{B_i : 1 \leq i \leq n\}$ are a collection of disjoint events whose union contains all of A. Further assume that $P(A) > 0$ and $P(B_i) > 0$ for all $1 \leq i \leq n$. Then for any $1 \leq i \leq n$,*

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{\sum_{j=1}^{n} P(A \mid B_j)P(B_j)}. \tag{1.3.1}$$

Proof. The left hand side of (1.3.1) can be written as

$$\begin{aligned} P(B_i \mid A) &= \frac{P(B_i \cap A)}{P(A)} = \frac{P(A \mid B_i)P(B_i)}{P\left(\bigcup_{j=1}^{n} A \cap B_j\right)} \\ &= \frac{P(A \mid B_i)P(B_i)}{\sum_{j=1}^{n} P(A \cap B_j)} = \frac{P(A \mid B_i)P(B_i)}{\sum_{j=1}^{n} P(A \mid B_j)P(B_j)}. \end{aligned}$$

■

Equation (1.3.1) is sometimes referred to as "Bayes' formula" or "Bayes' rule" as well. This result is originally due to Thomas Bayes (1701–1761).

Example 1.3.12. Shyam is randomly selected from the citizens of Hyderabad by the Health authorities. A laboratory test on his blood sample tells Shyam that he has tested positive for Swine Flu. It is found that 95% of people with Swine Flu test positive but 2% of people without the disease will also test positive. Suppose that 1% of the population has the disease. What is the probability that Shyam indeed has the Swine Flu ?

Consider the events $A = \{$ Shyam has Swine Flu $\}$ and $B = \{$ Shyam tested postive for Swine Flu $\}$. We are given:

$$P(B \mid A) = 0.95,\ P(B \mid A^c) = 0.02,\ \text{ and } P(A) = 0.01.$$

Using Bayes' Theorem, we have

$$\begin{aligned} P(A \mid B) &= \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)} \\ &= \frac{(0.95)(0.01)}{(0.95)(0.01) + (0.02)(0.99)} = 0.324 \end{aligned}$$

Despite testing positive, there is only a 32.4 percent chance that Shyam has the disease. ■

EXERCISES

Ex. 1.3.1. There are two dice, one red and one blue, sitting on a table. The red die is a standard die with six sides while the blue die is tetrahedral with four sides, so the outcomes 1, 2, 3, and 4 are all equally likely. A fair coin is flipped. If that coin comes up heads, the red die will be rolled, but if the coin comes up tails the blue die will be rolled.

(a) Find the probability that the rolled die will show a 1.

(b) Find the probability that the rolled die will show a 6.

Ex. 1.3.2. A pair of dice are thrown. It is given that the outcome on one die is a 3. what is the probability that the sum of the outcomes on both dice is greater than 7?

Ex. 1.3.3. Box A contains four white balls and three black balls and Box B contains three white balls and five black balls.

(a) Suppose a box is selected at random and then a ball is chosen from the box. If the ball drawn is black then what is the probability that it was from Box A?

(b) Suppose instead that one ball is drawn at random from Box A and placed (unseen) in Box B. What is the probability that a ball now drawn from Box B is black?

Ex. 1.3.4. Tomorrow the weather will either be sunny, cloudy, or rainy. There is a 60% chance tomorrow will be cloudy, a 30% chance tomorrow will be sunny, and a 10% chance that tomorrow will be rainy. If it rains, I will not go on a walk. But if it is cloudy, there is a 90% chance I will take a walk and if it's sunny there is a 70% chance I will take a walk. If I take a walk on a cloudy day, there is an 80% chance I will walk further than

five kilometers, but if I walk on a sunny day, there's only a 50% chance I will walk further than five kilometers. Using the percentages as given probabilities, answer the following questions:

(a) How likely is it that tomorrow will be cloudy and I will walk over five kilometers?

(b) How likely is it I will take a walk over five kilometers tomorrow?

Ex. 1.3.5. A box contains B black balls and W balls, where $W \geq 3, B \geq 3$. A sample of three balls is drawn at random with each drawn ball being discarded (not put back into the box) after it is drawn. For $j = 1, 2, 3$ let A_j denote the event that the ball drawn on the j^{th} draw is white. Find $P(A_1)$, $P(A_2)$ and $P(A_3)$.

Ex. 1.3.6. There are two sets of cards, one red and one blue. The red set has four cards - one that reads 1, two that read 2, and one that reads 3. An experiment involves flipping a fair coin. If the coin comes up heads a card will be randomly selected from the red set (and its number recorded) while if the coin comes up tails a card will be randomly selected from the blue set (and its number recorded). You can construct the blue set of cards in any way you see fit using any number of cards reading 1, 2, or 3. Explain how to build the blue set of cards to make each of the experimental outcomes $1, 2, 3$ equally likely.

Ex. 1.3.7. There are three tables, each with two drawers. Table 1 has a red ball in each drawer. Table 2 has a blue ball in each drawer. Table 3 has a red ball in one drawer and a blue ball in the other. A table is chosen at random, then a drawer is chosen at random from that table. Find the conditional probability that Table 1 is chosen, given that a red ball is drawn.

Ex. 1.3.8. In the G.R.E advanced mathematics exam, each multiple choice question has 4 choices for an answer. A prospective graduate student taking the test knows the correct answer with probability $\frac{3}{4}$. If the student does not know the answer, she guesses randomly. Given that a question was answered correctly, find the conditional probability that the student knew the answer.

Ex. 1.3.9. You first roll a fair die, then toss as many fair coins as the number that showed on the die. Given that 5 heads are obtained, what is the probability that the die showed 5?

Ex. 1.3.10. Manish is a student in a probability class. He gets a note saying, "I've organized a probability study group tonight at 7pm in the coffee shop. Come if you want." The note is signed "Hannah". However, Manish has class with two different Hannahs and he isn't sure which one sent the note. He figures that there is a 75% chance that Hannah A. would have organized such a study group, but only a 25% chance that Hannah B. would have done so. However, he also figures that if Hannah A. had organized the group, there is an

80% chance that she would have planned to meet on campus and only a 20% chance that she would have planned to meet in the coffee shop. While if Hannah B. had organized the group there is a 10% chance she would have planned for it on campus and a 90% chance she would have chosen the coffee shop. Given all this information, determine whether it is more likely that Manish should think the note came from Hannah A. or from Hannah B.

Ex. 1.3.11. State and prove a version of

(a) Theorem 1.3.5 when $\{B_i\}$ is a countably infinite collection of disjoint events.

(b) Theorem 1.3.11 when $\{B_i\}$ is a countably infinite collection of disjoint events.

Ex. 1.3.12. A bag contains 100 coins. Sixty of the coins are fair. The rest are biased to land heads with probability p (where $0 \leq p \leq 1$). A coin is drawn at random from the bag and tossed.

(a) Given that the outcome was a head what is the conditional probability that it is a biased coin?

(b) Evaluate your answer to (a) when $p = 0$. Can you explain why this answer should be intuitively obvious?

(c) Evaluate your answer to (a) when $p = \frac{1}{2}$. Can you explain why this answer should be fairly intuitive as well?

(d) View your answer to part (a) as a function $f(p)$. Show that $f(p)$ is an increasing function when $0 \leq p \leq 1$. Give an interpretation of this fact in the context of the problem.

Ex. 1.3.13. An urn contains b black balls and r red balls. A ball is drawn at random. The ball is replaced into the urn along with c balls of its colour and d balls of the opposite colour. Then another random ball is drawn and the procedure is repeated.

(a) What is the probability that the second ball drawn is a red ball?

(b) Assume $c = d$. What is the probability that the second ball drawn is a black ball?

(c) Still assuming $c = d$, what is the probability that the n^{th} ball drawn is a black ball?

(d) Assume $c > 0$ and $d = 0$, what is the probability that the n^{th} ball drawn is a black ball?

(6) Can you comment on the answers to (b) and/or (c) if the assumption that $c = d$ was removed?

Ex. 1.3.14. Use the following steps to prove Theorem 1.3.8.

(a) Prove Theorem 1.3.8 for the $n = 2$ case. (Hint: The proof should follow immediately from the definition of conditional probability).

(b) Prove Theorem 1.3.8 for the $n = 3$ case. (Hint: Rewrite the conditional probabilities in terms of ordinary probabilities).

(c) Prove Theorem 1.3.8 generally. (Hint: One method is to use induction, and parts (a) and (b) have already provided a starting point).

1.4 INDEPENDENCE

In the previous section we have seen instances where the probability of an event may change given the occurrence of a related event. However it is instructive and useful to study the case of two events where the occurrence of one has no effect on the probability of the other. Such events are said to be "independent".

Example 1.4.1. Suppose we toss a coin three times. Then the sample space

$$S = \{hhh, hht, hth, htt, thh, tht, tth, ttt\}.$$

Define $A = \{hhh, hht, hth, htt\} = \{\text{the first toss is a head}\}$, and similarly define $B = \{hhh, hht, thh, tht\} = \{\text{the second toss is a head}\}$. Note that $P(A) = \frac{1}{2} = P(B)$, while

$$P(A \mid B) = \frac{P(A \cap B)}{P(A)} = \frac{|A \cap B|}{|B|} = \frac{2}{4} = \frac{1}{2}$$

and

$$P(A \mid B^c) = \frac{P(A \cap B^c)}{P(B^c)} = \frac{|A \cap B^c|}{|B^c|} = \frac{2}{4} = \frac{1}{2}.$$

We have shown that $P(A) = P(A \mid B) = P(A \mid B^c)$. Therefore we conclude that the occurrence (or non-occurrence) of B has no effect on the probability of A. ■

This is the sort of condition we would want in a definition of independence. However, since defining $P(A \mid B)$ requires that $P(B) > 0$, our formal definition of "independence" will appear slightly different.

Definition 1.4.2. (Independence) *Two events A and B are independent if*

$$P(A \cap B) = P(A)P(B).$$

Example 1.4.3. Suppose we roll a die twice and denote as an ordered pair the result of the rolls. Suppose

$$E = \{ \text{ a six appears on the first roll } \} = \{(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}$$

and

$$F = \{ \text{ a six appears on the second roll } \} = \{(1,6),(2,6),(3,6),(4,6),(5,6),(6,6)\}.$$

As $E \cap F = \{(6,6)\}$, it is easy to see that

$$P(E \cap F) = \frac{1}{36}, P(E) = \frac{6}{36} = \frac{1}{6}, P(F) = \frac{6}{36} = \frac{1}{6}.$$

So E, F are independent as $P(E \cap F) = P(E)P(F)$. ■

Using the definition of conditional probability it is not hard to show (see Exercise 1.4.9) that if A and B are independent, and if $0 < P(B) < 1$ then

$$P(A \mid B) = P(A) = P(A \mid B^c). \tag{1.4.1}$$

If $P(A) > 0$ then the equations of (1.4.1) also hold with the roles of A and B reversed. Thus, independence implies four conditional probability equalities.

If we want to extend our definition of independence to three events A_1, A_2, and A_3, we would certainly want

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3) \tag{1.4.2}$$

to hold. We would also want any pair of the three events to be independent of each other. It is tempting to hope that pairwise independence is enough to imply (1.4.2). However, this is not true, as shown by the next example.

Example 1.4.4. Suppose we toss a fair coin two times. Consider the three events $A_1 = \{hh, tt\}$, $A_2 = \{hh, ht\}$, and $A_3 = \{hh, th\}$. Then it is easy to calculate that

$$\begin{aligned} P(A_1) = P(A_2) = P(A_3) &= \frac{1}{2}, \\ P(A_1 \cap A_2) = P(A_1 \cap A_3) = P(A_2 \cap A_3) &= \frac{1}{4}, \text{ and} \\ P(A_1 \cap A_2 \cap A_3) &= \frac{1}{4}. \end{aligned}$$

So even though A_1, A_2 and A_3 are pairwise independent, they do not satisfy (1.4.2). ■

It may also be tempting to hope that (1.4.2) is enough to imply pairwise independence, but that is not true either (see Exercise 1.4.6). The root of the problem is that, unlike the two event case, (1.4.2) does not imply that equality holds if any of the A_i are replaced by their complements. One solution is to insist that the multiplicative equality hold for any intersection of the events or their complements, which gives us the following definition.

Definition 1.4.5. (Mutual Independence) *A finite collection of events $A_1, A_2, \ldots, A_n$ is mutually independent if*

$$P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_1)P(E_2)\ldots P(E_n). \tag{1.4.3}$$

whenever E_j is either A_j or A_j^c.
An arbitrary collection of events A_t where $t \in I$ for some index set I is mutually independent if every finite subcollection is mutually independent.

Thus, mutual independence of n events is defined in terms of 2^n equations. It is a fact (see Exercise 1.4.10) that if a collection of events is mutually independent, then so is any subcollection.

EXERCISES

Ex. 1.4.1. In the first semifinal of an international volleyball tournament Brazil has a 60% chance to beat Pakistan. In the other semifinal Poland has a 70% chance to beat Mexico. If the results of the two matches are independent, what is the probability that Pakistan will meet Poland in the tournament final?

Ex. 1.4.2. A manufacturer produces nuts and markets them as having 50mm radius. The machines that produce the nuts are not perfect. From repeated testing, it was established that 15% of the nuts have radius below 49mm and 12% have radius above 51mm. If two nuts are randomly (and independently) selected, find the probabilities of the following events:

(a) The radii of both the nuts are between 49mm and 51mm;

(b) The radius of at least one nut exceeds 51mm.

Ex. 1.4.3. Four tennis players (Avinash, Ben, Carlos, and David) play a single-elimination tournament with Avinash playing David and Ben playing Carlos in the first round and the winner of each of those contests playing each other in the tournament final. Below is the

chart giving the percentage chance that one player will beat the other if they play. For instance, Avinash has a 30% chance of beating Ben if they happen to play.

	Avinash	Ben	Carlos	David
Avinash	-	30%	55%	40%
Ben	-	-	80%	45%
Carlos	-	-	-	15%
David	-	-	-	-

Suppose the outcomes of the games are independent. For each of the four players, determine the probability that player wins the tournament. Verify that the calculated probabilities sum to 1.

Ex. 1.4.4. Let A and B be events with $P(A) = 0.8$ and $P(B) = 0.7$.

(a) What is the largest possible value of $P(A \cap B)$?

(b) What is the smallest possible value of $P(A \cap B)$?

(c) What is the value of $P(A \cap B)$ if A and B are independent?

Ex. 1.4.5. Suppose we toss two fair dice. Let E_1 denote the event that the sum of the dice is six. E_2 denote the event that sum of the dice equals seven. Let F denote the event that the first die equals four. Is E_1 independent of F? Is E_2 independent of F?

Ex. 1.4.6. Suppose a bowl has twenty-seven balls. One ball is black, two are white, and eight each are green, red, and blue. A single ball is drawn from the bowl and its color is recorded. Define

$$\begin{aligned} A &= \{\text{the ball is either black or green}\} \\ B &= \{\text{the ball is either black or red}\} \\ C &= \{\text{the ball is either black or blue}\} \end{aligned}$$

(a) Calculate $P(A \cap B \cap C)$.

(b) Calculate $P(A)P(B)P(C)$.

(c) Are A, B, and C mutually independent? Why or why not?

Ex. 1.4.7. There are 150 students in the Probability 101 class. Of them, ninety are female, sixty use a pencil (instead of a pen), and thirty are wearing eye glasses. A student is chosen at random from the class. Define the following events:

$$\begin{aligned} A_1 &= \{\text{the student is a female}\} \\ A_2 &= \{\text{the student uses a pencil}\} \\ A_3 &= \{\text{the student is wearing eye glasses}\} \end{aligned}$$

(a) Show that it is impossible for these events to be mutually independent.

(b) Give an example to show that it may be possible for these events to be pairwise independent.

Ex. 1.4.8. When can an event be independent of itself? Do parts (a) and (b) below to answer this question.

(a) Prove that if an event A is independent of itself then either $P(A) = 0$ or $P(A) = 1$.

(b) Prove that if A is an event such that either $P(A) = 0$ or $P(A) = 1$ then A is independent of itself.

Ex. 1.4.9. This exercise explores the relationship between independence and conditional probability.

(a) Suppose A and B are independent events with $0 < P(B) < 1$. Prove that $P(A \mid B) = P(A)$ and that $P(A \mid B^c) = P(A)$.

(b) Suppose that A and B are independent events. Prove that A and B^c are also independent.

(c) Suppose that A and B are events with $P(B) > 0$. Prove that if $P(A \mid B) = P(A)$, then A and B are independent.

(d) Suppose that A and B are events with $0 < P(B) < 1$. Prove that if $P(A \mid B) = P(A)$, then $P(A \mid B^c) = P(A)$ as well.

Ex. 1.4.10. In this section we mentioned the following theorem: "If $E_1, E_2, \ldots, E_n$ is a collection of mutually independent events, then any subcollection of these events is mutually independent". Follow the steps below to prove the theorem.

(a) Suppose A, B, and C are mutually independent. In particular, this means that

$$\begin{aligned} P(A \cap B \cap C) &= P(A) \cdot P(B) \cdot P(C), \text{ and} \\ P(A \cap B \cap C^c) &= P(A) \cdot P(B) \cdot P(C^c). \end{aligned}$$

Use these two facts to conclude that A and B are pairwise independent.

(b) Suppose $E_1, E_2, \ldots, E_n$ is a collection of mutually independent events. Prove that $E_1, E_2, \ldots, E_{n-1}$ is also mutually independent.

(c) Use (b) and induction to prove the full theorem.

1.5 USING R FOR COMPUTATION

As we have already seen, and will see throughout this book, the general approach to solve problems in probability and statistics is to put them in an abstract mathematical framework. Many of these problems eventually simplify to computing some specific numbers. Usually these computations are simple and can be done using a calculator. For some computations however, a more powerful tool is needed. In this book, we will use a software called R to illustrate such computations. R is freely available open source software that runs on a variety of computer platforms, including Windows, macOS, and GNU/Linux.

R is many different things to different people, but for our purposes, it is best to think of it as a very powerful calculator. Once you install and start R,[1] you will be presented with a prompt that looks like the "greater than" sign (>). You can type expressions that you want to evaluate here and press the Enter key to obtain the answer. For example,

```
9 / 44
```

```
[1] 0.2045455
```

```
0.6 * 0.4 + 0.3 * 0.6
```

```
[1] 0.42
```

```
log(0.6 * 0.4 + 0.3 * 0.6)
```

```
[1] -0.8675006
```

It may seem odd to see a `[1]` at the beginning of each answer, but that is there for a good reason. R is designed for statistical computations, which often require working with a collection of numbers, which following standard mathematical terminology are referred to as *vectors*. For example, we may want to do some computations on a vector consisting

[1] Visit https://www.r-project.org/ to download R and learn more about it.

of the first 5 positive integers. Specifically, suppose we want to compute the squares of these integers, and then sum them up. Using R, we can do

```
c(1, 2, 3, 4, 5)^2
```

```
[1]  1  4  9 16 25
```

```
sum(c(1, 2, 3, 4, 5)^2)
```

```
[1] 55
```

Here the construct `c(...)` is used to create a vector containing the first five integers. Of course, doing this manually is difficult for larger vectors, so another useful construct is `m:n` which creates a vector containing all integers from `m` to `n`. Just as we do in mathematics, it is also convenient to use symbols (called "variables") to store intermediate values in long computations. For example, to do the same operations as above for the first 40 integers, we can do

```
x <- 1:40
x
```

```
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22
[23] 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
```

```
x^2
```

```
 [1]    1    4    9   16   25   36   49   64   81  100  121  144  169
[14]  196  225  256  289  324  361  400  441  484  529  576  625  676
[27]  729  784  841  900  961 1024 1089 1156 1225 1296 1369 1444 1521
[40] 1600
```

```
sum(x^2)
```

```
[1] 22140
```

We can now guess the meaning of the number in square brackets at the beginning of each line in the output: when R prints a vector that spans multiple lines, it prefixes each

line by the index of the first element printed in that line. The prefix appears for scalars too because R treats scalars as vectors of length one.

In the example above, we see two kinds of operations. The expression `x^2` is interpreted as an element-wise squaring operation, which means that the result will have the same length as the input. On the other hand, the expression `sum(x)` takes the elements of a vector `x` and computes their sum. The first kind of operation is called a *vectorized operation*, and most mathematical operations in R are of this kind.

To see how this can be useful, let us use R to compute factorials and binomial coefficients, which will turn up frequently in this book. Recall that the binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

represents the number of ways of choosing k items out of n, where for any positive integer m, $m!$ is the product of the first m positive integers. Just as `sum(x)` computes the sum of the elements of `x`, `prod(x)` computes their product. So, we can compute 10! and $\binom{10}{4}$ as

```
prod(1:10)
```

```
[1] 3628800
```

```
prod(1:10) / (prod(1:4) * prod(1:6))
```

```
[1] 210
```

Unfortunately, factorials can quickly become quite big, and may be beyond R's ability to compute precisely even for moderately large numbers. For example, trying to compute $\binom{200}{4}$, we get

```
prod(1:200)
```

```
[1] Inf
```

```
prod(1:200) / (prod(1:4) * prod(1:196))
```

```
[1] NaN
```

The first computation yields `Inf` because at some point in the computation of the product, the result becomes larger than the largest number R can store (this is often called "overflowing"). The second computation essentially reduces to computing `Inf/Inf`, and the resulting `NaN` indicates that the answer is ambiguous. The trivial mathematical fact that

$$\log m! = \sum_{i=1}^{m} \log i$$

comes to our aid here because it lets us do our computations on much smaller numbers. Using this, we can compute

```
logb <- sum(log(1:200)) - sum(log(1:4)) - sum(log(1:196))
logb
```

```
[1] 17.98504
```

```
exp(logb)
```

```
[1] 64684950
```

R actually has the ability to compute binomial coefficients built into it.

```
choose(200, 4)
```

```
[1] 64684950
```

These named operations, such as `sum()`, `prod()` `log()`, `exp()`, and `choose()`, are known as *functions* in R. They are analogous to mathematical functions in the sense that they map some inputs to an output. Vectorized functions map vectors to vectors, whereas summary functions like `sum()` and `prod()` map vectors to scalars. It is common practice in R to make functions vectorized whenever possible. For example, the `choose()` function is also vectorized:

```
choose(10, 0:10)
```

```
 [1]   1  10  45 120 210 252 210 120  45  10   1
```

```
choose(10:20, 4)
```

```
 [1]  210  330  495  715 1001 1365 1820 2380 3060 3876 4845
```

```
choose(2:15, 0:13)
```

```
 [1]   1   3   6  10  15  21  28  36  45  55  66  78  91 105
```

A detailed exposition of R is beyond the scope of this book. In this book, we will only use relatively basic R functions, which we will introduce as and when needed. There are many excellent introductions available for the interested reader. In particular, R is very useful for producing statistical plots, and most figures in this book are produced using R. We do not describe how to create these figures in the book itself, but R code to reproduce them is available on the website.

EXERCISES

Ex. 1.5.1. In R suppose we type in the following

```
x <- c(-15, -11, -4, 0, 7, 9, 16, 23)
```

Find out the output of the built-in functions given below:

`sum(x)` `length(x)` `mean(x)` `var(x)` `sd(x)` `max(x)` `min(x)` `median(x)`

Ex. 1.5.2. Obtain a six-sided die, and throw it ten times, keeping a record of the face that comes up each time. Store these values in a vector variable `x`. Find the output of the built-in functions given in the previous exercise when applied to this vector.

Ex. 1.5.3. Use R to verify the calculations done in Example 1.2.4.

Ex. 1.5.4. We return to the Birthday Problem given in Exercise 1.2.12. Using R, calculate the Probability that at least two from a group of N people share the same birthday, for $N = 10, 12, 17, 26, 34, 40, 41, 45, 75, 105$.

2

SAMPLING AND REPEATED TRIALS

Consider an experiment and an event A within the sample space. We say the experiment is a success if an outcome from A occurs and failure otherwise. Let us consider the following examples:

Experiment	**Sample Space**	**Event Description**	**Event A**	$\mathbf{P(A)}$
Toss a fair coin	$\{H, T\}$	Head appears	$\{H\}$	$\frac{1}{2}$
Roll a die	$\{1, 2, 3, 4, 5, 6\}$	Six appears	$\{6\}$	$\frac{1}{6}$
Roll a die	$\{1, 2, 3, 4, 5, 6\}$	A multiple of 3 appears	$\{3, 6\}$	$\frac{1}{3}$

In typical applications we would repeat an experiment several times independently and would be interested in the total number of successes achieved, a process that may be viewed as sampling from a large population. For instance, a manager in a factory making nuts and bolts, may devise an experiment to choose uniformly from a collection of manufactured bolts and call the experiment a success if the bolt is not defective. Then she would want to repeat such a selection every time and quantify the number of successes.

2.1 BERNOULLI TRIALS

We will now proceed to construct a mathematical framework for independent trials of an experiment where each trial is either a success or a failure. Let p be the probability of success at each trial. The sequence so obtained is called a sequence of Bernoulli trials with parameter p. The trials are named after James Bernoulli (1654-1705).

We will occasionally want to consider a single Bernoulli trial, so we will use the notation Bernoulli(p) to indicate such a distribution. Since we are only interested in the result of the trial, we may view this as a probability on the sample space $S = \{success, failure\}$

where $P(\{success\}) = p$, but more often we will be interested in multiple independent trials. We discuss this in the next example.

Example 2.1.1. Suppose we roll a die twice and ask how likely it is that we observe exactly one 6 between the two rolls. In the previous chapter (See Example 1.4.3) we would have viewed the sample space S as thirty-six equally likely outcomes, each of which was an ordered pair of results of the rolls. But since we are only concerned with whether the die roll is a 6 (success) or not a 6 (failure) we could also view it as two Bernoulli($\frac{1}{6}$) trials. Using notation from Example 1.4.3, note that $P(\text{success on the first roll}) = P(E) = \frac{1}{6}$ and $P(\text{success on the second roll}) = P(F) = \frac{1}{6}$. So

$$\begin{aligned} & P(\{success, success\}) \\ = \; & P(E \cap F) \\ & \text{(using independence)} \\ = \; & P(E)P(F) \\ = \; & P(\text{success on the first roll}) \cdot P(\text{success on the second roll}) \\ = \; & \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}. \end{aligned}$$

We could alternately view S as having only four elements - (success,success), (success,failure), (failure,success), and (failure,failure). The four outcomes are not equally likely, but the fact that the trials are independent allows us to easily compute the probability of each. Through similar computations,

$$\begin{aligned} & P(\{(success, failure)\}) = 5/36, \\ & P(\{(failure, success)\}) = 5/36, \\ & \text{and } P(\{(failure, failure)\}) = 25/36. \end{aligned}$$

To complete the problem, the event of rolling exactly one 6 among the two dice requires exactly one success and exactly one failure. From the list above, this can happen in either of two orders, so the probability of observing exactly one 6 is $\frac{5}{36} + \frac{5}{36} = \frac{10}{36}$. ■

For any two real numbers a, b and any integer $n \geq 1$, it is well known that

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}. \tag{2.1.1}$$

This is the binomial expansion due to Blaise Pascal(1623-1662). It turns out when a and b are positive numbers with $a + b = 1$, the terms in the right hand side above have a probabilistic interpretation. We illustrate it in the example below.

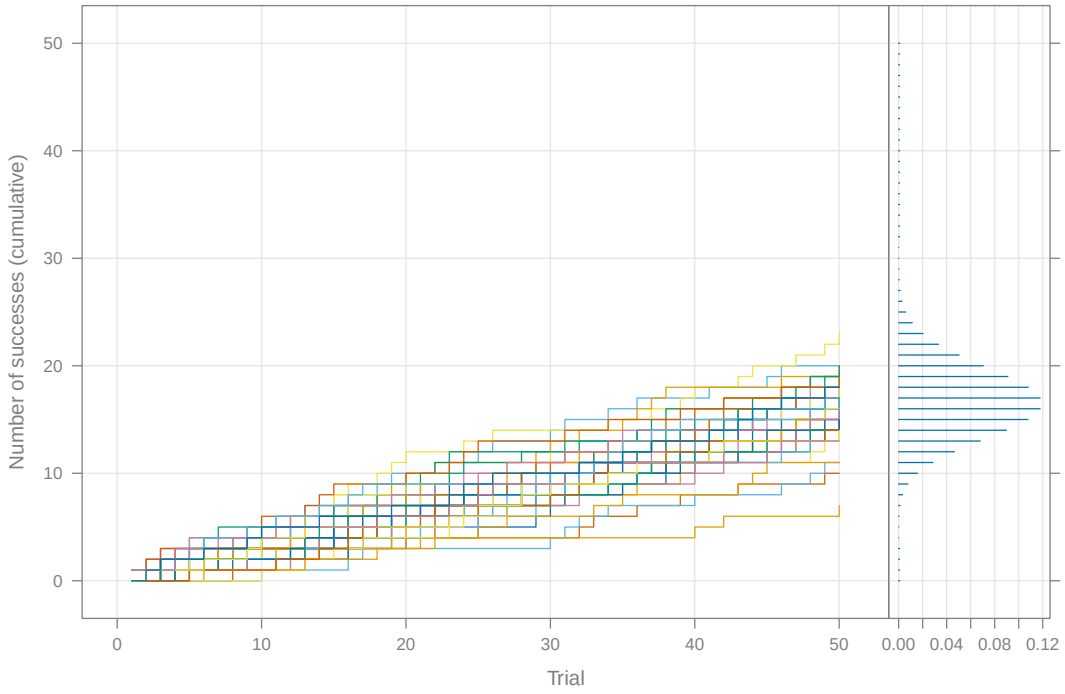


Figure 2.1: The Binomial distribution as number of successes in fifty Bernoulli ($\frac{1}{3}$) trials. The paths on the left count the cumulative successes in the fifty trials. The graph on the right show the actual probability given by the Binomial(50, $\frac{1}{3}$) distribution.

Example 2.1.2. After performing n independent Bernoulli(p) trials we are typically interested in the following questions:

(a) What is the probability of observing exactly k successes?

(b) What is the most likely number of successes?

(c) How many attempts must be made before the first success is observed?

(d) On average how many successes will there be?

Ans (a) - **Binomial(n,p)**: If $n = 1$, then the answer is clear, namely $P(\{\text{one success}\}) = p$ and $P(\{\text{zero successes}\}) = 1 - p$. For, $n > 1$ let $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ be an n-tuple of outcomes. So we may view the sample space S as the set of all ω where each ω_i is allowed to be either "success" or "failure". Let A_i represent either the event {the i^{th} trial is a success} or {the i^{th} trial is a failure}. Then by independence

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = \prod_{i=1}^{n} P(A_i). \tag{2.1.2}$$

Let B_k denote the event that there are k successes among the n trials. Then

$$P(B_k) = \sum_{\omega \in B_k} P(\{\omega\}).$$

But if $\omega \in B_k$, then in notation (2.1.2), exactly k of the A_i represent success trials and the other $n - k$ represent the failure trials. The order in which the successes and failures appear does not matter since the probabilities are being multiplied together. So for every $\omega \in B_k$,

$$P(\{\omega\}) = p^k (1-p)^{n-k}.$$

Consequently, we have

$$P(B_k) = |B_k| p^k (1-p)^{n-k}.$$

But B_k is the event of all outcomes for which there are k successes and the number of ways in which k successes can occur in n trials is known to be $\binom{n}{k}$. Therefore, for $0 \leq k \leq n$,

$$P(B_k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{2.1.3}$$

Note that if we are only interested in questions involving the number of successes, we could ignore the set S described above and simply use $\{0, 1, 2, \ldots, n\}$ as our sample space with $P(\{k\}) = \binom{n}{k} p^k (1-p)^{n-k}$. We call this a Binomial distribution with parameters n and p

(or a Binomial(n, p) for short). It is also worth noting that the binomial expansion (2.1.1) shows

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1,$$

which simply provides additional confirmation that we have accounted for all possible outcomes in our list of Bernoulli trials. See Figure 2.1 for a simulated example of fifty replications of Bernoulli($\frac{1}{3}$) trials.

Ans (b) - **Mode of a Binomial**: The problem is trivial if $p = 0$ or $p = 1$, so assume $0 < p < 1$. Using the same notation for B_k as in part (a), pick a particular number of successes k for which $0 \le k < n$. We want to determine the value of k that makes $P(B_k)$ as large as possible; such a value is called the "mode" . To find this value, it is instructive to compare the probability of $(k+1)$ successes to the probability of k successes –

$$\begin{aligned}
\frac{P(B_{k+1})}{P(B_k)} &= \frac{\binom{n}{k+1} p^{k+1}(1-p)^{n-(k+1)}}{\binom{n}{k} p^k (1-p)^{n-k}} \\
&= \frac{n!}{(k+1)!(n-(k+1))!} \cdot \frac{k!(n-k)!}{n!} \cdot \frac{p^{k+1}(1-p)^{n-(k+1)}}{p^k (1-p)^{n-k}} \\
&= \frac{p}{1-p} \cdot \frac{n-k}{k+1}.
\end{aligned}$$

If this ratio were to equal 1 we could conclude that $\{(k+1)$ successes$\}$ was exactly as likely as $\{k$ successes$\}$. Similarly if the ratio were bigger than 1 we would know that $\{(k+1)$ successes$\}$ was the more likely of the two and if the ratio were less than 1 we would see that $\{k$ successes$\}$ was the more likely case. Setting $\frac{P(B_{k+1})}{P(B_k)} \ge 1$ and solving for k yields the following sequence of equivalent inequalities:

$$\begin{aligned}
\frac{P(B_{k+1})}{P(B_k)} &\ge 1 \\
\frac{p}{1-p} \cdot \frac{n-k}{k+1} &\ge 1 \\
pn - pk &\ge k + 1 - pk - p \\
k &\le p(n+1) - 1.
\end{aligned}$$

In other words if k starts at 0 and begins to increase, the probability of achieving exactly k successes will increase while $k < p(n+1) - 1$ and then will decrease once $k > p(n+1) - 1$. As a consequence the most likely number of successes is the integer value of k for which

$k-1 \leq p(n+1)-1 < k$. This gives the critical value of $k = \lfloor p(n+1) \rfloor$, the greatest integer less than or equal to $p(n+1)$.

An unusual special case occurs if $p(n+1)$ is already an integer. Then the sequence of inequalities above is equality throughout, so if we let $k = \lfloor p(n+1) \rfloor = p(n+1)$ we find a ratio $P(B_k)/P(B_{k-1})$ exactly equal to 1. In this case $\{k-1 \text{ successes}\}$ and $\{k \text{ successes}\}$ share the distinction of being equally likely.

Ans (c) - **Geometric(p)**: It is possible we could see the first success as early as the first trial and, in fact, the probability of this occurring is just p, the probability that the first trial is a success. The probability of the first success coming on the k^{th} trial requires that the first $k-1$ trials be failures and the k^{th} trial be a success. Let A_i be the event $\{\text{the } i^{\text{th}} \text{ trial is a success}\}$ and let C_k be the event $\{\text{the first success occurs on the } k^{\text{th}} \text{ trial}\}$. So,

$$P(C_k) = P(A_1^c \cap A_2^c \cap \ldots \cap A_{k-1}^c \cap A_k).$$

As usual $P(A_i) = p$ and $P(A_i^c) = 1-p$, so by independence

$$P(C_k) = P(A_1^c)P(A_2^c)\ldots P(A_{k-1}^c)P(A_k) = (1-p)^{k-1}p$$

for $k > 0$. If we view these as probabilities of the outcomes of a sample space $\{1, 2, 3, \ldots\}$, we call this a geometric distribution with parameter p (or a Geometric(p) for short).

Ans (d) - **Average**: This is a natural question to ask but it requires a precise definition of what we mean by "average" in the context of probability. We shall do this in Chapter 4 and return to answer (d) at that point in time.

■

Bernoulli trials may also be used to determine probabilities associated with who will win a contest that requires a certain number of individual victories. Below is an example applied to a "best two out of three" situation.

Example 2.1.3. Jed and Sania play a tennis match. The match is won by the first player to win two sets. Sania is a bit better than Jed and she will win any given set with probability $\frac{2}{3}$. How likely is it that Sania will win the match? (Assume the results of each set are independent).

This can almost be viewed as three Bernoulli$(\frac{2}{3})$ trials where we view a success as a set won by Sania. One problem with that perspective is that an outcome such as (win,win,loss) never occurs since two wins put an end to the match and the third set will never be played. Nevertheless, the same tools used to solve the earlier problem can be used for this one as well. Sania wins the match if she wins the first two sets (which happens with probability

$\frac{4}{9}$). She also wins the match with either a (win,loss,win) or a (loss,win,win) sequence of sets, each of which has probability $\frac{4}{27}$ of occurring.

So the total probability of Sania winning the series is $\frac{4}{9} + \frac{4}{27} + \frac{4}{27} = \frac{20}{27}$.

Alternatively, it is possible to view this somewhat artificially as a genuine sequence of three Bernoulli($\frac{2}{3}$) trials where we pretend the players will play a third set even if the match is over by then. In effect the (win, win) scenario above is replaced by two different outcomes - (win, win, win) and (win, win, loss). Sania wins the match if she either wins all three sets (which has probability $\frac{8}{27}$) or if she wins exactly two of the three (which has probability $3 \cdot \frac{4}{27}$).

This perspective still leads us to the correct answer as $\frac{8}{27} + 3 \cdot \frac{4}{27} = \frac{20}{27}$. ■

2.1.1 *Using R to Compute Probabilities*

R can be used to compute probabilities of both the Binomial and Geometric distribution quite easily. We can compute them directly from the respective formulas. For example, with $n = 10$ and $p = 0.25$, all Binomial probabilities are given by

```
k <- 0:5
choose(5, k) * 0.25^k * 0.75^(5-k)
```

```
[1] 0.2373046875 0.3955078125 0.2636718750 0.0878906250 0.0146484375
[6] 0.0009765625
```

Similarly, the Geometric probabilities with $p = 0.25$ for $k = 0, 1, 2, \ldots, 10$ are given by

```
k <- 0:10
0.25 * 0.75^k
```

```
 [1] 0.25000000 0.18750000 0.14062500 0.10546875 0.07910156 0.05932617
 [7] 0.04449463 0.03337097 0.02502823 0.01877117 0.01407838
```

Actually, as both Binomial and Geometric are standard distributions, R has built-in functions to compute these probabilities. These can be used as follows.

```
dbinom(0:5, size = 5, prob = 0.25)
```

```
[1] 0.2373046875 0.3955078125 0.2636718750 0.0878906250 0.0146484375
[6] 0.0009765625
```

```
dgeom(0:10, prob = 0.25)
```

```
 [1] 0.25000000 0.18750000 0.14062500 0.10546875 0.07910156 0.05932617
 [7] 0.04449463 0.03337097 0.02502823 0.01877117 0.01407838
```

EXERCISES

Ex. 2.1.1. Three dice are rolled. How likely is it that exactly one of the dice shows a 6?

Ex. 2.1.2. A fair die is rolled repeatedly.

(a) What is the probability that the first 6 appears on the fifth roll?

(b) What is the probability that no 6's appear in the first four rolls?

(c) What is the probability that the second 6 appears on the fifth roll?

Ex. 2.1.3. Suppose that airplane engines operate independently in flight and fail with probability p $(0 \leq p \leq 1)$. A plane makes a safe flight if at least half of its engines are running. Kingfisher Air lines has a four–engine plane and Paramount Airlines has a two–engine plane for a flight from Bangalore to Delhi. Which airline has the higher probability for a successful flight?

Ex. 2.1.4. Two intramural volleyball teams have eight players each. There is a 10% chance that any given player will not show up to a game, independently of any another. The game can be played if each team has at least six members show up. How likely is it the game can be played?

Ex. 2.1.5. Mark is a 70% free throw shooter. Assume each attempted free throw is independent of every other attempt. If he attempts ten free throws, answer the following questions.

(a) How likely is it that Mark will make exactly seven of ten attempted free throws?

(b) What is the most likely number of free throws Mark will make?

(c) How do your answers to (a) and (b) change if Mark only attempts 9 free throws instead of 10?

Ex. 2.1.6. Continuing the previous exercise, Kalyani isn't as good a free throw shooter as Mark, but she can still make a shot 40% of the time. Mark and Kalyani play a game where the first one to sink a free throw is the winner. Since Kalyani isn't as skilled a player, she goes first to make it more fair.

(a) How likely is it that Kalyani will win the game on her first shot?

(b) How likely is it that Mark will win this game on his first shot? (Remember, for Mark even to get a chance to shoot, Kalyani must miss her first shot).

(c) How likely is it that Kalyani will win the game on her second shot?

(d) How likely is it that Kalyani will win the game?

Ex. 2.1.7. Recall from the text above that the R code

```
dbinom(0:5, size = 5, prob = 0.25)
```

```
[1] 0.2373046875 0.3955078125 0.2636718750 0.0878906250 0.0146484375
[6] 0.0009765625
```

produces a vector of six outputs corresponding to the probabilities that a Binomial$(5, 0.25)$ distribution takes on the six values 0-5. Specifically, the output indicates that the probability of the value 0 is approximately 0.2373046875, the probability of the value 1 is approximately 0.3955078125 and so on. In Example 2.1.2 we derived a formula for the most likely outcome of such a distribution. In the case of a Binomial$(5, 0.25)$ that formula gives the result $\lfloor (5+1)(0.25) \rfloor = 1$. We could have verified this via the R output above as well, since the second number on the list is the largest of the probabilities.

(a) Find the most likely outcome of a Binomial$(7, 0.34)$ distribution using the formula from example 2.1.2.

(b) Type an appropirate command into R to produce a vector of values corresponding to the probabilities that a Binomial$(7, 0.34)$ distribution takes on the possible values in its range. Use this list to verify your answer to part (a).

(c) Find the most likely outcome of a Binomial$(8, 0.34)$ distribution using the formula from Example 2.1.2.

(d) Type an appropirate command into R to produce a vector of values corresponding to the probabilities that a Binomial$(8, 0.34)$ distribution takes on the possible values in its range. Use this list to verify your answer to part (c).

Ex. 2.1.8. It is estimated that 0.8% of a large shipment of eggs to a certain supermarket are cracked. The eggs are packaged in cartons, each with a dozen eggs, with the cracked eggs being randomly distributed. A restaurant owner buys 10 cartons from the supermarket. Call a carton "defective" if it contains at least one cracked egg.

(a) If she notes the number of defective cartons, what are the possible outcomes for this experiment?

(b) If she notes the total number of cracked eggs, what are the possible outcomes for this experiment?

(c) How likely is it that she will find exactly one cracked egg among all of her cartons?

(d) How likely is it that she will find exactly one defective carton?

(e) Explain why your answer to (d) is close to, but slightly larger than, than your answer to (c).

(f) What is the most likely number of cracked eggs she will find among her cartons?

(g) What is the most likely number of defective cartons she will find?

(h) How do you reconcile your answers to parts (g) and (h)?

Ex. 2.1.9. Steve and Siva enter a bar with \$30 each. A round of drinks cost \$10. For each round, they roll a die. If the roll is even, Steve pays for the round and if the roll is odd, Siva pays for it. This continues until one of them runs out of money.

(a) What is the Probability that Siva runs out of money?

(b) What is the Probability that Siva runs out of money if Steve has cheated by bringing a die that comes up even only 40% of the time?

Ex. 2.1.10. Let $0 < p < 1$. Show that the mode of a Geometric(p) distribution is 1.

Ex. 2.1.11. Scott is playing a game where he rolls a standard die until it shows a 6. The number of rolls needed therefore has a Geometric($\frac{1}{6}$) distribution. Use the appropriate R commands to do the following:

(a) Produce a vector of values for $j = 1, \ldots, 6$ corresponding to the probabilities that it will take Scott j rolls before he observes a 6.

(b) Scott figures that since each roll has a $\frac{1}{6}$ probability of producing a 6, he's bound to get that result at some point after six rolls. Use the results from part (a) to determine the probability that Scott's expectations are met and a 6 will show up in one his first six rolls.

Ex. 2.1.12. Suppose a fair coin is tossed n times. Compute the following:

(a) $P(\{4 \text{ heads occur }\}|\{3 \text{ or } 4 \text{ heads occur}\})$;

(b) $P(\{k-1 \text{ heads occur}\}|\{k-1 \text{ or } k \text{ heads occur}\})$; and

(c) $P(\{k \text{ heads occur}\}|\{k-1 \text{ or } k \text{ heads occur}\})$.

Ex. 2.1.13. At a basketball tournament, each round is on a "best of seven games" basis. That is, Team I and Team 2 play until one of the teams has won four games. Suppose each game is won by Team I with probability p, independently of all previous games. Are the events $A = \{\text{Team I wins the round}\}$ and $B = \{\text{the round lasts exactly four games}\}$ independent?

Ex. 2.1.14. Two coins are sitting on a table. One is fair and the other is weighted so that it always comes up heads.

(a) If one coin is selected at random (each equally likely) and flipped, what is the probability the result is heads?

(b) One coin is selected at random (each equally likely) and flipped five times. Each flip shows heads. Given this information about the coin flip results, what is the conditional probability that the selected coin was the fair one?

Ex. 2.1.15. For $0 < p < 1$ we defined the geometric distribution as a probability on the set $\{1, 2, 3, \dots\}$ for which $P(\{k\}) = p(1-p)^{k-1}$. Show that these outcomes account for all possibilities by demonstrating that $\sum_{k=1}^{\infty} P(\{k\}) = 1$.

Ex. 2.1.16. The geometric distribution described the waiting time to observe a single success. A "Negative Binomial" distribution with parameters n and p (NegBinomial(n, p)) is defined the number of Bernoulli(p) trials needed before observing n successes. The following problem builds toward calculating some associated probabilities.

(a) If a fair die is rolled repeatedly and a number is recorded equal to the number of rolls until the second 6 is observed, what is the sample space of possible outcomes for this experiment?

(b) For k in the sample space you identified in part (a), what is $P(\{k\})$?

(c) If a fair die is rolled repeatedly and a number is recorded equal to the number of rolls until the n^{th} 6 is observed, what is the sample space of possible outcomes for this experiment?

(d) For k in the sample space you identified in part (c), what is $P(\{k\})$?

(e) If a sequence of Bernoulli(p) trials (with $0 < p < 1$) is performed and a number is recorded equal to the number of trials until the n^{th} success is observed, what is the sample space of possible outcomes for this experiment?

(f) For k in the sample space you identified in part (e), what is $P(\{k\})$?

(g) Show that you have accounted for all possibilities in part (f) by showing

$$\sum_{k \in S} P(\{k\}) = 1.$$

2.2 POISSON APPROXIMATION

Calculating Binomial probabilities can be challenging when n is large. Let us consider the following example:

Example 2.2.1. A small college has 1460 students. Assume that birthrates are constant throughout the year and that each year has 365 days. What is the probability that five or more students were born on Independence day?

The probability that any given student was born on Independence day is $\frac{1}{365}$. So the exact probability is

$$1 - \sum_{k=0}^{4} \binom{1460}{k} \left(\frac{1}{365}\right)^k \left(\frac{364}{365}\right)^{1460-k}.$$

Repeatedly dealing with large powers of fractions or large combinatorial computations is not so easy, so it would be convenient to find a faster way to estimate such a probability. ■

The example above can be thought of as a series of Bernoulli trials where a success means finding a student whose birthday is Independence day. In this case p is small ($\frac{1}{365}$) and n is large (1460). To approximate we will consider a limiting procedure where $p \to 0$ and $n \to \infty$, but with limits carried out in such a way that np is held constant. The computation below is called a Poisson approximation.

Theorem 2.2.2. *Let $\lambda > 0$, $k \geq 1$, $n \geq \lambda$ and $p = \frac{\lambda}{n}$. Defining A_k as*

$$A_k = \{k \text{ successes in } n \text{ Bernoulli}(p) \text{ Trials}\},$$

it then follows that

$$\lim_{n \to \infty} P(A_k) = \frac{e^{-\lambda} \lambda^k}{k!}. \tag{2.2.1}$$

Proof -

$$\begin{aligned} P(A_k) &= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n(n-1)\ldots(n-k+1)}{k!}\frac{\lambda^k}{n^k}\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!}\frac{n(n-1)\ldots(n-k+1)}{n^k}\left(1-\frac{\lambda}{n}\right)^{n}\left(1-\frac{\lambda}{n}\right)^{-k} \\ &= \frac{\lambda^k}{k!}1(1-\frac{1}{n})\ldots(1-\frac{k-1}{n})\left(1-\frac{\lambda}{n}\right)^{n}\left(1-\frac{\lambda}{n}\right)^{-k} \\ &= \frac{\lambda^k}{k!}\prod_{r=1}^{k-1}\left(1-\frac{r}{n}\right)\left(1-\frac{\lambda}{n}\right)^{n}\left(1-\frac{\lambda}{n}\right)^{-k}. \end{aligned}$$

Standard limit results imply that

$$\begin{aligned} \lim_{n\to\infty}(1-\frac{r}{n}) = 1 \quad &\text{for all} \quad r \geq 1; \\ \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-k} = 1 \quad &\text{for all} \quad \lambda \geq 0, k \geq 1; \ \textit{and} \\ \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n} = e^{-\lambda} \quad &\text{for all} \quad \lambda \geq 0. \end{aligned}$$

As $P(A_k)$ is a finite product of such expressions, the result is now immediate using the properties of limits. ■

Returning to Example 2.2.1 and using the above approximation, we would take $\lambda = pn = \frac{1460}{365} = 4$. So if E is the event {five or more Independence day birthdays},

$$\begin{aligned} P(E) &= 1-\sum_{k=0}^{4}\binom{1460}{k}\left(\frac{1}{365}\right)^k\left(\frac{364}{365}\right)^{1460-k} \\ &\approx 1-\left[e^{-4}+4e^{-4}+\frac{4^2}{2}e^{-4}+\frac{4^3}{6}e^{-4}+\frac{4^4}{24}e^{-4}\right]. \end{aligned}$$

Calculation demonstrates this is a good approximation. To seven digits of accuracy, the correct value is 0.37116294 while the Poisson approximation gives an answer of 0.37116306. These can be obtained using R as follows:

```
1 - sum(dbinom(0:4, size = 1460, prob = 1/365))
```

```
[1] 0.3711629
```

```
lambda <- 1460 / 365
1 - sum(exp(-lambda) * lambda^(0:4) / factorial(0:4))
```

```
[1] 0.3711631
```

It also turns out that the right hand side of (2.2.1) defines a probability on the sample space of non-negative integers. The distribution is named after Siméon Poisson (1781–1840).

Poisson (λ): Let $\lambda \geq 0$ and $S = \{0, 1, 2, 3, \ldots\}$ with probability P given by

$$P(\{k\}) = \frac{e^{-\lambda}\lambda^k}{k!}$$

for $k \in S$. This distribution is called Poisson with parameter λ (or Poisson(λ) for short).

As with Binomial and Geometric, R has a built-in function to evaluate Poisson probabilities as well. An alternative to the calculation above is the following.

```
1 - sum(dpois(0:4, lambda = 1460 / 365))
```

```
[1] 0.3711631
```

It is important to note that for this approximation to work well, p must be small and n must be large. For example, we may modify our question as follows:

Example 2.2.3. A class has 48 students. Assume that birthrates are constant throughout the year and that each year has 365 days. What is the probability that five or more students were born in September? ■

The correct answer to this question is

```
1 - sum(dbinom(0:4, size = 48, prob = 1/12))
```

```
[1] 0.3710398
```

However, the Poisson approximation remains unchanged at 0.3711631, because $np = 48/12 = 1460/365 = 4$, and only matches the correct answer up to 3 digits rather than 6. Figure 2.2 shows a point-by-point approximation of both Binomial distributions by Poisson.

At this point we have defined many named distributions. Frequently a problem will require the use of more than one of these as evidenced in the next example.

Example 2.2.4. A computer transmits three digital messages of 12 million bits of information each. Each bit has a probability of one one-billionth that it will be incorrectly received, independent of all other bits. What is the probability that at least two of the of the three messages will be received error free?

Since $n = 12,000,000$ is large and since $p = \frac{1}{1,000,000,000}$ is small it is appropriate to use a Poisson approximation where $\lambda = np = 0.012$. A message is error free if there isn't a single misread bit, so the probability that a given message will be received without an error is $e^{-0.012}$.

Now we can think of each message being like a Bernoulli trial with probability $e^{-0.012}$, so the number of messages correctly received is then like a $Binomial(3, e^{-0.012})$. Therefore the probability of receiving at least two error-free messages is

$$\binom{3}{3}(e^{-0.012})^3(1-e^{-0.012})^0 + \binom{3}{2}(e^{-0.012})^2(1-e^{-0.012})^1 \approx 0.9996.$$

There is about a 99.96% chance that at least two of the messages will be correctly received. ■

EXERCISES

Ex. 2.2.1. Do the problems below to familiarize yourself with the "sum" command in R.

(a) If a fair coin is tossed 100 times, what is the probability exactly 55 of the tosses show heads?

(b) Example 2.2.3 showed how to use R to add the probabilities of a range of outcomes for common distributions. Use this code as a guide to calculate the probability at least 55 tosses show heads.

Ex. 2.2.2. Consider an experiment described by a Poisson($\frac{1}{2}$) distribution and answer the following questions.

(a) What is the probability the experiment will produce a result of 0?

(b) What is the probability the experiment will produce a result larger than 1?

Ex. 2.2.3. Suppose we perform 500 independent trials with probability of success being 0.02.

(a) Use R to compute the probability that there are six or fewer successes. Obtain a decimal approximation accurate to five decimal places.

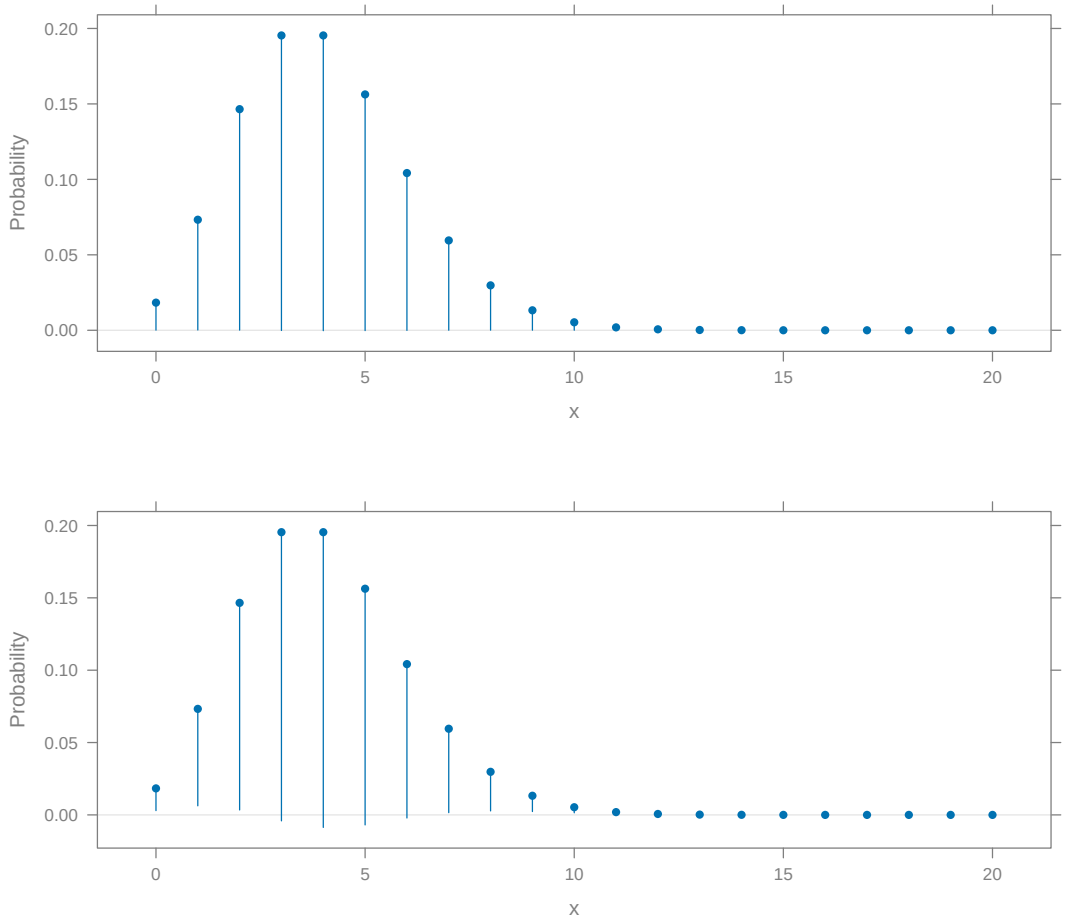


Figure 2.2: The Poisson approximation to the Binomial distribution. In both plots above, the points indicate Binomial probabilities for $k = 0, 1, 2, \ldots, 20$; the top plot for Binomial$(1460, \frac{1}{365})$, and the bottom for Binomial$(48, \frac{1}{12})$. The lengths of the vertical lines, "hanging" from the points, represent the corresponding probabilities for Poisson(4). For a good approximation, the bottom of the hanging lines should end up at the x-axis. As we can see, this happens in the top plot but not for the bottom plot, indicating that Poisson(4) is a good approximation for the first Binomial distribution, but not as good for the second.

(b) Use the Poisson approximation to estimate the probability that there are six or fewer successes and compare it to your answer to (a).

Now suppose we perform 5000 independent trials with probability of success being 0.002.

(c) Use R to compute the probability that there are six or fewer successes. Obtain a decimal approximation accurate to five decimal places.

(d) Use the Poisson approximation to estimate the probability that there are six or fewer successes and compare it to your answer to (c).

(e) Which approximation (b) or (d) is more accurate? Why?

Ex. 2.2.4. For a certain daily lottery, the probability is $\frac{1}{10000}$ that you will win. Suppose you play this lottery every day for three years. Use the Poisson approximation to estimate the chance that you will win more than once.

Ex. 2.2.5. A book has 200 pages. The number of mistakes on each page has a Poisson(1) distribution, and is independent of the number of mistakes on all other pages.

(a) What is the chance that there are at least 2 mistakes on the first page?

(b) What is the chance that at least eight of the first ten pages are free of mistakes?

Ex. 2.2.6. Let $\lambda > 0$. For the problems below, assume the probability space is a Poisson(λ) distribution.

(a) Let k be a non-negative integer. Calculate the ratio $\frac{P(\{k+1\})}{P(\{k\})}$.

(b) Use (a) to calculate the mode of a Poisson(λ).

Ex. 2.2.7. A number is to be produced as follows. A fair coin is tossed. If the coin comes up heads the number will be the outcome of an experiment corresponding to a Poisson(1) distribution. If the coin comes up tails the number will be the outcome of an experiment corresponding to a Poisson(2) distribution. Given that the number produced was a 2, determine the conditional probability that the coin came up heads.

Ex. 2.2.8. Suppose that the number of earthquakes that occur in a year in California has a Poisson distribution with parameter λ. Suppose that the probability that any given earthquake has magnitude at least 6 on the Richter scale is p.

(a) Given that there are exactly n earthquakes in a year, find an expression (in terms of n and p) for the conditional probability that exactly one of them is magnitude at least 6.

(b) Find an expression (in terms of λ and p) for the probability that there will be exactly one earthquake of magnitude at least 6 in a year.

(c) Find an expression (in terms of n, λ, and p) for the probability that there will be exactly k earthquakes of magnitude at least 6 in a year.

Ex. 2.2.9. We defined a Poisson distribution as a probability on $S = \{0, 1, 2, \dots\}$ for which

$$P(\{k\}) = \frac{e^{-\lambda}\lambda^k}{k!},$$

for $k \geq 1$. Prove that this completely accounts for all possibilities by proving that

$$\sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = 1.$$

(Hint: Consider the power series expansion of the exponential function).

Ex. 2.2.10. Consider n vertices labeled $\{1, 2, \dots, n\}$. Corresponding to each distinct pair $\{i, j\}$ we perform an independent Bernoulli (p) experiment and insert an edge between i and j with probability p. The graph constructed this way is denoted as $G(n, p)$.

(a) Let $1 \leq i \leq n$. We say j is a neighbour of i if there is an edge between i and j. For some $1 \leq k \leq n$ determine the probability that i has k neighbours ?

(b) Let $\lambda > 0$ and n large enough so that $0 < p = \frac{\lambda}{n} < 1$ and let $A_k = \{$ vertex 1 has k neighbours$\}$ what is the

$$\lim_{n\to\infty} P(A_k)?$$

2.3 SAMPLING WITH AND WITHOUT REPLACEMENT

Imagine a small town with 5000 residents, exactly 1000 of whom are under the age of eighteen. Suppose we randomly select four of these residents and ask how many of the four are under the age of eighteen. There is some ambiguity in how to interpret this idea of selecting four residents. One possibility is "sampling with replacement" where each selection could be any of the 5000 residents and the selections are all genuinely independent. With this interpretation, the sample is simply a series of four independent Bernoulli$(\frac{1}{5})$ trials, in which case the answer may be found using techniques from the previous sections. Note, however, that the assumption of independence allows for the possibility that the same individual will be chosen two or more times in separate trials. This is a situation that might seem peculiar when we think about choosing four people from a population of 5000,

since we may not have four different individuals at the end of the process. To eliminate this possibility consider "sampling without replacement" where it is assumed that if an individual is chosen for inclusion in the sample, that person is no longer available to be picked in a later selection. Equivalently we can consider all possible groups of four which might be selected and view each grouping as equally likely. This change means the problem can no longer be solved by viewing the situation as a series of independent Bernoulli trials. Nevertheless, other tools that have been previously developed will serve to answer this new problem.

Example 2.3.1. For the town described above, what is the probability that, of four residents randomly selected (without replacement), exactly two of them will be under the age of eighteen?

Since we are selecting four residents from the town of 5000, there are $\binom{5000}{4}$ ways this may be done. If each of these is equally likely, the desired probability may be calculated by determining how many of these selections result in exactly two people under the age of eighteen. This requires selecting two of the 1000 who are in that younger age group and also selecting two of the 4000 who are older. So there are $\binom{1000}{2}\binom{4000}{2}$ ways to make such choices and therefore the probability of selecting exactly two residents under age eighteen is $\binom{1000}{2}\binom{4000}{2}/\binom{5000}{4}$.

It is instructive to compare this to the solution if it is assumed the selection is done with replacement. In that case, the answer is the simply the probability that a Binomial$(4, \frac{1}{5})$ produces a result of two. From the previous sections, the answer is $\binom{4}{2}(\frac{1}{5})^2(\frac{4}{5})^2$.

To compare these answers we give decimal approximations of both. To six digits of accuracy

$$\frac{\binom{1000}{2}\binom{4000}{2}}{\binom{5000}{4}} \approx 0.153592 \quad \text{and} \quad \binom{4}{2}\left(\frac{1}{5}\right)^2\left(\frac{4}{5}\right)^2 = 0.1536,$$

so while the two answers are not equal, they are very close. This is a reflection of an important fact in statistical analysis—when samples are small relative to the size of the populations they came from, the two methods of sampling give very similar results. ■

2.3.1 *The Hypergeometric Distribution*

Analyzing such problems more generally, consider a population of N people. Suppose r of these N share a common characteristic and the remaining $N - r$ do not have this

characteristic. We take a sample of size m (without replacement) from the population and count the number of people among the sample that have the specified characteristic. This experiment is described by probabilities known as a hypergeometric distribution. Notice that the largest possible result is $\min\{m, r\}$ since the number cannot be larger than the size of the sample nor can it be larger than the number of people in the population with the characteristic. On the other extreme, it may be that the sample is so large it is guaranteed to select some people with the characteristic simply because the number of people without has been exhausted. More precisely, for every selection over $N - r$ in the sample we are guaranteed to select at least one person who has the characteristic. So the minimum possible result is the larger of 0 or $(m - (N - r))$.

HyperGeo(N, r, m): Let r and m be non-negative integers and let N be an integer with $N > \max\{r, m\}$. Let S be the set of integers ranging from $\max\{0, m - (N - r)\}$ to $\min\{m, r\}$ inclusive with probability P given by

$$P(\{k\}) = \frac{\binom{r}{k}\binom{N-r}{m-k}}{\binom{N}{m}}$$

for $k \in S$. Such a distribution is called hypergeometric with parameters N, r, and m (or HyperGeo(N, r, m)).

Of course, R can be used to compute hypergeometric probabilities as well. Example 2.3.1 can be phrased in terms of a HyperGeo$(5000, 1000, 4)$ distribution, with $P(\{k\})$ being the desired answer. This probability can be computed as:

```
dhyper(2, 4000, 1000, 4)
```

```
[1] 0.1535923
```

Note however, that instead of N, the parameter used by R is $N - r$.

2.3.2 *Hypergeometric Distribution as a Series of Dependent Trials*

It is also possible (and useful) to view sampling without replacement as a series of dependent Bernoulli trials for which each trial reduces the possible outcomes of subsequent trials. In this case each trial is described in terms of conditional probabilities based on the results of the preceding observations. We illustrate this by revisiting the previous example.

Example 2.3.1 Continued: We first solved this problem by considering every group of four as equally likely to be selected. Now consider the sampling procedure as a series of four separate Bernoulli trials where a success corresponds to the selection of a person under eighteen and a failure as the selection of someone older. We still want to determine the probability that a sample of size four will produce exactly two successes. One complication with this perspective is that the successes and failures could come in many different orders, so first consider the event where the series of selections follow the pattern "success-success-failure-failure". More precisely, for $j = 1, 2, 3, 4$ let

$$A_j = \{\text{The } j^{\text{th}} \text{ selection is a person younger than eighteen}\}.$$

Clearly $P(A_1) = \frac{1000}{5000}$. Given that the first selection is someone under eighteen, there are now only 4999 people remaining to choose among, and only 999 of them are under eighteen. Therefore $P(A_2|A_1) = \frac{999}{4999}$. Continuing with that same reasoning,

$$P(A_3^c|A_1 \cap A_2) = \frac{4000}{4998}$$

and

$$P(A_4^c|A_1 \cap A_2 \cap A_3^c) = \frac{3999}{4997}.$$

From those values, Theorem 1.3.8 may be used to calculate

$$\begin{aligned} P(success - success &- failure - failure) = P(A_1 \cap A_2 \cap A_3^c \cap A_4^c) \\ &= \frac{1000}{5000} \cdot \frac{999}{4999} \cdot \frac{4000}{4998} \cdot \frac{3999}{4997}. \end{aligned}$$

Next we must account for the fact that this figure only considers the case where the two younger people were chosen as the first two selections. There are $\binom{4}{2}$ different orderings that result in two younger and two older people, and it happens that each of these has the same probability calculated above. For example,

$$\begin{aligned} P(failure - success &- success - failure) = P(A_1^c \cap A_2 \cap A_3 \cap A_4^c) \\ &= \frac{4000}{5000} \cdot \frac{1000}{4999} \cdot \frac{999}{4998} \cdot \frac{3999}{4997}. \end{aligned}$$

The individual fractions are different, but their product is the same. This will always happen for different orderings of a specific number of successes since the denominators (5000 through 4997) reflect the steady reduction of one available choice with each additional selection. Similarly the numerators (1000 and 999 together with 4000 and 3999) reflect the number of people available from each of the two different categories and their reduction

as previous choices eliminate possible candidates. Therefore the total probability is the product of the number of orderings and the probability of each ordering.

$$P(\text{two under eighteen}) = \binom{4}{2} \cdot \frac{4000}{5000} \cdot \frac{1000}{4999} \cdot \frac{999}{4998} \cdot \frac{3999}{4997}.$$

We leave it to the reader to verify that this is equal to $\binom{1000}{2}\binom{4000}{2}/\binom{5000}{4}$, the answer we found when we originally solved the problem via a different method.

The following theorem generalizes this previous example.

Theorem 2.3.2. *Let S be a sample space with a hypergeometric distribution with parameters N, r, and m. Then $P(\{k\})$ equals*

$$\binom{m}{k}\left[\frac{r}{N}\frac{r-1}{N-1}\cdots\frac{r-(k-1)}{N-(k-1)}\right]\left[\frac{N-r}{N-k}\frac{N-r-1}{N-k-1}\cdots\frac{N-r-(m-1-k)}{N-(m-1)}\right]$$

for any $k \in S$.

Proof. Following the previous example as a model, this can be proven by viewing the hypergeometric distribution as a series of dependent trials. The first k fractions are the probabilities the first k trials each result in successes conditioned on the successes of the preceding trials. The remaining $m-k$ fractions are the conditional probabilities the remaining trials result in failures. The leading factor of $\binom{m}{k}$ accounts for the number of different patterns of k successes and $m-k$ failures, each of which is equally likely. It is also possible to prove the equality directly using combinatorial identities and we leave this as Exercise 2.3.4. ■

2.3.3 *Binomial Approximation to the Hypergeometric Distribution*

We saw with Example 2.3.1 that sampling with and without replacement may give very similar results. The following theorem makes a precise statement to this effect.

Theorem 2.3.3. *Let N, m, and r be positive integers for which $m < r < N$ and let k be a positive integer between 0 and m. Define*

$$p = \frac{r}{N}, \quad p_1 = \frac{r-k}{N-k}, \quad \text{and} \quad p_2 = \frac{r-k}{N-m}.$$

Letting H denote the probability that a hypergeometric distribution with parameters N, r, and m takes on the value k, the following inequalities give bounds on this probability:

$$\binom{m}{k} p_1^k (1-p_2)^{m-k} < H \leq \binom{m}{k} p^k (1-p_1)^{m-k}.$$

Proof- The inequalities may be verified by comparing p, p_1, and p_2 to the fractions from Theorem 2.3.2. Specifically note that the k fractions

$$\frac{r}{N}, \frac{r-1}{N-1}, \ldots, \frac{r-(k-1)}{N-(k-1)}$$

are all less than or equal to p. Likewise the $m-k$ fractions

$$\frac{N-r}{N-k}, \frac{N-r-1}{N-k-1}, \ldots, \frac{N-r-(m-1-k)}{N-(m-1)}$$

are all less than or equal to $\frac{N-r}{N-k}$ which itself equals $1-p_1$. Combining these facts proves the right hand inequality. The left hand inequality may be similarly shown by noting that the fractions

$$\frac{r}{N}, \frac{r-1}{N-1}, \ldots, \frac{r-(k-1)}{N-(k-1)}$$

are all greater than p_1 while the fractions

$$\frac{N-r}{N-k}, \frac{N-r-1}{N-k-1}, \ldots, \frac{N-r-(m-1-k)}{N-(m-1)}$$

all exceed $\frac{N-r-(m-k)}{N-m}$ which equals $1-p_2$. ■

When m is small relative to r and N, both fractions p_1 and p_2 are approximately equal to p. So this theorem justifies the earlier statement that sampling with and without replacement yield similar results when samples are small relative to the populations from which they were derived.

EXERCISES

Ex. 2.3.1. Suppose there are thirty balls in an urn, ten of which are black and the remaining twenty of which are red. Suppose three balls are selected from the urn (without replacement).

(a) What is the probability that the sequence of draws is red-red-black?

(b) What is the probability that the three draws result in exactly two red balls?

Ex. 2.3.2. This exercise explores how to use R to investigate the Binomial approximation to the Hypergeometric distribution.

(a) A jar contains forty marbles – thirty white and ten black. Ten marbles are drawn at random from the jar. Use R to calculate the probability that exactly five of the marbles drawn are black. Do two separate computations, one under the assumption that the draws are with replacement and the other under the assumption that the draws are without replacement.

(b) Repeat part (a) except now assume the jar contains 400 marbles – 300 wihite and 100 black.

(c) Repeat part (a) excpet now assume the jar contains 4000 marbles – 3000 white and 1000 black.

(d) Explain what you are observing with your results of parts (a), (b), and (c).

Ex. 2.3.3. Consider a room of one hundred people – forty men and sixty women.

(a) If ten people are selected from the room, find the probability that exactly six are women. Calculate this probability with and without replacement and compare the decimal approximations of your two results.

(b) If ten people are selected from the room, find the probability that exactly seven are women. Calculate this probability with and without replacement and compare the decimal approximations of your two results.

(c) If 100 people are selected from the room, find the probability that exactly sixty are women. Calculate this probability with and without replacement and compare the two answers.

(d) If 100 people are selected from the room, find the probability that exactly sixty-one are women. Calculate this probability with and without replacement and compare the two answers.

Ex. 2.3.4. Use the steps below to prove Theorem 2.3.2

(a) Prove that $\frac{r!(N-k)!}{N!(r-k)!}$ equals

$$\frac{r}{N} \cdot \frac{r-1}{N-1} \cdots \frac{r-(k-1)}{N-(k-1)}.$$

(b) Prove that $\frac{(N-r)!(N-m)!}{(N-k)!((N-r-(m-k))!}$ equals

$$\frac{N-r}{N-k} \cdot \frac{N-r-1}{N-k-1} \cdots \frac{N-r-(m-1-k)}{N-(m-1)}.$$

(c) Use (a) and (b) to prove Theorem 2.3.2.

Ex. 2.3.5. A box contains W white balls and B black balls. A sample of n balls is drawn at random for some $n \leq \min(W, B)$. For $j = 1, 2, \cdots, n$, let A_j denote the event that the ball drawn on the j^{th} draw is white. Let B_k denote the event that the sample of n balls contains exactly k white balls.

(a) Find $P(A_j|B_k)$ if the sample is drawn with replacement.

(b) Find $P(A_j|B_k)$ if the sample is drawn without replacement.

Ex. 2.3.6. For the problems below, assume a HyperGeo(N, r, m) distribution.

(a) Calculate the ratio $\frac{P(\{k+1\})}{P(\{k\})}$.
(Assume that $max\{0, m-(N-r)\} \leq k \leq min\{r, m\}$ to avoid zero in the denominator).

(b) Use (a) to calculate the mode of a HyperGeo(N, r, m).

Ex. 2.3.7. Biologists use a technique called "capture-recapture" to estimate the size of the population of a species that cannot be directly counted. The following exercise illustrates the role a hypergeometric distribution plays in such an estimate.

Suppose there is a species of unknown population size N. Suppose fifty members of the species are selected and given an identifying mark. Sometime later a sample of size twenty is taken from the population and it is found that four of the twenty were previously marked. The basic idea behind mark-recapture is that since the sample showed $\frac{4}{20} = 20\%$ marked members, that should also be a good estimate for the fraction of marked members of the species as a whole. However, for the whole species that fraction is $\frac{50}{N}$ which provides a population estimate of $N \approx 250$.

Looking more deeply at the problem, if the second sample is assumed to be done at random without replacement and with each member of the population equally likely to be selected, the resulting number of marked members should follow a HyperGeo$(N, 50, 20)$ distribution.

Under these assumptions use the formula for the mode calculated in the previous exercise to determine which values of N would cause a result of four marked members to be the most likely of the possible outcomes.

Ex. 2.3.8. The geometric distribution was first developed to determine the number of independent Bernoulli trials needed to observe the first success. When viewing the hypergeometric distribution as a series of dependent trials, the same question may be asked. Suppose we have a population of N people for which r have a certain characteristic and the remaining $N - r$ do not have that characteristic. Suppose an experiment consists of sampling (without replacement) repeatedly and recording the number of the sample that first corresponds to selecting someone with the specified characteristic. Answer the questions below.

(a) What is S, the list of possible outcomes of this experiment?

(b) For each $k \in S$, what is $P(\{k\})$?

(c) Define $p = \frac{r}{N}$ and $p_1 = \frac{r-(k-1)}{N-(k-1)}$. Using the result from (b) prove the following bounds on the probability distribution:

$$p(1-p_1)^{k-1} \leq P(\{k\}) \leq p_1(1-p)^{k-1}$$

(As a consequence, when k is much smaller than r and N, the values of p_1 and p are approximately equal and the probabilities from (b) are closely approximated by a geometric distribution).

3

DISCRETE RANDOM VARIABLES

In the previous chapter many different distributions were developed out of Bernoulli trials. In that chapter we proceeded by creating new sample spaces for each new distribution, but when faced with many questions related to the same basic framework, it is usually clearer to maintain a single sample space and to define functions on that space whose outputs relate them to questions under consideration. Such functions are known as "random variables" and they will be the focus of this chapter.

3.1 RANDOM VARIABLES AS FUNCTIONS

Example 3.1.1. Suppose a coin is flipped three times. Consider the probabilities associated with the following two questions:

(a) How many coins will come up heads?

(b) Which will be the first flip (if any) that shows heads?

At this point the answers to these questions should be easy to determine, but the purpose of this example is to emphasize how functions could be used to answer *both* questions within the context of a single sample space. Let S be a listing of all eight possible orderings of heads and tails on the three flips, so that $S = \{hhh, hht, hth, htt, thh, tht, tth, ttt\}$. Now define two functions on S. Let X be the function that describes the total number of heads among the three flips and let Y be the function that describes the first flip that produces heads. Then X and Y are given by the table

ω	$X(\omega)$	$Y(\omega)$
hhh	3	1
hht	2	1
hth	2	1
htt	1	1
thh	2	2
tht	1	2
tth	1	3
ttt	0	none

where $Y(ttt)$ is defined as "none" as there is no first time the coin produces heads.

Suppose we want to know the probability that exactly two of the three coins will be heads. The relevant event is $E = \{hht, hth, thh\}$, but in the pre-image notation of function theory this set may also be described as $X^{-1}(\{2\})$, the elements of S for which X produces an output of 2. This allows us to describe the probability of the event as:

$$P(\text{two heads}) = P(X^{-1}(\{2\})) = P(\{hht, hth, thh\}) = \frac{3}{8}.$$

Rather than use the standard pre-image notation, it is more common in probability to write $(X = 2)$ for the set $X^{-1}(\{2\})$ as this emphasizes that we are considering outcomes for which the function X equals 2.

Similarly, if we wanted to know the probability that the first result of heads showed up on the third flip, that is a question that involves the function Y. Using the notation $(Y = 3)$ in place of $Y^{-1}(\{3\})$ the probability may be calculated as

$$P(\text{first heads on flip three}) = P(Y = 3) = P(\{tth\}) = \frac{1}{8}.$$

As above we can compute the

$$P(X = 0) = \frac{1}{8}, P(X = 1) = \frac{3}{8}, \text{ and } P(X = 3) = \frac{1}{8}$$

and

$$P(Y = 1) = \frac{1}{2}, P(Y = 2) = \frac{1}{4}, \text{ and } P(Y = \text{none}) = \frac{1}{8},$$

thus giving a complete description of how X and Y distribute the probabilities onto their range. For both cases only a single sample space was needed. Two different questions were approached by defining two different functions on that sample space. ■

The following theorem explains how the mechanism of the previous example may be more generally applied.

Theorem 3.1.2. *Let S be a sample space with probability P and let $X : S \to T$ be a function. Then X generates a probability Q on T given by*

$$Q(B) = P(X^{-1}(B)).$$

The probability Q is called the "distribution of X" as it describes how X distributes the probability from S onto T. The proof relies on two set-theoretic facts that we will take as given. The first is that $X^{-1}(\bigcup_{i=1}^{\infty} B_i) = \bigcup_{i=1}^{\infty} X^{-1}(B_i)$ and the second is the fact that if B_i and B_j are disjoint, then so are $X^{-1}(B_i)$ and $X^{-1}(B_j)$.

Proof. Let $B \subset T$. As P is known to be a probability, $0 \leq P(X^{-1}(B)) \leq 1$, and so Q maps subsets of T into $[0,1]$. As X is a function into T, we know $X^{-1}(T) = S$. Therefore $Q(T) = P(X^{-1}(T)) = P(S) = 1$ and Q satisfies the first probability axiom.

To show that Q satisfies the second axiom, suppose $B_1, B_2, \ldots$ are a countable collection of disjoint subsets of T.

$$\begin{aligned} Q(\bigcup_{i=1}^{\infty} B_i) &= P(X^{-1}(\bigcup_{i=1}^{\infty} B_i)) \\ &= P(\bigcup_{i=1}^{\infty} X^{-1}(B_i)) \\ &= \sum_{i=1}^{\infty} P(X^{-1}(B_i)) \\ &= \sum_{i=1}^{\infty} Q(B_i). \end{aligned}$$

■

As in the previous example, it is typical to write $(X \in B)$ in place of the notation $X^{-1}(B)$ to emphasize the fact that we are computing the probability that the function X takes a value in the set B. In practice, the new probability Q would rarely be used explicity, but would be calculated in terms of the original probability P via the relationship described in the theorem.

EXAMPLE 3.1.3. A board game has a wheel that is to be spun periodically. The wheel can stop in one of ten equally likely spots. Four of these spots are red, three are blue, two are green, and one is black. Let X denote the color of the spot. Determine the distribution of X.

The function X is defined on a sample space S that consists of the ten spots the wheel could stop, and it takes values on the set of colors $T = \{red, blue, green, black\}$. Its distribution is a probability Q on the set of colors which can be determined by calculating the probability of each color.

For instance $Q(\{red\}) = P(X = red) = P(X^{-1}(\{red\})) = \frac{4}{10}$ as four of the ten spots on the wheel are red and all spots are equally likely. Similarly,

$$\begin{aligned} Q(\{blue\}) &= P(X = blue) = \frac{3}{10} \\ Q(\{green\}) &= P(X = green) = \frac{2}{10} \\ Q(\{black\}) &= P(X = black) = \frac{1}{10} \end{aligned}$$

completing the description of the distribution. ■

Example 3.1.4. For a certain lottery, a three-digit number is randomly selected (from 000 to 999). If a ticket matches the number exactly, it is worth \$200. If the ticket matches exactly two of the three digits, it is worth \$20. Otherwise it is worth nothing. Let X be the value of the ticket. Find the distribution of X.

The function X is defined on $S = \{000, 001, \ldots, 998, 999\}$ - the set of all one thousand possible three digit numbers. The function takes values on the set $\{0, 20, 200\}$, so the distribution Q is a probability on $T = \{0, 20, 200\}$.

First, $Q(\{200\}) = P(X = 200) = \frac{1}{1000}$ as only one of the one thousand three digit numbers is going to be an exact match.

Next, $Q(\{20\}) = P(X = 20)$, so it must be determined how many of the one thousand possibilities will have exactly two matches. There are $\binom{3}{2} = 3$ different ways to choose the two digits that will match. Those digits are determined at that point and the remaining digit must be one of the nine digits that do not match the third spot, so there are $3 \cdot 9 = 27$ three digit numbers that match exactly two digits. So $Q(\{20\}) = P(X = 20) = \frac{27}{1000}$.

Finally, as Q is a probability, $Q(\{0\}) = 1 - Q(\{20\}) - Q(\{200\}) = \frac{972}{1000}$. ■

It is frequently the case that we are interested in functions that have real-valued outputs and we reserve the term "random variable" for such a situation.

Definition 3.1.5. *A "discrete random variable" is a function $X : S \to T$ where S is a sample space equipped with a probability P, and T is a countable (or finite) subset of the real numbers.*
From Theorem 3.1.2, P generates a probability on T. As it is a discrete space, the distribution may be determined by knowing the likelihood of each possible value of X. Because of this we define a function $f_X : T \to [0, 1]$ given by

$$f_X(t) = P(X = t)$$

referred to as a "probability mass function". Then for any event $A \subset T$ the quantity $P(X \in A)$ may be computed via

$$P(X \in A) = \sum_{t \in A} f_X(t) = \sum_{t \in A} P(X = t).$$

The function from Example 3.1.4 is a discrete random variable because it takes on one of the real values 0, 20, or 200. We calculated its probability mass function when describing its distribution and it is given by

$$f_X(0) = \frac{972}{1000}, \quad f_X(20) = \frac{27}{1000}, \quad f_X(200) = \frac{1}{1000}.$$

The function from Example 3.1.3 is not a discrete random variable by the above definition because its range is a collection of colors, not real numbers.

3.1.1 *Common Distributions*

When studying random variables it is often more important to know how they distribute probability onto their range than how they actually act as functions on their domains. As such it is useful to have a notation that recognizes the fact that two functions may be very different in terms of where they map domain elements, but nevertheless have the same range and produce the same distribution on this range.

Definition 3.1.6. *Let $X : S \to T$ and $Y : S \to T$ be discrete random variables. We say X and Y have equal distriubtion provided $P(X = t) = P(Y = t)$ for all $t \in T$.*

There are many distributions which appear frequently enough they deserve their own special names for easy identification. We shall use the symbol $\sim$ to mean "is distributed as" or "is equal in distribution to". For example, in the definition below $X \sim \text{Bernoulli}(p)$ should be read as "X has a Bernoulli(p) distribution". This says nothing explicit about how X behaves as a function on its domain, but completely describes how X distributes probability onto its range.

The following are common discrete distributions which we have seen arise previously in the text.

Definition 3.1.7. *$X \sim$ **Uniform**$(\{1, 2, \ldots, n\})$: Let $n \geq 1$ be an integer. If X is a random variable such that $P(X = k) = \frac{1}{n}$ for all $1 \leq k \leq n$ then we say that X is a uniform random variable on the set $\{1, 2, \ldots, n\}$.*

Definition 3.1.8. *$X \sim$ **Bernoulli**(p): Let $0 \leq p \leq 1$. When X is a random variable such that $P(X = 1) = p$ and $P(X = 0) = 1 - p$ we say that X is a Bernoulli random variable with parameter p. This takes the concept of a "Bernoulli trial" which we have previously discussed and puts it in the context of a random variable where 1 corresponds to success and 0 corresponds to failure.*

Definition 3.1.9. $X \sim$ **Binomial**(n,p)*: Let* $0 \leq p \leq 1$ *and let* $n \geq 1$ *be an integer. If* X *is a random variable taking values in* $\{0,1,\ldots,n\}$ *having a probability mass function*

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for all $0 \leq k \leq n$*, then* X *is a Binomial random variable with parameters* n *and* p*. We have seen that such a quantity describes the number of successes in* n *Bernoulli trials.*

Definition 3.1.10. $X \sim$ **Geometric**(p)*: Let* $0 < p < 1$*. If* X *is a random variable with values in* $\{1,2,3,\ldots\}$ *and a probability mass function*

$$P(X=k) = p \cdot (1-p)^{k-1}$$

for all $k \geq 1$*, then* X *is a geometric random variable with parameter* p*. Such a random variable arises when determining how many Bernoulli trials must be attempted before seeing the first success.*

Definition 3.1.11. $X \sim$ **Negative Binomial**(r,p)*: Let* $0 < p < 1$*. If* X *is a random variable with values in* $\{r, r+1, r+2, \ldots\}$ *and a probability mass function*

$$P(X=k) = \binom{k-1}{r-1} p^r \cdot (1-p)^{k-r}$$

for all $k \geq r$*, then* X *is a Negative Binomial random variable with parameters* (r,p)*. Such a random variable arises when determining how many Bernoulli trials must be attempted before seeing* r *successes.*

Definition 3.1.12. $X \sim$ **Poisson**(λ)*: Let* $\lambda > 0$*. When* X *is a random variable with values in* $\{0,1,2,\ldots\}$ *such that its probability mass function is*

$$P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

for all $k \geq 0$*, then* X *is called a Poisson random variable with parameter* λ*. We first used these distributions as approximations to a Binomial*(n,p) *when* n *was large and* p *was small.*

Definition 3.1.13. $X \sim \textbf{HyperGeo}(N, r, m)$*: Let* N*,* r*, and* m *be positive integers for which* $r < N$ *and* $m < N$*. Let* X *be a random variable taking values in the integers between* $\min\{m, r\}$ *and* $\max\{0, m - (N - r)\}$ *inclusive with probability mass function*

$$P(X = k) = \frac{\binom{r}{k}\binom{N-r}{m-k}}{\binom{N}{m}}$$

The random variable X *is called hypergeoemtric with parameters* N*,* r*, and* m*. Such quantities occur when sampling without replacement.*

EXERCISES

Ex. 3.1.1. Consider the experiment of flipping a coin four times and recording the sequence of heads and tails. Let S be the sample space of all sixteen possible orderings of the results. Let X be the function on S describing the number of tails among the flips. Let Y be the function on S describing the first flip (if any) to come up tails.

(a) Create a table as in Example 3.1.1 describing functions X and Y.

(b) Use the table to calculate $P(X = 2)$.

(c) Use the table to calculate $P(Y = 3)$.

Ex. 3.1.2. A pair of fair dice are thrown. Let X represent the larger of the two values on the dice and let Y represent the smaller of the two values.

(a) Describe S, the domain of functions X and Y. How many elements are in S?

(b) What are the ranges of X and Y. Do X and Y have the same range? Why or why not?

(c) Describe the distribution of X and describe the distribution of Y by finding the probability mass function of each. Is it true that X and Y have the same distribution ?

Ex. 3.1.3. A pair of fair dice are thrown. Let X represent the number of the first die and let Y represent the number of the second die.

(a) Describe S, the domain of functions X and Y. How many elements are in S?

(b) Describe T, the range of functions X and Y. How many elements are in T?

(c) Describe the distribution of X and describe the distribution of Y by finding the probability mass function of each. Is it true that X and Y have the same distribution ?

(d) Are X and Y the same function? Why or why not?

Ex. 3.1.4. Use the $\sim$ notation to classify the distributions of the random variables described by the scenarios below. For instance, if a scenario said, "let X be the number of heads in three flips of a coin" the apprORpriate answer would be $X \sim \text{Binomial}(3, \frac{1}{2})$ as that describes the number of successes in three Bernoulli trials.

(a) Let X be the number of 5's seen in four die rolls. What is the distribution of X?

(b) Each ticket in a certain lottery has a 20% chance to be a prize-winning ticket. Let Y be the number of tickets that need to be purchased before seeing the first prize-winner. What is the distribution of Y?

(c) A class of ten students is comprised of seven women and three men. Four students are randomly selected from the class. Let Z denote the number of men among the four randomly selected students. What is the distribution of Z?

Ex. 3.1.5. Suppose X and Y are random variables.

(a) Explain why $X + Y$ is a random variable.

(b) Theorem 3.1.2 does not require that X be real-valued. Why do you suppose that our definition of "random variable" insisted that such functions should be real-valued?

Ex. 3.1.6. Let $X : S \to T$ be a discrete random variable. Suppose $\{B_i\}_{i \geq 1}$ are sequence of events in T then show that $X^{-1}(\bigcup_{i=1}^{\infty} B_i) = \bigcup_{i=1}^{\infty} X^{-1}(B_i)$ and that if B_i and B_j are disjoint, then so are $X^{-1}(B_i)$ and $X^{-1}(B_j)$.

3.2 INDEPENDENT AND DEPENDENT VARIABLES

Most interesting problems require the consideration of several different random variables and an analysis of the relationships among them. We have already discussed what it means for a collection of events to be independent, and it is useful to extend this notion to random variables as well. As with events, we will first describe the notion of pairwise independence of two objects, before defining mutual independence of an arbitrary collection of objects.

3.2.1 *Independent Variables*

Definition 3.2.1. (Independence of a Pair of Random Variables) *Two random variables X and Y are independent if $(X \in A)$ and $(Y \in B)$ are independent for every event A in the range of X and every event B in the range of Y.*

As events become more complicated and involve multiple random variables, a notational shorthand will become useful. It is common in probability to write $(X \in A, Y \in B)$ for the event $(X \in A) \cap (Y \in B)$ and we will begin using this convention at this point.

Further, even though the definition of $X : S \to T$ and $Y : S \to U$ being independent random variables requires that $(X \in A)$ and $(Y \in B)$ be independent for all events $A \subset T$ and $B \subset U$, for discrete random variables it is enough to verify the events $(X = t)$ and $(Y = u)$ are independent events for all $t \in T$ and $u \in U$ to conclude they are independent (See Exercise 3.2.12).

Example 3.2.2. When we originally considered the example of rolling a pair of dice, we viewed the results as thirty-six equally likely outcomes. However, it is also possible to view the result of each die as a random variable in its own right, and then consider the possible results of the pair of random variables. Let $X, Y \sim \text{Uniform}(\{1, 2, 3, 4, 5, 6\})$ and suppose X and Y are independent. If $x, y \in \{1, 2, 3, 4, 5, 6\}$ what is $P(X = x, Y = y)$?

By indpendence $P(X = x, Y = y) = P(X = x)P(Y = y) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$. Therefore, the result is identical to the original perspective – each of the thirty-six outcomes of the pair of dice is equally likely. ■

Definition 3.2.3. (Mutual Independence of Random Variables) *A finite collection of random variables $X_1, X_2, \ldots, X_n$ is mutually independent if the sets $(X_j \in A_j)$ are mutually independent for all events A_j in the ranges of the corresponding X_j.*
An arbitrary collection of random variables X_t where $t \in I$ for some index set I is mutually independent if every finite sub-collection is mutually independent.

For many problems it is useful to think about repeating a single experiment many times with the results of each repetition being independent from every other. Though the results are assumed to be independent, the experiment itself remains the same, so the random variables produced all have the same distribution. The resulting sequence of random variables $X_1, X_2, X_3, \ldots$ is referred to as "i.i.d." (standing for "independent and identically distributed"). When considering such sequences we will sometimes write $X_1, X_2, X_3, \ldots$

are i.i.d. with distribution X, where X is a random variable that shares their common distribution.

Example 3.2.4. Let $X_1, X_2, \ldots, X_n$ be i.i.d. with a Geometric(p) distribution. What is the probabilty that all of these random variables are larger than some positive integer j?

As a preliminary calculation, if $X \sim$ Geometric(p) and if $j \geq 1$ is an integer we may determine $P(X > j)$.

$$\begin{aligned} P(X > j) &= \sum_{i=j+1}^{\infty} P(X = i) = \sum_{i=j+1}^{\infty} p(1-p)^{i-1} \\ &= \frac{p \cdot (1-p)^j}{1-(1-p)} = (1-p)^j. \end{aligned}$$

But each of X_1, X_2, ..., X_n have this distribution, so using the computatation above, together with independence, we have

$$\begin{aligned} P(X_1 > j, X_2 > j, \ldots, X_n > j) &= P(X_1 > j)P(X_2 > j)\ldots P(X_n > j) \\ &= (1-p)^j \cdot (1-p)^j \cdot \cdots \cdot (1-p)^j \\ &= (1-p)^{nj}. \end{aligned}$$

■

3.2.2 *Conditional, Joint, and Marginal Distributions*

Consider a problem involving two random variables. Let X be the number of centimeters of rainfall in a certain forest in a given year, and let Y be the number of square meters of the forest burned by fires that same year. It seems these variables should be related; knowing one should affect the probabilities associated with the values of the other. Such random variables are not independent of each other and we now introduce several ways to compute probabilities under such circumstances. An important concept toward this end is the notion of a "conditional distribution" which reflects the fact that the occurrence of an event may affect the likely values of a random variable.

Definition 3.2.5. *Let X be a random variable on a sample space S and let $A \subset S$ be an event such that $P(A) > 0$. Then the probability Q described by*

$$Q(B) = P(X \in B \mid A) \tag{3.2.1}$$

is called the "conditional distribution" of X given the event A.

As with any discrete random variable, the distribution is completely determined by the probabilities associated with each possible value the random variable may assume. This means the conditional distribution may be considered known provided the values of $P(X = a|A)$ are known for every $a \in \text{Range}(X)$. Though this definition allows for A to be any sort of event, in this section we will mainly consider examples where A describes the outcome of some random variable. So a notation like $P(X|Y = b)$ will be the conditional distribution of the random variable X given that the random variable Y is known to have the value b.

In many cases random variables are dependent in such a way that the distribution of one variable is known in terms of the values taken on by another.

Example 3.2.6. Let $X \sim \text{Uniform}(\{1, 2\})$ and let Y be the number of heads in X tosses of a fair coin. Clearly X and Y should not be independent. In particular, a result of $Y = 0$ could occur regardless of the value of X, but a result of $Y = 2$ guarantees that $X = 2$ as two heads could not be observed with just one flip on the coin. Any information regarding X or Y may influence the distribution of the other, but the description of the variables makes it clearest how Y depends on X. If $X = 1$ then Y is the number of heads in one flip of a fair coin. Letting A be the event $(X = 1)$ and using the terminology of (3.2.1) from Definition 3.2.5, we can say the conditional distribution of Y given that $X = 1$ is a Bernoulli($\frac{1}{2}$). We will use the notation

$$(Y \mid X = 1) \sim Bernoulli(\frac{1}{2})$$

to indicate this fact. In other words, this notation means the same thing as the pair of equations

$$\begin{aligned} P(Y = 0 \mid X = 1) &= \frac{1}{2} \\ P(Y = 1 \mid X = 1) &= \frac{1}{2} \end{aligned}$$

If $X = 2$ then Y is the number of heads in two flips of a fair coin and therefore $(Y \mid X = 2) \sim \text{Binomial}(2, \frac{1}{2})$ which means the following three equations hold:

$$\begin{aligned} P(Y = 0 \mid X = 2) &= \frac{1}{4} \\ P(Y = 1 \mid X = 2) &= \frac{1}{2} \\ P(Y = 2 \mid X = 2) &= \frac{1}{4} \end{aligned}$$

■

The conditional probabilities of the previous example were easily determined in part because the description of Y was already given in terms of X, but frequently random

variables may be dependent in some way that is not so explicitly described. A more general method of expressing the dependence of two (or more) variables is to present the probabilities associated with all combinations of possible values for every variable. This is known as their joint distribution.

Definition 3.2.7. *If X and Y are discrete random variables, the "joint distribution" of X and Y is the probability Q on pairs of values in the ranges of X and Y defined by*

$$Q((a,b)) = P(X = a, Y = b).$$

The definition may be expanded to a finite collection of discrete random variables X_1, X_2, ..., X_n for which the joint distribution of all n variables is the probability defined by

$$Q((a_1, a_2, \ldots, a_n)) = P(X_1 = a_1, X_2 = a_2, \ldots, X_n = a_n).$$

In the above definition as discussed before for any event D,

$$Q(D) = \sum_{(a_1,a_2,\ldots,a_n)\in D} Q((a_1, a_2, \ldots, a_n)).$$

For a pair of random variables with few possible outcomes, it is common to describe the joint distribution using a chart for which the columns correspond to possible X values, the rows to possible Y values, and for which the entries of the chart are probabilities.

Example 3.2.8. Let X and Y be the dependent variables described in Example 3.2.6. The X variable will be either 1 or 2. The Y variable could be as low as 0 (if no heads are flipped) or as high as 2 (if two coins are flipped and both show heads). As $\text{Range}(X) = \{1, 2\}$ and as $\text{Range}(Y) = \{0, 1, 2\}$, the pair (X, Y) could potentially be any of the six possible pairings (though, in fact, one of the pairings has probability zero). To find the joint distribution of X and Y we must calculate the probabilities of each possibility. In this case the values may be obtained using the definition of conditional probability. For instance,

$$P(X = 1, Y = 0) = P(Y = 0|X = 1) \cdot P(X = 1) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

and

$$P(X = 1, Y = 2) = P(Y = 2|X = 1) \cdot P(X = 1) = 0 \cdot \frac{1}{2} = 0.$$

The entire joint distribution $P(X = a, Y = b)$ is described by the following chart.

	$X = 1$	$X = 2$
$Y = 0$	$1/4$	$1/8$
$Y = 1$	$1/4$	$1/4$
$Y = 2$	0	$1/8$

■

Knowing the joint distribution of random variables gives a complete picture of the probabilities associated with those variables. From that information it is possible to compute all conditional probabilities of one variable from another. For instance, in the example analyzed above, the variable Y was originally described in terms of how it depended on X. However, this also means that X should be dependent on Y. The joint distribution may be used to determine how.

Example 3.2.9. Let X and Y be the variables of Example 3.2.6. How may the conditional distributions of X given values of Y be determined?

There will be three different conditional distributions depending on whether $Y = 0$, $Y = 1$, or $Y = 2$. Below we will solve the $Y = 0$ case. The other two cases will be left as exercises. The conditional distribution of X given that $Y = 0$ is determined by the values of $P(X = 1|Y = 0)$ and $P(X = 2|Y = 0)$ both of which may be computed using Bayes' rule.

$$\begin{aligned} P(X = 1|Y = 0) &= \frac{P(Y = 0|X = 1) \cdot P(X = 1)}{P(Y = 0)} \\ &= \frac{P(Y = 0|X = 1) \cdot P(X = 1)}{P(Y = 0|X = 1) \cdot P(X = 1) + P(Y = 0|X = 2) \cdot P(X = 2)} \\ &= \frac{(1/2)(1/2)}{(1/2)(1/2) + (1/4)(1/2)} = \frac{2}{3} \end{aligned}$$

As the only values for X are 1 and 2 it must be that $P(X = 2|Y = 0) = \frac{1}{3}$. ■

Just because X and Y are dependent on each other doesn't mean they need to be thought of as a pair. It still makes sense to talk about the distribution of X as a random variable in its own right while ignoring its dependence on the variable Y. When there are two or more variables under discussion, the distribution of X alone is sometimes called the "marginal" distribution of X because it can be computed using the margins of the chart describing the joint distribution of X and Y.

Example 3.2.10. Continue with X and Y as described in Example 3.2.6. Below is the chart describing the joint distribution of X and Y that was created in Example 3.2.8, but with the addition of one column on the right and one row at the bottom. The entries in the extra column are the sums of the values in the corresponding row; likewise the entries in the extra row are the sums of the values in the corresponding column.

	$X=1$	$X=2$	Sum
$Y=0$	1/4	1/8	3/8
$Y=1$	1/4	1/4	4/8
$Y=2$	0	1/8	1/8
Sum	1/2	1/2	

The values in the right hand margin (column) exactly describe the distribution of Y. For instance the event $(Y=0)$ can be partitioned into two disjoint events $(X=1, Y=0) \cup (X=2, Y=0)$ each of which is already described in the joint distribution chart. Adding them together gives the result that $P(Y=0) = \frac{3}{8}$. In a similar fashion, the bottom margin (row) describes the distribution of X. This extended chart also makes it numerically clearer why these two random variables are dependent. For instance,

$$P(X=1, Y=0) = \frac{1}{4} \qquad \text{while} \qquad P(X=1) \cdot P(Y=0) = \frac{3}{16}$$

As these quantities are unequal, the random variables cannot be independent. ■

In general, knowing the marginal distributions of X and Y is not sufficient information to reconstruct their joint distribution. This is because the marginal distributions do not provide any information about how the random variables relate to each other. However, if X and Y happen to be independent, then their joint distribution may easily be computed from the marginals as

$$P(X=x, Y=y) = P(X=x)P(Y=x)$$

3.2.3 *Memoryless Property of the Geometric Random Variable*

It is also possible to calculate conditional probabilities of a random variable based on subsets of its own values. A particularly important example of this is the "memoryless property" of geometric random variables.

Example 3.2.11. Suppose we toss a fair coin until the first head appears. Let X be the number of tosses performed. We have seen in Example 2.1.2 that $X \sim \text{Geometric}(\frac{1}{2})$. Note that if m is a positive integer,

$$P(X > m) = \sum_{k=m+1}^{\infty} P(X=k) = \sum_{k=m+1}^{\infty} \frac{1}{2^k} = \frac{1}{2^m}$$

Now let n be a positive integer and suppose we take the event $(X > n)$ as given. In other words, we assume we know that none of the first n flips resulted in heads. What is the conditional distribution of X given this new information? A routine calculation shows

$$P(X > n+m \mid X > n) = \frac{P(X > n+m)}{P(X > n)} = \frac{\frac{1}{2^{m+n}}}{\frac{1}{2^n}} = \frac{1}{2^m}$$

As a consequence,

$$P(X > n+m \mid X > n) = P(X > m). \tag{3.2.2}$$

Given that a result of heads has not occurred by the n-th flip, the probability that such a result will require at least m more flips is identical to the (non-conditional) probability the result would have required more than m flips from the start. In other words, if we know that the first n flips have not yet produced a head, the number of additional flips required to observe the first head still is a Geometric$(\frac{1}{2})$ random variable. This is called the "memoryless property" of the geometric distribution as it can be interpreted to mean that when waiting times are geometrically distributed, no matter how long we wait for an event to occur, the future waiting time always looks the same given that the event has not occurred yet. The result remains true of geometric variables of any parameter p, a fact which we leave as an exercise. ■

3.2.4 *Multinomial Distributions*

Consider a situation similar to that of Bernoulli trials, but instead of results of each attempt limited to success or failure, suppose there are many different possible results for each trial. As with the Bernoulli trial cases we assume that the trials are mutually independent, but identically distributed. In the next example we will show how to calculate the joint distribution for the random variables representing the number of times each outcome occurs.

Example 3.2.12. Suppose we perform n i.i.d. trials each of which has k different possible outcomes. For $j = 1, 2, \ldots, k$, let p_j represent the probability any given trial results in the j-th outcome and let X_j represent the number of the n trials that result in the j-th outcome. The joint distribution of all of the random variables X_1, X_2, …, X_k is called a "multinomial distribution".

Let $B(x_1, x_2, \ldots, x_k) = \{X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k\}$. Then,

$$P(B(x_1, x_2, \ldots, x_k)) = \sum_{\omega \in B(x_1, x_2, \ldots, x_k)} P(\{\omega\})$$

Each $\omega \in B(x_1, x_2, \ldots, x_k)$ is an element in the sample space corresponding to the j-th outcome occuring exactly x_j times. As the trials are independent, and as an outcome j occurs in x_j trials, each of which had probability p_j, this means

$$P(\{\omega\}) = \prod_{j=1}^{k} p_j^{x_j}$$

Consequently, each outcome in $B(x_1, x_2, \ldots, x_k)$ has the same probability. So to determine the likelihood of the event, we need only determine $|B(x_1, x_2, \ldots, x_k)|$, the number of outcomes the event contains. The calculation of this quantity is a combinatorial problem; it is the number of ways of allocating n balls in k boxes, such that x_j of them fall into box j. We leave it as an exercise to prove that

$$|B(x_1, x_2, \ldots, x_k)| = \frac{n!}{x_1!\, x_2!\, \ldots\, x_k!}$$

With that computation complete, the joint distribution of $X_1, X_2, \ldots X_k$ is given by

$$P(X_1 = x_1, \ldots, X_k = x_k) = \begin{cases} \frac{n!}{x_1!\, x_2! \ldots x_k!} \prod_{j=1}^{k} p_j^{x_j} & \text{if } x_j \in \{0, 1, \ldots, n\} \\ & \text{and } \sum_{j=1}^{k} x_j = n \\ 0 & \text{otherwise} \end{cases}$$

■

EXERCISES

Ex. 3.2.1. An urn has four balls labeled $1, 2, 3$, and 4. A first ball is drawn and its number is denoted by X. A second ball is then drawn from the three remaining balls in the urn and its number is denoted by Y.

(a) Calculate $P(X = 1)$.

(b) Calculate $P(Y = 2 \mid X = 1)$.

(c) Calculate $P(Y = 2)$.

(d) Calculate $P(X = 1, Y = 2)$.

(e) Are X and Y independent? Why or why not?

Ex. 3.2.2. Two dice are rolled. Let X denote the sum of the dice and let Y denote the value of the first die.

(a) Calculate $P(X = 7)$ and $P(Y = 4)$.

(b) Calculate $P(X = 7, Y = 4)$.

(c) Calculate $P(X = 5)$ and $P(Y = 4)$.

(d) Calculate $P(X = 5, Y = 4)$.

(e) Are X and Y independent? Why or why not?

Ex. 3.2.3. Let X and Y be the variables described in Example 3.2.6.

(a) Determine the conditional distribution of X given that $Y = 1$.

(b) Determine the conditional distribution of X given that $Y = 2$.

Ex. 3.2.4. Let X and Y be random variables with joint distribution given by the chart below.

	$X = 0$	$X = 1$	$X = 2$
$Y = 0$	1/12	0	3/12
$Y = 1$	2/12	1/12	0
$Y = 2$	3/12	1/12	1/12

(a) Compute the marginal distributions of X and Y.

(b) Compute the conditional distribution of X given that $Y = 2$.

(c) Compute the conditional distribution of Y given that $X = 2$.

(d) Carry out a computation to show that X and Y are not independent.

Ex. 3.2.5. Let X be a random variable with range $\{0, 1\}$ and distribution

$$P(X = 0) = \frac{1}{3} \qquad \text{and} \qquad P(X = 1) = \frac{2}{3}$$

and let Y be a random variable with range $\{0, 1, 2\}$ and distribution

$$P(Y = 0) = \frac{1}{5}, \qquad P(Y = 1) = \frac{1}{5}, \qquad \text{and} \qquad P(Y = 2) = \frac{3}{5}$$

Suppose that X and Y are independent. Create a chart describing the joint distribution of X and Y.

Ex. 3.2.6. Consider six independent trials each of which are equally likely to produce a result of 1, 2, or 3. Let X_j denote the number of trials that result in j. Calculate $P(X_1 = 1, X_2 = 2, X_3 = 3)$.

Ex. 3.2.7. Prove the combinatorial fact from Example 3.2.12 in the following way. Let $A_n(x_1, x_2, \ldots, x_k)$ denote the number of ways of putting n balls into k boxes in such a way that exactly x_j balls wind up in box j for $j = 1, 2, \ldots, k$.

(a) Prove that $A_n(x_1, x_2, \ldots, x_k) = \binom{n}{x_1} A_{n-x_1}(x_2, x_3, \ldots, x_k)$.

(b) Use part (a) and induction to prove that $A_n(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1!\, x_2! \ldots x_k!}$.

Ex. 3.2.8. Let X be the result of a fair die roll and let Y be the number of heads in X coin flips.

(a) Both X and $(Y|X = n)$ can be written in terms of common distributions using the $\sim$ notation. What is the distribution of X? What is the distribution of $(Y|X = n)$ for $n = 1, \ldots 6$?

(b) Determine the joint distribution for X and Y.

(c) Calculate the marginal distribution of Y.

(d) Compute the conditional distribution of X given that $Y = 6$.

(e) Compute the conditional distribution of X given that $Y = 0$.

(f) Perform a computation to prove that X and Y are not independent.

Ex. 3.2.9. Suppose the number of earthquakes that occur in a year, anywhere in the world, is a Poisson random variable with mean λ. Suppose the probability that any given earthquake has magnitude at least 5 on the Richter scale is p independent of all other quakes. Let $N \sim \text{Poisson}(\lambda)$ be the number of earthquakes in a year and let M be the number of earthquakes in a year with magnitude at least 5, so that $(M|N = n) \sim \text{Binomial}(n, p)$.

(a) Calculate the joint distribution of M and N.

(b) Show that the marginal distribution of M is determined by

$$P(M = m) = \frac{1}{m!} e^{-\lambda} (\lambda p)^m \sum_{n=m}^{\infty} \frac{\lambda^{n-m}}{(n-m)!} (1-p)^{n-m}$$

for $m > 0$.

(c) Perform a change of variables (where $k = n - m$) in the infinite series from part (b) to prove

$$P(M = m) = \frac{1}{m!} e^{-\lambda} (\lambda p)^m \sum_{k=0}^{\infty} \frac{(\lambda(1-p))^k}{k!}$$

(d) Use part (c) together with the infinite series equality $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$ to conclude that $M \sim \text{Poisson}(\lambda p)$.

Ex. 3.2.10. Let X be a discrete random variable which has $\mathbb{N} = \{1, 2, 3, \dots\}$ as its range. Suppose that for all positive integers m and n, X has the memoryless property – $P(X > n + m \mid X > n) = P(X > m)$. Prove that X must be a geometric random variable. [Hint: Define $p = P(X = 1)$ and use the memoryless property to calculate $P(X = n)$ inductively].

Ex. 3.2.11. A discrete random variable X is called "constant" if there is a single value c for which $P(X = c) = 1$.

(a) Prove that if X is a constant discrete random variable then X is independent of itself.

(b) Prove that if X is a discrete random variable which is independent of itself, then X must be constant. [Hint: It may help to look at Exercise 1.4.8].

Ex. 3.2.12. Let $X : S \to T$ and $Y : S \to U$ be discrete random variables. Show that if

$$P(X = t, Y = u) = P(X = t)P(Y = u)$$

for all $t \in T$ and $u \in U$ then X and Y are independent random variables.

3.3 FUNCTIONS OF RANDOM VARIABLES

There are many circumstances where we want to consider functions applied to random variables as inputs of functions. For a simple geometric example, suppose a rectangle is selected in such a way that its width X and its length Y are both random variables with known joint distribution. The area of the rectangle is $A = XY$, and as X and Y are random, A should be random as well. How may the distribution of A be calculated from the joint distribution of X and Y? In general, if a new random variable Z depends on random variables $X_1, X_2, \dots, X_n$ which have a given joint distribution, how may the distribution of Z be calculated from what is already known? In this section we discuss the answers to such questions and also address related issues surrounding independence.

If $X : S \to T$ is a random variable and if $f : T \to R$ is a function, then the quantity $f(X)$ makes sense as a composition of functions $f \circ X : S \to R$. In fact, as $f(X)$ is defined on the sample space S, this new composition is itself a random variable.

The same reasoning holds for functions of more than one variable. If $X_1, X_2, \dots, X_n$ are random variables then $f(X_1, X_2, \dots, X_n)$ is a random variable provided f is defined for

the values the X_j variables produce. Below we illustrate how to calculate the distribution of $f(X_1, X_2, \ldots, X_n)$ in terms of the joint distribution of the X_j input variables. We demonstrate the method with several examples followed by a general theorem.

3.3.1 *Distribution of $f(X)$ and $f(X_1, X_2, \ldots, X_n)$*

The distribution of $f(X)$ involves the probability of events such as $(f(X) = a)$ for values of a that the function may produce. The key to calculating this probability is that these events may be rewritten in terms of the input values of X instead of the output values of $f(X)$.

EXAMPLE 3.3.1. Let $X \sim \text{Uniform}(\{-2, -1, 0, 1, 2\})$ and let $f(x) = x^2$. Determine the range and distribution of $f(X)$.

As $f(X) = X^2$, the values that $f(X)$ produces are the squares of the values that X produces. Squaring the values in $\{-2, -1, 0, 1, 2\}$ shows the range of $f(X)$ is $\{0, 1, 4\}$. The probabilities that $f(X)$ takes on each of these three values determine the distribution of $f(X)$ and these probabilities can be easily calculated from the known probabilities associated with X.

$$\begin{aligned} P(f(X) = 0) &= P(X = 0) = \frac{1}{5} \\ P(f(X) = 1) &= P((X = 1) \cup (X = -1)) = \frac{1}{5} + \frac{1}{5} = \frac{2}{5} \\ P(f(X) = 4) &= P((X = 2) \cup (X = -2)) = \frac{1}{5} + \frac{1}{5} = \frac{2}{5} \end{aligned}$$

■

A complication with this method is that there may be many different inputs that produce the same output. Sometimes a problem requires careful consideration of all ways that a given output may be produced. For instance,

EXAMPLE 3.3.2. What is the probability the sum of three dice will equal six? Let X, Y, and Z be the results of the first, second, and third die respectively. These are i.i.d. random variables each distributed as $\text{Uniform}(\{1, 2, 3, 4, 5, 6\})$. A sum of six can be arrived at in three distinct ways:

Case I: through three rolls of 2;

Case II: through one roll of 3, one roll of 2, and one roll of 1; or

Case III: through one roll of 4 and two rolls of 1

The first of these is the simplest to deal with as independence gives

$$P(X=2,Y=2,Z=2) = P(X=2)\cdot P(Y=2)\cdot P(Z=2) = \frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{6} = \frac{1}{216}$$

The other cases involve a similar computation, but are complicated by the consideration of which number shows up on which die. For instance, both events $(X=1,Y=2,Z=3)$ and $(X=3,Y=2,Z=1)$ are included as part of Case II as are four other permutations of the numbers. Likewise Case III includes three permutations, one of which is $(X=4,Y=1,Z=1)$. Putting all three cases together,

$$\begin{aligned} P(\text{sum of 6}) &= P(\text{Case I}) + P(\text{Case II}) + P(\text{Case III}) \\ &= \frac{1}{216} + 6\cdot\frac{1}{216} + 3\cdot\frac{1}{216} = \frac{5}{108}. \end{aligned}$$

So there is slightly less than a 5% chance three rolled dice will produce a sum of six. ■

This method may also be used to show relationships among the common (named) distributions that have been previously described, as in the next two examples.

Example 3.3.3. Let $X,Y \sim \text{Bernoulli}(p)$ be two independent random variables. If $Z = X+Y$, show that $Z \sim \text{Binomial}(2,p)$.

This result should not be surpirsing given how Bernoulli and Binomial distributions arose in the first place. Each of X and Y produces a value of 0 if the corresponding Bernoulli trial was a failure and 1 if the trial was a success. Therefore $Z = X+Y$ equals the total number of successes in two independent Bernoulli trials, which is exactly what led us to the Binomial distribution in the first place. However, it is instructive to consider how this problem relates to the current topic of discussion.

As each of X and Y is either 0 or 1 the possible values of Z are in the set {0,1,2}. A result of $Z=0$ can only occur if both X and Y are zero. So,

$$\begin{aligned} P(Z=0) &= P(X=0,Y=0) \\ &= P(X=0)\cdot P(Y=0) \\ &= (1-p)(1-p) \\ &= (1-p)^2. \end{aligned}$$

Similarly, $P(Z=2) = P(X=1)\cdot P(Y=1) = p^2$.

There are two different ways that Z could equal 1, either $X = 1$ and $Y = 0$, or $X = 0$ and $Y = 1$. So,

$$\begin{aligned} P(Z = 1) &= P((X = 1, Y = 0) \cup (X = 0, Y = 1)) \\ &= P(X = 1, Y = 0) + P(X = 0, Y = 1) \\ &= p(1-p) + (1-p)p \\ &= 2p(1-p) \end{aligned}$$

These values of $P(Z = 0), P(Z = 1)$, and $P(Z = 2)$ are exactly what define $Z \sim \text{Binomial}(2, p)$. ■

Two of the previous three examples involve adding random variables together. In fact, addition is one of the most common examples of applying functions to random quantities. In the previous situations, calculating the distribution of the sum was relatively simple because the component variables only had finitely many outcomes. But now suppose X and Y are random variables taking values in $\{0, 1, 2, \dots\}$ and suppose $Z = X + Y$. How could $P(Z = n)$ be calculated?

As both X and Y are non-negative and as $Z = X + Y$, the value of Z must be at least as large as either X or Y individually. If $Z = n$, then X could take on any value $j \in \{0, 1, \dots, n\}$, but once that value is determed, the value of Y is compelled to be $n - j$ to give the appropriate sum. In other words, the event $(Z = n)$ partitions into the following union.

$$(Z = n) = \bigcup_{j=0}^{n} (X = j, Y = n - j).$$

When X and Y are independent, this means

$$\begin{aligned} P(Z = n) &= P\left(\bigcup_{j=0}^{n} (X = j, Y = n - j)\right) \\ &= \sum_{j=0}^{n} P(X = j, Y = n - j) \\ &= \sum_{j=0}^{n} P(X = j) \cdot P(Y = n - j) \end{aligned}$$

Such a computation is usually referred to as a "convolution" which will be addressed more generally later in the text. It occurs regularly when determining the distribution of sums of independent random variables.

Example 3.3.4. Let $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$ be independent random variables. Let $Z = X + Y$.

(a) Find the distribution of Z.

(b) Find the conditional distribution of $X \mid Z$.

For $x, y \in \{0, 1, 2, \dots\}$, we have

$$\begin{aligned} P(X = x, Y = y) &= P(X = x) \cdot P(Y = y) \\ &= e^{-\lambda_1} \frac{\lambda_1^x}{x!} \cdot e^{-\lambda_2} \frac{\lambda_2^y}{y!}. \end{aligned}$$

(a) As computed above, the distribution of Z is given by the convolution. For any $n = 0, 1, 2, \dots$ we have

$$\begin{aligned} P(Z = n) &= P(X + Y = n) \\ &= \sum_{j=0}^{n} P(X = j) \cdot P(Y = n - j) \\ &= \sum_{j=0}^{n} e^{-\lambda_1} \frac{\lambda_1^j}{j!} \cdot e^{-\lambda_2} \frac{\lambda_2^{n-j}}{(n-j)!} \\ &= e^{-(\lambda_1+\lambda_2)} \sum_{j=0}^{n} \frac{\lambda_1^j \lambda_2^{n-j}}{j!(n-j)!} \\ &= e^{-(\lambda_1+\lambda_2)} \frac{1}{n!} \sum_{j=0}^{n} \frac{n!}{j!(n-j)!} \lambda_1^j \lambda_2^{n-j} \\ &= e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1 + \lambda_2)^n}{n!} \end{aligned}$$

where in the last line we have used the binomial expansion (2.1.1). Hence we can conclude that $Z \sim \text{Poisson}\ (\lambda_1 + \lambda_2)$.

The above calculation is easily extended by an induction argument to obtain the fact that if $\lambda_i > 0$, X_i, $1 \le i \le k$ are independent Poisson(λ_i) distributed random variables (respectively). Then $Z = \sum_{i=1}^{k} X_i$ has Poisson $(\sum_{i=1}^{k} \lambda_i)$ distribution. Thus if we have k independent Poisson (λ) random variables then $\sum_{i=1}^{k} X_i$ has Poisson$(k\lambda)$ distribution.

(b) We readily observe that X and Z are dependent. We shall now try to understand the conditional distribution of $(X|Z = n)$. As the ranges of X and Y do not have any negative numbers, given that $Z = X + Y = n$, X can only take values in $\{0, 1, 2, 3, \dots, n\}$.

For $k \in \{0, 1, 2, 3, \ldots, n\}$ we have,

$$\begin{aligned} P(X = k \mid Z = n) &= \frac{P(X = k, X + Y = n)}{P(X + Y = n)} = \frac{P(X = k, Y = n - k)}{P(X + Y = n)} \\ &= \frac{P(X = k)P(Y = n - k)}{P(X + Y = n)} \\ &= \frac{e^{-\lambda_1}\frac{\lambda_1^k}{k!} \cdot e^{-\lambda_2}\frac{\lambda_2^{n-k}}{(n-k)!}}{e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^n}{n!}} = \frac{n!}{k!(n-k)!}\frac{\lambda_1^k\lambda_2^{n-k}}{(\lambda_1+\lambda_2)^n} \\ &= \binom{n}{k}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k}. \end{aligned}$$

Hence $(X \mid Z = n) \sim \text{Binomial}(n, \frac{\lambda_1}{\lambda_1+\lambda_2})$. ■

The point of the examples above is that a probability associated with a functional value $f(X_1, X_2, \ldots, X_n)$ may be calculated directly from the probabilities associated with the input variables $X_1, X_2, \ldots, X_n$. The following theorem explains how this may be accomplished generally for any number of variables.

Theorem 3.3.5. *Let X_1, X_2, …, X_n be random variables defined on a single sample space S. Let f be a function of n variables for which $f(X_1, X_2, \ldots, X_n)$ is defined in the range of the X_j variables. Let B be a subset of the range of f. Then,*

$$P(f(X_1, X_2, \ldots, X_n) \in B) = P((X_1, X_2, \ldots, X_n) \in f^{-1}(B)).$$

Proof. Note that the events $(f(X_1, X_2, \ldots, X_n) \in B)$ and $((X_1, X_2, \ldots, X_n) \in f^{-1}(B))$ are both subsets of S, as outcomes $s \in S$ determine the values of the X_j variables which in turn determine the output of f. The theorem follows immediately from the set theoretic fact that

$$f(X_1(s), X_2(s), \ldots, X_n(s)) \in B \iff (X_1(s), X_2(s), \ldots X_n(s)) \in f^{-1}(B)$$

This is because the expression $f(X_1(s), X_2(s), \ldots, X_n(s)) \in B$ is what defines s to be an outcome in the event $(f(X_1, X_2, \ldots, X_n) \in B)$. Likewise, the expression

$$(X_1(s), X_2(s), \ldots X_n(s)) \in f^{-1}(B)$$

defines s to be in the event $((X_1, X_2, \ldots X_n) \in f^{-1}(B))$. As these events are equal, they have the same probability. ■

3.3.2 *Functions and Independence*

If X and Y are independent random variables, does that guarantee that functions $f(X)$ and $g(Y)$ of these random variables are also indpendent? If we take the intuitive view of independence as saying "knowing information about X does not affect the probabilities associated with Y" then it seems the answer should be "yes". After all, X determines the value of $f(X)$ and Y determines the value of $g(Y)$. So information about $f(X)$ should translate to information about X and infromation about $g(Y)$ should translate to information about Y. Therefore if information about $f(X)$ affected probabilities associated with $g(Y)$, then it seems there should be information about X that would affect the probability assoicated with Y. We generalize this argument and make it more rigorous in the following result.

Theorem 3.3.6. *Fix $n \geq 1$. For each $j \in \{1, 2, \ldots, n\}$ let $i \in \{1, 2, \ldots, m_j\}$ for some positive integer m_j. Suppose $X_{i,j}$ is an array of mutually independent discrete random variables and we define*

$$Y_j = f_j(X_{1,j}, X_{2,j}, \ldots X_{m_j,j}),$$

where $f_j : \mathbb{R}^{m_j} \to \mathbb{R}$ are continuous functions. Then the resulting variables $Y_1, Y_2, \ldots, Y_n$ are mutually independent.

Informally this theorem says that random quantities produced from independent inputs will, themselves, be independent.

Proof. Let $B_1, B_2, \ldots, B_n$ be sets in the ranges of $Y_1, Y_2, \ldots, Y_n$ respectively. Using independence and some set-theoretic identities, we have

$$\begin{aligned}
&P(Y_1 \in B_1, \ldots, Y_n \in B_n) \\
&\quad = \quad P(f_1(X_{1,1}, \ldots, X_{m_1,1}) \in B_1, \ldots, f_n(X_{1,n}, \ldots, X_{m_n,n}) \in B_n) \\
&\quad = \quad P((X_{1,1}, \ldots, X_{m_1,1}) \in f_1^{-1}(B_1), \ldots, (X_{1,n}, \ldots, X_{m_n,n}) \in f_n^{-1}(B_n)) \\
&\quad = \quad \prod_{i=1}^{n} P((X_{i,1}, \ldots, X_{m_i,i}) \in f_i^{-1}(B_i)) \\
&\quad = \quad \prod_{i=1}^{n} P(f_i(X_{i,1}, \ldots, X_{m_i,i}) \in B_i) \\
&\quad = \quad P(Y_1 \in B_1) \cdots P(Y_n \in B_n)
\end{aligned}$$

It follows that $Y_1, Y_2, \ldots, Y_n$ are mutually independent. ∎

EXERCISES

Ex. 3.3.1. Let $X \sim \text{Uniform}(\{1,2,3\})$ and $Y \sim \text{Uniform}(\{1,2,3\})$ be independent and let $Z = X + Y$.

(a) Determine the range of Z.

(b) Determine the distriubtion of Z.

(c) Is Z uniformly distributed over its range?

Ex. 3.3.2. Consider the experiment of rolling three dice and calculating the sum of the rolls. Answer the following questions.

(a) What is the range of possible results of this experiment?

(b) Calculate the probability the sum equals three.

(c) Calculate the probability the sum equals four.

(d) Calculate the probability the sum equals five.

(e) Calculate the probability the sum equals ten.

Ex. 3.3.3. Let $X \sim \text{Bernoulli}(p)$ and $Y \sim \text{Bernoulli}(q)$ be independent.

(a) Prove that XY is a Bernoulli random variable. What is its parameter?

(b) Prove that $(1 - X)$ is a Bernoulli random variable. What is its parameter?

(c) Prove that $X + Y - XY$ is a Bernoulli random variable. What is its parameter?

Ex. 3.3.4. Let $X \sim \text{Binomial}(n,p)$ and $Y \sim \text{Binomial}(m,p)$. Assume X and Y are independent and let $Z = X + Y$. Prove that $Z \sim \text{Binomial}(m+n,p)$.

Ex. 3.3.5. Let $X \sim \text{Negative Binomial}(r,p)$ and $Y \sim \text{Negative Binomial}(s,p)$. Assume X and Y are independent and let $Z = X + Y$. Prove that $Z \sim \text{Negative Binomial}(r+s,p)$.

Ex. 3.3.6. Consider one flip of a single fair coin. Let X denote the number of heads on the flip and let Y denote the number of tails on the flip.

(a) Show that $X, Y \sim Bernoulli(\frac{1}{2})$.

(b) Let $Z = X + Y$ and explain why $P(Z = 1) = 1$.

(c) As (b) clearly says that Z cannot be a $Binomial(2, \frac{1}{2})$, explain why this result does not conflict with the conclusion of Example 3.3.3.

Ex. 3.3.7. Let $X \sim \text{Geometric}(p)$ and $Y \sim \text{Geometric}(p)$ be independent. Let $Z = X + Y$.

(a) Determine the range of Z.

(b) Use a convolution to prove that $P(Z = n) = (n-1)p^2(1-p)^{n-2}$.

(c) Recall from the discussion of Geometric distributions that $(X = 1)$ is the most likely result for X and $(Y = 1)$ is the most likely result for Y. This does not imply that $(Z = 2)$ is the most likely outcome for Z. Determine the values of p for which $P(Z = 3)$ is larger than $P(Z = 2)$.

Ex. 3.3.8. Let X_1, X_2, X_3, X_4 be an i.i.d. sequence of Bernouli(p) random variables. Let $Y = X_1 + X_2 + X_3 + X_4$. Prove that $P(Y = 2) = 6p^2(1-p)^2$.

Ex. 3.3.9. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sequence of Bernoulli(p) random variables. Let $Y = X_1 + X_2 + \cdots + X_n$. Prove that $Y \sim \text{Binomial}(n, p)$.

Ex. 3.3.10. Let $X_1, X_2, \ldots, X_r$ be an i.i.d. sequence of Geometric (p) random variables. Let $Y = X_1 + X_2 + \cdots + X_r$. Prove that $Y \sim$ Negative Binomial(r, p).

Ex. 3.3.11. Let X_1, X_2, X_3, X_4 be an i.i.d. sequence of Bernoulli(p) random variables. Let $Y = X_1 + X_2$ and let $Z = X_3 + X_4$. Note that Example 3.3.3 guarantees that $Y, Z \sim \text{Binomial}(2, p)$.

(a) Create a chart describing the joint distribution of Y and Z.

(b) Use the chart from (a) to explain why Y and Z are independent.

(c) Explain how you could use Theorem 3.3.6 to reach the conclusion that Y and Z are independent without calculating their joint distribution.

Ex. 3.3.12. Let X_1, X_2, X_3 be an i.i.d. sequence of Bernoulli(p) random variables. Let $Y = X_1 + X_2$ and let $Z = X_2 + X_3$. Note that Example 3.3.3 guarantees that $Y, Z \sim \text{Binomial}(2, p)$.

(a) Create a chart describing the joint distribution of Y and Z.

(b) Use the chart from (a) to explain why Y and Z are not independent.

(c) Explain why the conclusion from (b) is not inconsistant with Theorem 3.3.6.

Ex. 3.3.13. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sequence of discrete random variables and let Z be the maximum of these n variables. Let r be a real number and let $R = P(X_1 \leq r)$. Prove that $P(Z \leq r) = R^n$.

Ex. 3.3.14. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sequence of discrete random variables and let Z be the minimum of these n variables. Let r be a real number and let $R = P(X_1 \leq r)$. Prove that $P(Z \leq r) = 1 - (1 - R)^n$.

Ex. 3.3.15. Let $X \sim \text{Geometric}(p)$ and let $Y \sim \text{Geometric}(q)$ be independent random variables. Let Z be the smaller of X and Y. It is a fact that Z is also geometrically distributed. This problem asks you to prove this fact using two different methods.
METHOD I:

(a) Explain why the event $(Z = n)$ can be written as the disjoint union

$$(Z = n) = (X = n, Y = n) \cup (X = n, Y > n) \cup (X > n, Y = n)$$

(b) Recall from the proof of the memoryless property of geometric random variables that $P(X > m) = \frac{1}{2^m}$. Use this fact and part (a) to prove that

$$P(Z = n) = [(1 - p)(1 - q)]^{n-1}(pq + p(1 - q) + (1 - p)q)$$

(c) Use (b) to conclude that $Z \sim \text{Geometric}(r)$ for some quantity r and calculate the value of r in terms of the p and q.

METHOD II: Recall that geometric random variables first arose from noting the time it takes for a sequence of Bernoulli trials to first produce a success. With that in mind, let $A_1, A_2, \ldots$ be Bernoulli(p) random variables and let $B_1, B_2, \ldots$ be Bernoulli(q) random variables. Further assume the A_j and B_k variables collectively are mutually independent. The variable X may be viewed as the number of the first A_j that produces a result of 1 and the variable Y may be viewed similarly for the B_k sequence.

(a) Let C_j be a random variable that is 1 if either $A_j = 1$ or $B_j = 1$ (or both), and is equal to 0 otherwise. Prove that $C_j \sim \text{Bernoulli}(r)$ for some quantity r and calculate the value of r in terms of p and q.

(b) Explain why the sequence $C_1, C_2, \ldots$ are mutually independent random variables.

(c) Let Z be the random variable that equals the number of the first C_j that results in a 1 and explain why Z is the smaller of X and Y.

(d) Use (c) to conclude that $Z \sim \text{Geometric}(r)$ for the value of r calculated in part (a).

Ex. 3.3.16. Each day during the hatching season along the Odisha and Northern Tamil Nadu coast line a Poisson (λ) number of turtle eggs hatch giving birth to young turtles. As these turtles swim into the sea the probability that they will survive each day is p. Assume

that number of hatchings on each day and the life of the turtles born are all independent. Let $X_1 = 0$ and for $i \geq 2$, X_i be the total number of turtles alive at sea on the i^{th} morning of the hatching season before the hatchings on the i-th day. Find the distribution of X_n.

4

SUMMARIZING DISCRETE RANDOM VARIABLES

When we first looked at Bernoulli trials in Example 2.1.2 we asked the question "On average how many successes will there be after n trials?" In order to answer this question, a specific definition of "average" must be developed.

To begin, consider how to extend the basic notion of the average of a list of numbers to the situation of equally likely outcomes. For instance, if we want to know what the average roll of a die will be, it makes sense to declare it to be 3.5, the average value of $1, 2, 3, 4, 5$, and 6. A motivation for a more general definition of average comes from a rewriting of this calculation.

$$\frac{1+2+3+4+5+6}{6} = 1(\frac{1}{6}) + 2(\frac{1}{6}) + 3(\frac{1}{6}) + 4(\frac{1}{6}) + 5(\frac{1}{6}) + 6(\frac{1}{6}).$$

From the perspective of the right hand side of the equation, the results of all outcomes are added together after being weighted, each according to its probability. In the case of a die, all six outcomes have probability $\frac{1}{6}$.

4.1 EXPECTED VALUE

Definition 4.1.1. *Let $X : S \to T$ be a discrete random variable (so T is countable). Then the expected value (or average) of X is written as $E[X]$ and is given by*

$$E[X] = \sum_{t \in T} t \cdot P(X = t)$$

provided that the sum converges absolutely. In this case we say that X has "finite expectation". If the sum diverges to $\pm\infty$ we say the random variable has infinite expectation. If the sum diverges, but not to infinity, we say the expected value is undefined.

Example 4.1.2. In the previous chapter, Example 3.1.4 described a lottery for which a ticket could be worth nothing, or it could be worth either \$20 or \$200. What is the average value of such a ticket?

We calculated the distribution of ticket values as $P(X = 200) = \frac{1}{1000}$, $P(X = 20) = \frac{27}{1000}$, and $P(X = 0) = \frac{972}{1000}$. Applying the definition of expected value results in

$$E[X] = 200(\frac{1}{1000}) + 20(\frac{27}{1000}) + 0(\frac{972}{1000}) = 0.74,$$

so a ticket has an expected value of 56 cents. ■

It is possible to think of a constant as a random variable. If $c \in \mathbb{R}$ then we could define a random variable X with a distribution such that $P(X = c) = 1$. It is a slight abuse of notation, but in this case we will simply write c for both the real number as well as the constant random variable. Such random variables have the obvious expected value.

Theorem 4.1.3. *Let c be a real number. Then $E[c] = c$.*

Proof - By definition $E[c]$ is a sum over all possible values of c, but in this case that is just a single value, so $E[c] = c \cdot P(c = c) = c \cdot 1 = c$. ■

When the range of X is finite, $E[X]$ always exists since it is a finite sum. When the range of X is infinite there is a possibility that the infinite series will not be absolutely convergent and therefore that $E[X]$ will be infinite or undefined. In fact, when proving theorems about how expected values behave, most of the complications arise from the fact that one must know that an infinite sum converges absolutely in order to rearrange terms within that sum with equality. The next examples explore ways in which expected values may misbehave.

Example 4.1.4. Suppose X is a random variable taking values in the range $T = \{2, 4, 8, 16, \dots\}$ such that $P(X = 2^n) = \frac{1}{2^n}$ for all integers $n \geq 1$.

This is the distribution of a random variable since

$$\sum_{n=1}^{\infty} P(X = 2^n) = \sum_{n=1}^{\infty} \frac{1}{2^n} = 1.$$

But note that

$$\sum_{n=1}^{\infty} 2^n \cdot P(X = 2^n) = \sum_{n=1}^{\infty} 2^n \frac{1}{2^n} = \sum_{n=1}^{\infty} 1$$

which diverges to infinity, so this random variable has an infinite expected value. ■

Example 4.1.5. Suppose X is a random variable taking values in the range $T = \{-2, 4, -8, 16, \dots\}$ such that $P(X = (-2)^n) = \frac{1}{2^n}$ for all integers $n \geq 1$.

$$\sum_{n=1}^{\infty} (-2)^n \cdot P(X = 2^n) = \sum_{n=1}^{\infty} (-2)^n \frac{1}{2^n} = \sum_{n=1}^{\infty} (-1)^n.$$

This infinite sum diverges (not to $\pm\infty$), so the expected value of this random variable is undefined. ■

The examples above were specifically constructed to produce series which clearly diverged, but in general it can be complicated to check whether an infinite sum is absolutely convergent or not. The next technical lemma provides a condition that is often simpler to check. The convenience of this lemma is that, since $|X|$ is always positive, the terms of the series for $E[|X|]$ may be freely rearranged without changing the value of (or the convergence of) the sum.

Lemma 4.1.6. *$E[X]$ is a real number if and only if $E[|X|] < \infty$.*

Proof - Let T be the range of X. So $U = \{|t| : t \in T\}$ is the range of $|X|$. By definition

$$E[|X|] = \sum_{u \in U} u \cdot P(|X| = u), \quad \text{while}$$

$$E[X] = \sum_{t \in T} t \cdot P(X = t).$$

To more easilly relate these two sums, define $\hat{T} = \{t : |t| \in U\}$. Since every $u \in U$ came from some $t \in T$ the new set $\hat{T}$ contains every element of T. For every $t \in \hat{T}$ for which $t \notin T$, the element is outside of the range of X and so $P(X = t) = 0$ for such elements. Because of this $E[X]$ may be written as

$$E[X] = \sum_{t \in \hat{T}} t \cdot P(X = t)$$

since any additional terms in the series are zero.

Note that for each $u \in U$, the event $(|X| = u)$ is equal to $(X = u) \cup (X = -u)$ where each of u and $-u$ is an element of $\hat{T}$. Therefore,

$$\begin{aligned} u \cdot P(U = u) &= u \cdot (P(X = u) + P(X = -u)) \\ &= u \cdot P(X = u) + u \cdot P(X = -u) \\ &= |u| \cdot P(X = u) + |-u| \cdot P(X = -u) \end{aligned}$$

(When $u = 0$ the quantities $P(|X| = 0)$ and $P(X = 0) + P(X = -0)$ are typically not equal, but the equation is still true since both sides of the equation are zero). Summing over all $u \in U$ then yields

$$\begin{aligned}\sum_{u\in U} u \cdot P(|X| = u) &= \sum_{u\in U} |u| \cdot P(X = u) + |-u| \cdot P(X = -u) \\ &= \sum_{t\in \hat{T}} |t| \cdot P(X = t) \\ &= \sum_{t\in T} |t \cdot P(X = t)|.\end{aligned}$$

Therefore the series describing $E[X]$ is absolutely convergent exactly when $E[|X|] < \infty$. ■

4.1.1 *Properties of the Expected Value*

We will eventually wish to calculate the expected values of functions of multiple random variables. Of particular interest to statistics is an understanding of expected values of sums and averages of i.i.d. sequences. That understanding will be made easier by first learning something about how expected values behave for simple combinations of variables.

Theorem 4.1.7. *Suppose that X and Y are discrete random variables, both with finite expected value and both defined on the same sample space S. If a and b are real numbers then*

(1) $E[aX] = aE[X]$;

(2) $E[X + Y] = E[X] + E[Y]$; *and*

(3) $E[aX + bY] = aE[X] + bE[Y]$.

(4) If $X \geq 0$ *then* $E[X] \geq 0$.

Proof of (1) - If $a = 0$ then both sides of the equation are zero, so assume $a \neq 0$. We know that X is a function from S to some range U. So aX is also a random variable and its range is $T = \{au : u \in U\}$.

By definition $E[aX] = \sum_{t \in T} t \cdot P(aX = t)$, but because of how T is defined, adding values indexed by $t \in T$ is equivalent to adding values indexed by $u \in U$ where $t = au$. In other words

$$\begin{aligned} E[aX] &= \sum_{t \in T} t \cdot P(aX = t) \\ &= \sum_{u \in U} au \cdot P(aX = au) \\ &= a \cdot \sum_{u \in U} u \cdot P(X = u) \\ &= aE[X]. \end{aligned}$$

Proof of (2) - We are assuming that X and Y have the same domain, but they typically have different ranges. Suppose $X : S \to U$ and $Y : S \to V$. Then the random variable $X + Y$ is also defined on S and takes values in $T = \{u + v : u \in U, v \in V\}$. Therefore, adding values indexed by $t \in T$ is equivalent to adding values indexed by u and v as they range over U and V respectively. So,

$$\begin{aligned} E[X+Y] &= \sum_{t \in T} t \cdot P(X + Y = t) \\ &= \sum_{u \in U, v \in V} (u+v) \cdot P(X = u, Y = v) \\ &= \sum_{u \in U} \sum_{v \in V} (u+v) \cdot P(X = u, Y = v) \\ &= \sum_{u \in U} \sum_{v \in V} u \cdot P(X = u, Y = v) + \sum_{u \in U} \sum_{v \in V} v \cdot P(X = u, Y = v) \\ &= \sum_{u \in U} \sum_{v \in V} u \cdot P(X = u, Y = v) + \sum_{v \in V} \sum_{u \in U} v \cdot P(X = u, Y = v) \end{aligned}$$

where the rearrangement of summation is legitimate since the series converges absolutely. Notice that as u ranges over all of U the sets $(X = u, Y = v)$ partition the set $(Y = v)$ into disjoint pieces based on the value of X. Likewise the event $(X = u)$ is partitioned by $(X = u, Y = v)$ as v ranges over all values of $v \in V$. Therefore, as a disjoint union,

$$(Y = v) = \bigcup_{u \in U} (X = u, Y = v) \qquad \text{and} \qquad (X = u) = \bigcup_{v \in V} (X = u, Y = v),$$

and so

$$P(Y = v) = \sum_{u \in U} P(X = u, Y = v) \text{ and } P(X = u) = \sum_{v \in V} P(X = u, Y = v).$$

From there the proof may be completed, since

$$\begin{aligned} E[X+Y] &= \sum_{u\in U} u \sum_{v\in V} P(X=u, Y=v) + \sum_{v\in V} v \sum_{u\in U} P(X=u, Y=v) \\ &= \sum_{u\in U} u \cdot P(X=u) + \sum_{v\in V} v \cdot P(Y=v) \\ &= E[X] + E[Y]. \end{aligned}$$

Proof of (3) - This is an easy consequence of (1) and (2). From (2) the expected value $E[aX+bY]$ may be rewritten as $E[aX]+E[bY]$. From there, applying (1) shows this is also equal to $aE[X]+bE[Y]$. (Using induction this theorem may be extended to any finite line ar combination of random variables, a fact which we leave as an exercise below).
Proof of (4) - We know that X is a function from S to T where $t \in T$ implies that $t \geq 0$. As,

$$E[X] = \sum_{t\in T} t \cdot P(X=t),$$

it follows by definition of series (in the case T is countable) that $E[X] \geq 0$. ■

Example 4.1.8. What is the average value of the sum of a pair of dice?

To answer this question by appealing to the definition of expected value would require summing over the eleven possible outcomes $\{2, 3, \ldots, 12\}$ and computing the probabilities of each of those outcomes. Theorem 4.1.7 makes things much simpler. We began this section by noting that a single die roll has an expected value of 3.5. The sum of two dice is $X+Y$ where each of X and Y represents the outcome of a single die. So the average value of the sum of a pair of dice is $E[X+Y] = E[X]+E[Y] = 3.5+3.5 = 7$. ■

Example 4.1.9. Consider a game in which a player might either gain or lose money based on the result. A game is considered "fair" if it is described by a random variable with an expected value of zero. Such a game is fair in the sense that, on average, the player will have no net change in money after playing.

Suppose a particular game is played with one player (the roller) throwing a die. If the die comes up an even number, the roller wins that dollar amount from his opponent. If the die is odd, the roller wins nothing. Obviously the game as stated is not "fair" since the roller cannot lose money and may win something. How much should the roller pay his opponent to play this game in order to make it a fair game?

Let X be the amount of money the rolling player gains by the result on the die. The set of possible outcomes is $T = \{0, 2, 4, 6\}$ and it should be routine at this point to verify that $E[X] = 2$. Let c be the amount of money the roller should pay to play in order to

make the game fair. Since X is the amount of money gained by the roll, the net change of money for the roller is $X - c$ after accounting for how much was paid to play. A fair game requires

$$0 = E[X - c] = E[X] - E[c] = 2 - c.$$

So the roller should pay his opponent \$2 to make the game fair. ■

4.1.2 *Expected Value of a Product*

Theorem 4.1.7 showed that $E[X + Y] = E[X] + E[Y]$. It is natural to ask whether a similar rule exists for the product of variables. While it is not generally the case that the expected value of a product is the product of the expected values, if X and Y happen to be independent, the result is true.

Theorem 4.1.10. *Suppose that X and Y are discrete random variables, both with finite expected value and both defined on the same sample space S. If X and Y are independent, then $E[XY] = E[X]E[Y]$.*

Proof - Suppose $X : S \to U$ and $Y : S \to V$. Then the random variable XY takes values in $T = \{uv : u \in U, v \in V\}$. So,

$$\begin{aligned}
E[XY] &= \sum_{t \in T} t \cdot P(XY = t) \\
&= \sum_{u \in U} \sum_{v \in V} (uv) \cdot P(X = u, Y = v) \\
&= \sum_{u \in U} \sum_{v \in V} (uv) \cdot P(X = u) P(Y = v) \\
&= \sum_{u \in U} u \cdot P(X = u) \sum_{v \in V} v \cdot P(Y = v) \\
&= \left(\sum_{u \in U} u \cdot P(X = u) \right) \left(\sum_{v \in V} v \cdot P(Y = v) \right) \\
&= E[X]E[Y].
\end{aligned}$$

Before showing an example of how this theorem might be used, we provide a demonstration that the result will not typically hold without the assumption of independence.

Example 4.1.11. Let $X \sim \text{Uniform}(\{1, 2, 3\})$ and let $Y = 4 - X$. It is easy to verify $Y \sim \text{Uniform}(\{1, 2, 3\})$ as well, but X and Y are certainly dependent. A routine computation shows $E[X] = E[Y] = 2$, and so $E[X]E[Y] = 4$.

However, the random variable XY can only take on two possible values. It may equal 3 (if either $X = 1$ and $Y = 3$ or vica versa) or it may equal 4 (if $X = Y = 2$). So, $P(XY = 3) = \frac{2}{3}$ and $P(XY = 4) = \frac{1}{3}$. Therefore,

$$E[XY] = 3(\frac{2}{3}) + 4(\frac{1}{3}) = \frac{10}{3} \neq 4.$$

The conclusion of Theorem 4.1.10 fails since X and Y are dependent. ■

Example 4.1.12. Suppose an insurance company assumes that, for a given month, both the number of customer claims X and the average cost per claim Y are independent random variables. Suppose further the company is able to estimate that $E[X] = 100$ and $E[Y] = \$1,250$. How should the company estimate the total cost of all claims that month?

The total cost should be the number of claims times the average cost per claim, or XY. Using Theorem 4.1.10 the expected value of XY is simply the product of the separate expected values.

$$E[XY] = E[X]E[Y] = 100 \cdot \$1,250 = \$125,000.$$

Notice, though, that the assumption of independence played a critical role in this computation. Such an assumption might not be valid for many practical problems. Consider, for example, if a weather event such as a tornado tends to cause both a larger-than-average number of claims and also a larger-than-average value per claim. This could cause the variables X and Y to be dependent and, in such a case, estimating the total cost would not be as simple as taking the product of the separate expected values. ■

4.1.3 *Expected Values of Common Distributions*

A quick glance at the definition of expected value shows that it only depends on the distribution of the random variable. Therefore one can compute the expected values for the various common distributions we defined in the previous chapter.

Example 4.1.13. **(Expected Value of a Bernoulli(p))**

Let $X \sim \text{Bernoulli}(p)$. So $P(X = 0) = 1 - p$ and $P(X = 1) = p$.

Therefore $E[X] = 0(1 - p) + 1(p) = p$. ■

Example 4.1.14. **(Expected Value of a Binomial(n,p))**

We will show two ways to calculate this expected value – the first is more computationally complicated, but follows from the definition of the Binomial distribution directly; the

second is simpler, but requires using the relationship between the Binomial and Bernoulli random variables. In algebraic terms, if $Y \sim \text{Binomial}(n, p)$ then

$$\begin{aligned} E[Y] &= \sum_{k=0}^{n} k \cdot P(Y = k) \\ &= \sum_{k=1}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^{n} k \cdot \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &= np \cdot \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} p^{k-1} (1-p)^{(n-1)-(k-1)} \\ &= np \cdot \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} \\ &= np \cdot \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{(n-1)-k} \end{aligned}$$

where the last equality is a shift of variables. But now, by the binomial theorem, the sum $\sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{(n-1)-k}$ is equal to 1 and therefore $E[Y] = np$.

Alternatively, recall that the Binomial distribution first came about as the total number of successes in n independent Bernoulli trials. Therefore a Binomial(n, p) distribution results from adding together n independent Bernoulli(p) random variables. Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli(p) and let $Y = X_1 + X_2 + \cdots + X_n$. Then $Y \sim \text{Binomial}(n, p)$ and

$$\begin{aligned} E[Y] &= E[X_1 + X_2 + \cdots + X_n] \\ &= E[X_1] + E[X_2] + \cdots + E[X_n] \\ &= p + p + \cdots + p = np. \end{aligned}$$

This also provides the answer to part (d) of Example 2.1.2. The expected number of successes in a series of n independent Bernoulli(p) trials is np. ■

In the next example we will calculate the expected value of a geometric random variable. The computation illustrates a common technique from calculus for simplifying power series by differentiating the sum term-by-term in order to rewrite a complicated series in a simpler way.

EXAMPLE 4.1.15. **(Expected Value of a Geometric(p))**

If $X \sim Geometric(p)$ and $0 < p < 1$, then

$$E[X] = \sum_{k=1}^{\infty} k \cdot p(1-p)^{k-1}$$

To evaluate the sum of the series we will need to work the partial sums of the same. For any $n \geq 1$, let

$$\begin{aligned} T_n &= \sum_{k=1}^{n} kp(1-p)^{k-1} = \sum_{k=1}^{n} k(1-(1-p))(1-p)^{k-1} \\ &= \sum_{k=1}^{n} k(1-p)^{k-1} - \sum_{k=1}^{n} k(1-p)^{k} \\ &= \sum_{k=1}^{n} (1-p)^{k-1} - n(1-p)^n = \frac{1-(1-p)^n}{p} - n(1-p)^n. \end{aligned}$$

Using standard results from analysis we know that for $0 < p < 1$,

$$\lim_{n\to\infty} (1-p)^n = 0 \text{ and } \lim_{n\to\infty} n(1-p)^n = 0.$$

Therefore $T_n \to \frac{1}{p}$ as $n \to \infty$. Hence

$$E[X] = \frac{1}{p}.$$

For instance, suppose we wanted to know on average how many rolls of a die it would take before we observed a 5. Each roll is a Bernoulli trial with a probability $\frac{1}{6}$ of success. The time it takes to observe the first success is distributed as a $Geometric(\frac{1}{6})$ and so has expected value $\frac{1}{1/6} = 6$. On average it should take six rolls before observing this outcome. ■

Example 4.1.16. **(Expected Value of a Poisson(λ))**

We can make a reasonable guess at the expected value of a Poisson(λ) random variable by recalling that such a distribution was created to approximate a Binomial when n was large and p was small. The parameter $\lambda = np$ remained fixed as we took a limit. Since we showed above that a $Binomial(n, p)$ has an expected value of np, it seems plausible that a $Poisson(\lambda)$ should have an expected value of λ. This is indeed true and it is possible to prove the fact by using the idea that the Poisson random variable is the limit of a sequence of Binomial random variables. However, this proof requires an understanding of how limits and expected values interact, a concept that has not yet been introduced in the text. Instead we leave a proof based on a direct algebraic computation as Exercise 4.1.12.

Taking the result as a given, we will illustrate how this expected value might be used for an applied problem. Suppose an insurance company wants to model catastrophic floods using a Poisson(λ) random variable. Since floods are rare in any given year, and since the company is considering what might occur over a long span of years, this may be a reasonable assumption.

As its name implies, a "50-year flood" is a flood so substantial that it should occur, on average, only once every fifty years. However, this is just an average; it may be possible to have two "50-year floods" in consecutive years, though such an event would be quite rare. Suppose the insurance company wants to know how likely it is that there will be two or more "50-year floods" in the next decade, how should this be calculated?

There is an average of one such flood every fifty years, so by proportional reasoning, in the next ten years there should be an average of 0.2 floods. In other words, the number of floods in the next ten years should a random variable $X \sim Poisson(0.2)$ and we wish to calculate $P(X \geq 2)$.

$$\begin{aligned} P(X \geq 2) &= 1 - P(X = 0) - P(X = 1) \\ &= 1 - e^{-0.2} - e^{-0.2}(0.2) \\ &\approx 0.0002. \end{aligned}$$

So assuming the Poisson random variable is an accurate model, there is only about a 0.02% chance that two or more such disastrous floods would occur in the next decade. ■

For a Hypergeometric random variable, we will demonstrate another proof technique common to probability. An expected value may involve a complicated (or infinite) sum which must be computed. However, this sum includes within it the probabilities of each outcome of the random variable, and those probabilities must therefore add to 1. It is sometimes possible to simplify the sum describing the expected value using the fact that a related sum is already known.

EXAMPLE 4.1.17. **(Expected Value of a HyperGeo(N, r, m))** Let m and r be positive integers an d let N be an integer for which $N > \max\{m, r\}$. Let X be a random variable

with $X \sim \text{HyperGeo}(N, r, m)$. To calculate the expected value of X, we begin with two facts. The first is an identity involving combinations. If $n \geq k > 0$ then

$$\begin{aligned}\binom{n}{k} &= \frac{n!}{k!(n-k)!} \\ &= \frac{n}{k}\frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} \\ &= \frac{n}{k}\binom{n-1}{k-1}.\end{aligned}$$

The second comes from the consideration of the probabilities associated with a $\text{HyperGeo}(N-1, r-1, m-1)$ distribution. Specifically, as k ranges over all possible values of such a distribution, we have

$$\sum_k \frac{\binom{r-1}{k}\binom{(N-1)-(r-1)}{(m-1)-k}}{\binom{N-1}{m-1}} = 1$$

since this is the sum over all outcomes of the random variable.

To calculate $E[X]$, let j range over the possible values of X. Recall that the minimum value of j is $\max\{0, m-(N-r)\}$ and the maximum value of j is $\min\{r, m\}$. Now let $k = j-1$. This means that the maximum value for k is $\min\{r-1, m-1\}$. If the minimum value for j was $m-(N-r)$ then the minimum value for k is $m-(N-r)-1 = ((m-1)-((N-1)-(r-1)))$. If the minimum value for j was 0 then the minimum value for k is -1.

The key to the computation is to note that as j ranges over all of the values of X, the values of k cover all possible values of a $\text{HyperGeo}(N-1, m-1, r-1)$ distribution. In fact, the only possible value k may assume that is not in the range of such a distribution is if $k = -1$ as a minimum value. Now,

$$E[X] = \sum_j j \cdot \frac{\binom{r}{j}\binom{N-r}{m-j}}{\binom{N}{m}},$$

and if $j = 0$ is in the range of X, then that term of the sum is zero and it may be deleted without affecting the value. That is equivalent to deleting the $k = -1$ term, so

the remaining values of k exactly describe the range of a HyperGeo$(N-1, m-1, r-1)$ distriubtion. From there, the expected value may be calculated as

$$\begin{aligned} E[X] &= \sum_j j \cdot \frac{\binom{r}{j}\binom{N-r}{m-j}}{\binom{N}{m}} \\ &= \sum_j j \cdot \frac{\frac{r}{j}\binom{r-1}{j-1}\binom{(N-1)-(r-1)}{(m-1)-(j-1)}}{\frac{N}{m}\binom{N-1}{m-1}} \\ &= (\frac{rm}{N}) \cdot \sum_j \frac{\binom{r-1}{j-1}\binom{(N-1)-(r-1)}{(m-1)-(j-1)}}{\binom{N-1}{m-1}} \\ &= (\frac{rm}{N}) \cdot \sum_k \frac{\binom{r-1}{k}\binom{(N-1)-(r-1)}{(m-1)-k}}{\binom{N-1}{m-1}} \\ &= (\frac{rm}{N}) \cdot (1) = \frac{rm}{N}. \end{aligned}$$

This nearly completes the goal of calculating the expected values of hypergeometric distributions. The only remaining issues are the cases when $m = 0$ and $r = 0$. Since the hypergoemetric distribution was only defined when m and r were non-negative integers, and since the proof above requires the consideration of such a distribution for the values $m-1$ and $r-1$, the remaining cases must be handled separately. However, they are fairly easy and yield the same result, a fact we leave it to the reader to verify. ■

4.1.4 *Expected Value of* $f(X_1, X_2, \ldots, X_n)$

As we have seen previously, if X is a random variable and if f is a function defined on the possible outputs of X, then $f(X)$ is a random variable in its own right. The expected value of this new random variable may be computed in the usual way from the distribution of $f(X)$, but it is an extremely useful fact that it may also be computed from the distribution of X itself. The next example and theorems illustrate this fact.

Example 4.1.18. Returning to a setting first seen in Example 3.3.1 we will let $X \sim$ Uniform$(\{-2, -1, 0, 1, 2\})$, and let $f(x) = x^2$. How may $E[f(X)]$ be calculated?

We will demonstrate this in two ways – first by appealing directly to the definition, and then using the distribution of X instead of the distribution of $f(X)$. To use the definition of expected value, recall that $f(X) = X^2$ takes values in $\{0, 1, 4\}$ with the following probabilities: $P(f(X) = 0) = \frac{1}{5}$ while $P(f(X) = 1) = P(f(X) = 4) = \frac{2}{5}$. Therefore,

$$E[f(X)] = 0(\frac{1}{5}) + 1(\frac{2}{5}) + 4(\frac{2}{5}) = 2.$$

However, the values of $f(X)$ are completely determined from the values of X. For instance, the event $(f(X) = 4)$ had a probability of $\frac{2}{5}$ because it was the disjoint union of two other events $(X = 2) \cup (X = -2)$, each of which had probability $\frac{1}{5}$. So the term $4(\frac{2}{5})$ in the computation above could equally well have been thought of in two pieces

$$\begin{aligned} 4 \cdot P(f(X) = 4) &= 4 \cdot P((X = 2) \cup (X = -2)) \\ &= 4 \cdot (P(X = 2) + P(X = -2)) \\ &= 4 \cdot P(X = 2) + 4 \cdot P(X = -2) \\ &= 2^2 \cdot P(X = 2) + (-2)^2 \cdot P(X = -2), \end{aligned}$$

where the final expression emphasizes that the outcome of 4 resulted either from 2^2 or $(-2)^2$ depending on the value of X. Following a similar plan for the other values of $f(X)$ allows $E[f(X)]$ to be calcualted directly from the probabilities of X as

$$\begin{aligned} E[f(X)] &= (-2)^2 \cdot P(X = -2) + (-1)^2 \cdot P(X = -1) + 0^2 \cdot P(X = 0) \\ &\qquad + 1^2 \cdot P(X = 1) + 2^2 \cdot P(X = 2) \\ &= 4(\frac{1}{5}) + 1(\frac{1}{5}) + 0(\frac{1}{5}) + 1(\frac{1}{5}) + 4(\frac{1}{5}) \\ &= 2, \end{aligned}$$

which gives the same result as the previous computation. ■

The technique of the example above works for any functions as demonstrated by the next two theorems. We first state and prove a version for functions of a single random variable and then deal with the multivariate case.

Theorem 4.1.19. *Let $X : S \to T$ be a discrete random variable and define a function $f : T \to U$. Then the expected value of $f(X)$ may be computed as*

$$E[f(X)] = \sum_{t \in T} f(t) \cdot P(X = t).$$

Proof - By definition $E[f(X)] = \sum_{u \in U} u \cdot P(f(X) = u)$. However, as in the previous example, the event $(f(X) = u)$ may be partitioned according to the input values of X which cause $f(X)$ to equal u. Recall that $f^{-1}(u)$ describes the set of values in T which, when input into the function f, produce the value u. That is, $f^{-1}(u) = \{t \in T : f(t) = u\}$. Therefore,

$$(f(X) = u) = \bigcup_{t \in f^{-1}(u)} (X = t), \quad \text{and so}$$

$$P(f(X) = u) = \sum_{t \in f^{-1}(u)} P(X = t).$$

Putting this together with the definition of $E[f(X)]$ shows

$$\begin{aligned} E[f(X)] &= \sum_{u \in U} u \cdot P(f(X) = u) \\ &= \sum_{u \in U} u \cdot \sum_{t \in f^{-1}(u)} P(X = t) \\ &= \sum_{u \in U} \sum_{t \in f^{-1}(u)} u \cdot P(X = t) \\ &= \sum_{u \in U} \sum_{t \in f^{-1}(u)} f(t) \cdot P(X = t) \\ &= \sum_{t \in T} f(t) \cdot P(X = t), \end{aligned}$$

where the final step is simply the fact that $T = f^{-1}(U)$ and so summing over the values of $t \in T$ is equivalent to grouping them together in the sets $f^{-1}(u)$ and summing over all values in U that may be achieved by $f(X)$. ■

Theorem 4.1.20. *Let X_1, X_2, $\ldots X_n$ be random variables defined on a common sample space S. The X_j variables may have different ranges, so let $X_j : S \to T_j$. Let f be a function defined for all possible outputs of the X_j variables. Then*

$$E[f(X)] = \sum_{t_1 \in T_1, \ldots t_n \in T_n} f(t_1, \ldots, t_n) \cdot P(X_1 = t_1, \ldots, X_n = t_n).$$

The proof is nearly the same as for the one-variable case. The only diference is that $f^{-1}(u)$ is now a set of vectors of values $(t_1, \ldots, t_n)$, so that the event $(f(X) = u)$ decomposes into events of the form $(X_1 = t_1, \ldots, X_n = t_n)$. However, this change does not interfere with the logic of the proof. We leave the details to the reader.

EXERCISES

Ex. 4.1.1. Let X, Y be discrete random variables. If $X \leq Y$, then show that $E[X] \leq E[Y]$.

Ex. 4.1.2. A lottery is held every day, and on any given day there is a 30% chance that someone will win, with each day independent of every other. Let X denote the random variable describing the number of times in the next five days that the lottery will be won.

(a) What type of random variable (with what parameter) is X?

(b) On average (expected value), how many times in the next five days will the lottery be won?

(c) When the lottery occurs for each of the next five days, what is the most likely number (mode) of days there will be a winner?

(d) How likely is it the lottery will be won in either one or two of the next five days?

Ex. 4.1.3. A game show contestant is asked a series of questions. She has a probability of 0.88 of knowing the answer to any given question, independently of every other. Let Y denote the random variable describing the number of questions asked until the contestant does not know the correct answer.

(a) What type of random variable (with what parameter) is Y?

(b) On average (expected value), how many questions will be asked until the first question for which the contestant does not know the answer?

(c) What is the most likely number of questions (mode) that will be asked until the contestant does not know a correct answer?

(d) If the contestant is able to answer twelve questions in a row, she will win the grand prize. How likely is it that she will know the answers to all twelve questions?

Ex. 4.1.4. Sonia sends out invitations to eleven of her friends to join her on a hike she's planning. She knows that each of her friends has a 59% chance of deciding to join her independently of each other. Let Z denote the number of friends who join her on the hike.

(a) What type of random variable (with what parameter) is Z?

(b) What is the average (expected value) number of her friends that will join her on the hike?

(c) What is the most likely number (mode) of her friends that will join her on the hike?

(d) How do your answers to (b) and (c) change if each friend has only a 41% chance of joining her?

Ex. 4.1.5. A player rolls three dice and earns \$1 for each die that shows a 6. How much should the player pay to make this a fair game?

Ex. 4.1.6. **("The St.Petersburg Paradox")** Suppose a game is played whereby a player begins flipping a fair coin and continues flipping it until it comes up heads. At that time the player wins a 2^n dollars where n is the total number of times he flipped the coin. Show that there is no amount of money the player could pay to make this a fair game. (Hint: See Example 4.1.4).

Ex. 4.1.7. Two different investment strategies have the following probabilities of return on \$10,000.

Strategy A has a 20% chance of returning \$14,000, a 35% chance of returning \$12,000, a 20% chance of returning \$10,000, a 15% chance of returning \$8,000, and a 10% chance of returning only \$6,000.

Strategy B has a 25% chance of returning \$12,000, a 35% chance of returning \$11,000, a 25% chance of returning \$10,000, and a 15% chance of returning \$9,000.

(a) Which strategy has the larger expected value of return?

(b) Which strategy is more likely to produce a positive return on investment?

(c) Is one strategy clearly preferable to the other? Explain your reasoning.

Ex. 4.1.8. Calculate the expected value of a Uniform$(\{1, 2, \ldots, n\})$ random variable by following the steps below.

(a) Prove the numerical fact that $\sum_{j=1}^{n} j = \frac{n^2+n}{2}$. (Hint: There are many methods to do this. One uses induction).

(b) Use (a) to show that if $X \sim$ Uniform$(\{1, 2, \ldots, n\})$, then $E[X] = \frac{n+1}{2}$.

Ex. 4.1.9. Use induction to extend the result of Theorem 4.1.7 by proving the following:

If $X_1, X_2, \ldots, X_n$ are random variables with finite expectation all defined on the same sample space S and if $a_1, a_2, \ldots a_n$ are real numbers, then

$$E[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = a_1E[X_1] + a_2E[X_2] + \cdots + a_nE[X_n].$$

Ex. 4.1.10. Suppose X and Y are random variables for which X has finitie expected value and Y has infinite expected value. Prove that $X + Y$ has infinite expected value.

Ex. 4.1.11. Suppose X and Y are random variables. Suppose $E[X] = \infty$ and $E[Y] = -\infty$.

(a) Provide an example to show that $E[X + Y] = \infty$ is possible.

(b) Provide an example to show that $E[X + Y] = -\infty$ is possible.

(c) Provide an example to show that $E[X+Y]$ may have finite expected value.

Ex. 4.1.12. Let $X \sim Poisson(\lambda)$.

(a) Write an expression for $E[X]$ as an infinite sum.

(b) Every non-zero term in your answer to (a) should have a λ in it. Factor this λ out and explain why the remaining sum equals 1. (Hint: One way to do this is through the use of infinite series. Another way is to use the idea from Example 4.1.17).

Ex. 4.1.13. A daily lottery is an event that many people play, but for which the likelihood of any given person winning is very small, making a Poisson approximation appropriate. Suppose a daily lottery has, on average, two winners every five weeks. Estimate the probability that next week there will be more than one winner.

4.2 VARIANCE AND STANDARD DEVIATION

As a single number, the average of a random variable may or may not be a good approximation of the values that variable is likely to produce. For example, let X be defined such that $P(X=10)=1$, let Y be defined so that $P(Y=9)=P(Y=11)=\frac{1}{2}$, and let Z be defined such that $P(Z=0)=P(Z=20)=\frac{1}{2}$. It is easy to check that all three of these random variables have an expected value of 10. However the number 10 exactly describes X, is always off from Y by an absolute value of 1 and is always off from Z by an absolute value of 10.

It is useful to be able to quantify how far away a random variable typically is from its average. Put another way, if we think of the expected value as somehow measuring the "center" of the random variable, we would like to find a way to measure the size of the "spread" of the variable about its center. Quantities useful for this are the variance and standard deviation.

Definition 4.2.1. *Let X be a random variable with finite expected value. Then the variance of the random variable is written as $Var[X]$ and is defined as*

$$Var[X] = E[(X - E[X])^2]$$

The standard deviation of X is written as $SD[X]$ and is defined as

$$SD[X] = \sqrt{Var[X]}$$

Notice that $Var[X]$ is the average of the square distance of X from its expected value. So if X has a high probability of being far away from $E[X]$ the variance will tend to be large, while if X is very near $E[X]$ with high probability the variance will tend to be small. In either case the variance is the expected value of a squared quantity, and as such is always non-negative. Therefore $SD[X]$ is defined whenever $Var[X]$ is defined.

If we were to associate units with the random variable X (say *meters*), then the units of $Var[X]$ would be $meters^2$ and the units of $SD[X]$ would be *meters*. We will see that the standard deviation is more meaningful as a measure of the "spread" of a random variable while the variance tends to be a more useful quantity to consider when carrying out complex computations.

Informally we will view the standard deviation as a typical distance from average. So if X is a random variable and we calculate that $E[X] = 12$ and $SD[X] = 3$, we might say, "The variable X will typically take on values that are in or near the range $9 - 15$, one standard deviation either side of the average". A goal of this section is to make that language more precise, but at this point it will help with intuition to understand this informal view.

The variance and standard deviation are described in terms of the expected value. Therefore $Var[X]$ and $SD[X]$ can only be defined if $E[X]$ exists as a real number. However, it is possible that $Var[X]$ and $SD[X]$ could be infinite even if $E[X]$ is finite (see Exercises). In practical terms, if X has a finite expected value and infinite standard deviation, it means that the random variable has a clear average, but is so spread out that any finite number underestimates the typical distance of the random variable from its average.

Example 4.2.2. As above, let X be a constant varaible with $P(X = 10) = 1$. Let Y be such that $P(Y = 9) = P(Y = 11) = \frac{1}{2}$ and let Z be such that $P(Z = 0) = P(Z = 20) = \frac{1}{2}$.

Since X always equals $E[X]$, the quantity $(X - E[X])^2$ is always zero and we can conclude that $Var[X] = 0$ and $SD[X] = 0$. This makes sense given the view of $SD[X]$ as an estimate of how spread out the variable is. Since X is constant it is not at all spread out and so $SD[X] = 0$.

To calculate $Var[Y]$ we note that $(Y - E[Y])^2$ is always equal to 1. Therefore $Var[Y] = 1$ and $SD[Y] = 1$. Again this reaffirms the informal description of the standard deviation; the typical distance between Y and its average is 1.

Likewise $(Z - E[Z])^2$ is always equal to 100. Therefore $Var[Z] = 100$ and $SD[Z] = 10$. The typical distance between Z and its average is 10. ■

Example 4.2.3. What are the variance and standard deviation of a die roll?

Before we carry out the calculation, let us use the informal idea of standard deviation to estimate an answer and help build intuition. We know the average of a die roll is 3.5. The closest a die could possibly be to this average is 0.5 (if it were to roll a 3 or a 4) and

the furthest it could possibly be is 2.5 (if it were to roll a 1 or a 6). Therefore the standard deviation, a typical distance from average, should be somewhere between 0.5 and 2.5.

To calculate the quantity exactly, let X represent the roll of a die. By definition, $Var[X] = E[(X-3.5)^2]$, and the values that $(X-3.5)^2$ may assume are determined by the six values X may take on.

$$\begin{aligned} Var[X] &= E[(X-3.5)^2] \\ &= \frac{1}{6}(2.5)^2 + \frac{1}{6}(1.5)^2 + \frac{1}{6}(0.5)^2 + \frac{1}{6}(-0.5)^2 + \frac{1}{6}(-1.5)^2 + \frac{1}{6}(-2.5)^2 \\ &= \frac{35}{12}. \end{aligned}$$

So, $SD[X] = \sqrt{\frac{35}{12}} \approx 1.71$ which is near the midpoint of the range of our estimate above. ■

4.2.1 *Properties of Variance and Standard Deviation*

Theorem 4.2.4. *Let $a \in \mathbb{R}$ and let X be a random variable with finite variance (and thus, with finite expected value as well). Then,*

(a) $Var[aX] = a^2 \cdot Var[X]$;

(b) $SD[aX] = |a| \cdot SD[X]$;

(c) $Var[X+a] = Var[X]$; *and*

(d) $SD[X+a] = SD[X]$.

Proof of (a) and (b) - $Var[aX] = E[(aX - E[aX])^2]$. Using known properties of expected value this may be rewritten as

$$\begin{aligned} Var[aX] &= E[(aX - aE[X])^2] \\ &= E[a^2(X-E[X])^2] \\ &= a^2 E[(X-E[X])^2] \\ &= a^2 Var[X]. \end{aligned}$$

That concludes the proof of (a). The result from (b) follows by taking square roots of both sides of this equation.

Proof of (c) and (d) - (See Exercises) ■

The variance may also be computed using a different (but equivalent) formula if $E[X]$ and $E[X^2]$ are known.

Theorem 4.2.5. *Let X be a random variable for which $E[X]$ and $E[X^2]$ are both finite. Then*

$$Var[X] = E[X^2] - (E[X])^2 .$$

Proof -

$$\begin{aligned} Var[X] &= E[(X - E[X])^2] \\ &= E[X^2 - 2XE[X] + (E[X])^2] \\ &= E[X^2] - 2E[XE[X]] + E[(E[X])^2]. \end{aligned}$$

But $E[X]$ is a constant, so

$$\begin{aligned} Var[X] &= E[X^2] - 2E[XE[X]] + E[(E[X])^2] \\ &= E[X^2] - 2E[X]E[X] + (E[X])^2 \\ &= E[X^2] - (E[X])^2 . \end{aligned}$$

■

In statistics we frequently want to consider the sum or average of many random variables. As such it is useful to know how the variance of a sum relates to the variances of each variable separately. Toward that goal we have

Theorem 4.2.6. *If X and Y are independent random variables, both with finite expectation and finite variance, then*

(a) $Var[X+Y] = Var[X] + Var[Y]$; *and*

(b) $SD[X+Y] = \sqrt{(SD[X])^2 + (SD[Y])^2}$.

Proof - Using Theorem 4.2.5,

$$\begin{aligned} Var[X+Y] &= E[(X+Y)^2] - (E[X+Y])^2 \\ &= E[X^2 + 2XY + Y^2] - \left((E[X])^2 + 2E[X]E[Y] + (E[Y])^2\right) \\ &= E[X^2] + 2E[XY] + E[Y^2] - (E[X])^2 - 2E[X]E[Y] - (E[Y])^2. \end{aligned}$$

But by Theorem 4.1.10, $E[XY] = E[X]E[Y]$ since X and Y are independent. So,

$$\begin{aligned} Var[X+Y] &= E[X^2] - (E[X])^2 + E[Y^2] - (E[Y])^2 \\ &= Var[X] + Var[Y]. \end{aligned}$$

Part (b) follows immediately after rewriting the variances in terms of standard deviations and taking square roots. As with expected values, this theorem may be generalized to a sum of any finite number of independent random variables using induction. The proof of that fact is left as Exercise 4.2.11. ■

Example 4.2.7. What is the standard deviation of the sum of two dice?

We previously found that if X represents one die, then $Var[X] = \frac{35}{12}$. If X and Y are two independent dice, then $Var[X+Y] = Var[X] + Var[Y] = \frac{35}{12} + \frac{35}{12} = \frac{35}{6}$. Therefore $SD[X+Y] = \sqrt{\frac{35}{6}} \approx 2.42$. ■

4.2.2 *Variances of Common Distributions*

As with expected value, the variances of the common discrete random variables can be calculated from their corresponding distributions.

Example 4.2.8. **(Variance of a Bernoulli(p))**

Let $X \sim Bernoulli(p)$. We have already calculated that $E[X] = p$. Since X only takes on the values 0 or 1 it is always true that $X^2 = X$. Therefore $E[X^2] = E[X] = p$.

So, $Var[X] = E[X^2] - (E[X])^2 = p - p^2 = p(1-p)$. ■

Example 4.2.9. **(Variance of a Binomial(n,p))**

We will calculate the variance of a Binomial random variable using the fact that it may be viewed as the sum of n independent Bernoulli random variables. A strictly algebraic computation is also possible (see Exercises).

Let $X_1, X_2, \ldots, X_n$ be independent $Bernoulli(p)$ random variables. Therefore, if $Y = X_1 + X_2 + \cdots + X_n$ then $Y \sim$ Binomial(n,p) and

$$\begin{aligned} Var[Y] &= Var[X_1 + X_2 + \cdots + X_n] \\ &= Var[X_1] + Var[X_2] + \cdots + Var[X_n] \\ &= p(1-p) + p(1-p) + \cdots + p(1-p) \\ &= np(1-p). \end{aligned}$$

For an application of this computation we return to the idea of sampling from a population where some members of the population have a certain characteristic and others

do not. The goal is to provide an estimate of the number of people in the sample that have the characteristic. For this example, suppose we were to randomly select 100 people from a large city in which 20% of the population works in a service industry. How many of the 100 people from our sample should we expect to be service industry workers?

If the sampling is done without replacement (so we cannot pick the same person twice), then strictly speaking the desired number would be described by a Hypergeometric random variable. However, we have also seen that there is little difference between the Binomial and Hypergeometric distributions when the size of the sample is small relative to the size of the population. So since the sample is only 100 people from a "large city", we will assume this situation is modeled by a binomial random variable. Specifically, since 20% of the population consits of service workers, we will assume $X \sim Binomial(100, 0.2)$.

The simplest way to answer to the question of how many service industy workers to expect within the sample is to compute the expected value of X. In this case $E[X] = 100(0.2) = 20$, so we should expect around 20 of the 100 people in the sample to be service workers. However, this is an incomplete answer to the question since it only provides an average value; the actual number of service workers in the sample is probably not going to be exactly 20, it's only likely to be around 20 on average. A more complete answer to the question would give an estimate as to how far away from 20 the actual value is likely to be. But this is precisely what the standard deviation describes – an estimate of the likely difference between the actual result of the random variable and its expected value.

In this case $Var[X] = 100(0.2)(0.8) = 16$ and so $SD[X] = \sqrt{16} = 4$. This means that the actual number of service industry workers in the sample will typically be about 4 or so away from the expected value of 20, so a more complete answer to the question would be "The sample is likely to have around $16 - 24$ service workers in it". That is not to say that the actual number of service workers is guaranteed to fall in the that range, but the range provides s a sort of likely error associated with the estimate of 20. Results in the $16 - 24$ range should be considered fairly common. Results far outside that range, while possible, should be considered fairly unusual. ■

Recall in Example 4.1.17 we calculated $E[X]$ using a technique in which the sum describing $E[X]$ was computed based on another sum which only involved the distribution of X directly. This second sum equalled 1 since it simply added up the probabilities that X assumed each of its possible values. In a similar fashion, it is sometimes possible to calculate a sum describing $E[X^2]$ in terms of a sum for $E[X]$ which is already known. From that point, Theorem 4.2.5 may be used to calculate the variance and standard deviation of X. This technique will be illustrated in the next example in which we calculate the spread associated with a geometric random variable.

Example 4.2.10. **(Variance of a Geometric(p))**

Let $0 < p < 1$. $X \sim Geometric(p)$ for which we know $E[X] = \frac{1}{p}$. Then,

$$E[X^2] = \sum_{k=1}^{\infty} k^2 p(1-p)^{k-1}$$

To evaluate the sum of the series we will need to work the partial sums of the same. For any $n \geq 1$, let

$$\begin{aligned}
S_n &= \sum_{k=1}^{n} k^2 p(1-p)^{k-1} = \sum_{k=1}^{n} k^2 (1-(1-p))(1-p)^{k-1} \\
&= \sum_{k=1}^{n} k^2 (1-p)^{k-1} - \sum_{k=1}^{n} k^2 (1-p)^k \\
&= 1 + \sum_{k=2}^{n} (2k-1)(1-p)^{k-1} - n^2(1-p)^n \\
&= 1 - \sum_{k=2}^{n} (1-p)^{k-1} + 2\sum_{k=2}^{n} k(1-p)^{k-1} - n^2(1-p)^n \\
&= 2 - \sum_{k=1}^{n} (1-p)^{k-1} + 2(-1 + \sum_{k=1}^{n} k(1-p)^{k-1}) - n^2(1-p)^n \\
&= -\frac{1-(1-p)^n}{p} + \frac{2}{p}\sum_{k=1}^{n} kp(1-p)^{k-1} - n^2(1-p)^n
\end{aligned}$$

Using standard results from analysis and result from Example 4.1.15 we know that for $0 < p < 1$,

$$\lim_{n\to\infty} \sum_{k=1}^{n} kp(1-p)^{k-1} = \frac{1}{p}, \lim_{n\to\infty} (1-p)^n = 0, \text{ and } \lim_{n\to\infty} n^2(1-p)^n = 0.$$

Therefore $S_n \to -\frac{1}{p} + \frac{2}{p^2}$ as $n \to \infty$. Hence

$$E[X^2] = -\frac{1}{p} + \frac{2}{p^2}.$$

Using Theorem 4.2.5 the variance may then be calculated as

$$\begin{aligned}
Var[X] &= E[X^2] - (E[X])^2 \\
&= \frac{2}{p^2} - \frac{1}{p} - (\frac{1}{p})^2 \\
&= \frac{1}{p^2} - \frac{1}{p}
\end{aligned}$$

■

A similar technique may be used for calculating the variance of a Poisson random variable, a fact which is left as an exercise. We finish this subsection with a computation of the variance of a hypergeometric distribution using an idea similar to how we calculated its expected value in Example 4.1.17.

Example 4.2.11. Let m and r be positive integers and let N be an integer with $N > \max\{m, r\}$ and let $X \sim \text{HyperGeo}(N, r, m)$. To calculate $E[X^2]$, as j ranges over the values of X,

$$\begin{aligned} E[X^2] &= \sum_j j^2 \cdot \frac{\binom{r}{j}\binom{N-r}{m-j}}{\binom{N}{m}} \\ &= \sum_j j^2 \cdot \frac{\frac{r}{j}\binom{r-1}{j-1}\binom{(N-1)-(r-1)}{(m-1)-(j-1)}}{\frac{N}{m}\binom{N-1}{m-1}} \\ &= (\frac{rm}{N}) \sum_j j \cdot \frac{\binom{r-1}{j-1}\binom{(N-1)-(r-1)}{(m-1)-(j-1)}}{\binom{N-1}{m-1}} \\ &= (\frac{rm}{N}) \cdot \sum_k (k+1) \frac{\binom{r-1}{k}\binom{(N-1)-(r-1)}{(m-1)-k}}{\binom{N-1}{m-1}} \end{aligned}$$

where k ranges over the values of $Y \sim \text{HyperGeo}(N-1, r-1, m-1)$. Therefore,

$$\begin{aligned} E[X^2] &= (\frac{rm}{N})E[Y+1] \\ &= (\frac{rm}{N})(E[Y]+1) \\ &= (\frac{rm}{N})(\frac{(r-1)(m-1)}{(N-1)}+1). \end{aligned}$$

Now the variance may be easily computed as

$$\begin{aligned} Var[X] &= E[X^2] - (E[X])^2 \\ &= (\frac{rm}{N})(\frac{(r-1)(m-1)}{(N-1)}+1) - (\frac{rm}{N})^2 \\ &= \frac{N^2rm - Nrm^2 - Nr^2m + r^2m^2}{N^2(N-1)}. \end{aligned}$$

As with the computation of expected value, the cases of $m = 0$ and $r = 0$ must be handled separately, but yield the same result. ■

4.2.3 *Standardized Variables*

Many random variables may be rescaled into a standard format by shifting them so that they have an average of zero and then rescaling them so that they have a variance (and standard deviation) of one. We introduce this idea now, though its chief importance will not be realized until later.

Definition 4.2.12. *A standardized random variable X is one for which*

$$E[X] = 0 \quad \textit{and} \quad Var[X] = 1.$$

Theorem 4.2.13. *Let X be a discrete random variable with finite expected value and finite, non-zero variance. Then $Z = \frac{X - E[X]}{SD[X]}$ is a standardized random variable.*

Proof - The expected value value of Z is

$$\begin{aligned} E[Z] &= E[\frac{X - E[X]}{SD[X]}] \\ &= \frac{E[X - E[X]]}{SD[X]} \\ &= \frac{E[X] - E[X]}{SD[X]} = 0 \end{aligned}$$

while the variance of Z is

$$\begin{aligned} Var[Z] &= Var[\frac{X - E[X]}{SD[X]}] \\ &= \frac{Var[X - E[X]]}{(SD[X])^2} \\ &= \frac{Var[X]}{Var[X]} = 1. \end{aligned}$$

For easy reference we finish off this section by providing a chart of values associated with common discrete distributions.

Distribution	Expected Value	Variance
Bernoulli(p)	p	$p(1-p)$
Binomial(n,p)	np	$np(1-p)$
Geometric(p)	$\frac{1}{p}$	$\frac{1-p}{p^2}$
HyperGeo(N,r,m)	$\frac{rm}{N}$	$\frac{N^2rm-Nrm^2-Nr^2m+r^2m^2}{N^2(N-1)}$
Poisson(λ)	λ	λ
Uniform($\{1,2,\ldots,n\}$)	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$

EXERCISES

Ex. 4.2.1. A random variable X has a probability mass function given by

$$P(X=0)=0.2, P(X=1)=0.5, P(X=2)=0.2, \text{ and } P(X=3)=0.1.$$

Calculate the expected value and standard deviation of this random variable. What is the probability this random variable will produce a result more than one standard deviation from its expected value?

Ex. 4.2.2. Answer the following questions about flips of a fair coin.

(a) Calculate the standard deviation of the number of heads that show up in 100 flips of a fair coin.

(b) Show that if the number of coins is quadrupled (to 400) the standard deviation only doubles.

Ex. 4.2.3. Suppose we begin rolling a die, and let X be the number of rolls needed before we see the first 3.

(a) Show that $E[X]=6$.

(b) Calculate $SD[X]$.

(c) Viewing $SD[X]$ as a typical distance of X from its expected value, would it seem unusual to roll the die more than nine times before seeing a 3?

(d) Calculate the actual probability $P(X>9)$.

(e) Calculate the probability X produces a result within one standard deviation of its expected value.

Ex. 4.2.4. A key issue in statistical sampling is the determination of how much a sample is likely to differ from the population it came from. This exercise explores some of these ideas.

(a) Suppose a large city is exactly 50% women and 50% men and suppose we randomly select 60 people from this city as part of a sample. Let X be the number of women in the sample. What are the expected value and standard deviation of X? Given these values, would it seem unusual if fewer than 45% of the individuals in the sample were women?

(b) Repeat part (a), but now assume that the sample consists of 600 people.

Ex. 4.2.5. Calculate the variance and standard deviation of the value of the lottery ticket from Example 3.1.4.

Ex. 4.2.6. Prove parts (c) and (d) of Theorem 4.2.4.

Ex. 4.2.7. Let $X \sim Binomial(n, p)$. Show that for $0 < p < 1$, this random variable has the largest standard deviation when $p = \frac{1}{2}$.

Ex. 4.2.8. Follow the steps below to calculate the variance of a random variable with a Uniform$(\{1, 2, \ldots, n\})$ distribution.

(a) Prove that $\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$. (Induction is one way to do this).

(b) Let $X \sim$ Uniform$(\{1, 2, \ldots, n\})$. Use (a) to calculate $E[X^2]$.

(c) Use (b) and the fact that $E[X] = \frac{n+1}{2}$ to calculate $Var[X]$.

Ex. 4.2.9. This exercise provides an example of a random variable with finite expected value, but infinite variance. Let X be a random variable for which $P(X = \frac{2^n}{n(n+1)}) = \frac{1}{2^n}$ for all integers $n \geq 1$.

(a) Prove that X is a well-defined variable by showing $\sum_{n=1}^{\infty} P(X = \frac{2^n}{n(n+1)}) = 1$.

(b) Prove that $E[X] = 1$.

(c) Prove that $Var[X]$ is infinite.

Ex. 4.2.10. Recall that the hypergeometric distribution was first developed to answer questions about sampling without replacement. With that in mind, answer the following questions using the chart of expected values and variances.

(a) Use the formula in the chart to calculate the variance of a hypergeometric distribution if $m = 0$. Explain this result in the context of what it means in terms of sampling.

(b) Use the formula in the chart to calculate the variance of a hypergeometric distribution if $r = 0$. Explain this result in the context of what it means in terms of sampling.

(c) Though we only defined a hypergeometric distrbiution if $N > \max\{r, m\}$, the definition could be extended to $N = \max\{r, m\}$. Use the chart to calculate the variance of a hypergeometric distribution if $N = m$. Explain this result in the context of what it means in terms of sampling without replacement.

Ex. 4.2.11. Prove the following facts about independent random variables.

(a) Use Theorem 4.2.6 and induction to prove that if X_1, X_2, ..., X_n are independent, then

$$Var[X_1 + \cdots + X_n] = Var[X_1] + \cdots + Var[X_n].$$

(b) Suppose X_1, X_2, ..., X_n are i.i.d. Prove that

$$E[X_1 + \cdots + X_n] = n \cdot E[X_1] \quad \text{and} \quad SD[X_1 + \cdots + X_n] = \sqrt{n} \cdot SD[X_1].$$

(c) Suppose X_1, X_2, ..., X_n are mutually independent standardized random variables (not necessarilly identically distributed). Let

$$Y = \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}}.$$

Prove that Y is a standardized random variable.

Ex. 4.2.12. Let X be a discrete random variable which takes on only non-negative values. Show that if $E[X] = 0$ then $P(X = 0) = 1$.

Ex. 4.2.13. Suppose X is a discrete random variable with finite variance (and thus finite expected value as well) and suppose there are two different numbers $a, b \in \mathbb{R}$ for which $P(X = a)$ and $P(X = b)$ are both positive. Prove that $Var[X] > 0$.

Ex. 4.2.14. Let X be a discrete random variable with finite variance (and thus finite expected value as well).

(a) Prove that $E[X^2] \geq (E[X])^2$.

(b) Suppose there are two different numbers $a, b \in \mathbb{R}$ for which $P(X = a)$ and $P(X = b)$ are both positive. Prove that $E[X^2] > (E[X])^2$.

Ex. 4.2.15. Let $X \sim \text{Binomial}(n, p)$ for $n > 1$ and $0 < p < 1$. Using the steps below, provide an algebraic proof of the fact that $Var[X] = np(1 - p)$ without appealing to the fact that such a variable is the sum of Bernoulli trials.

(a) Begin by writing an expression for $E[X^2]$ in summation form.

(b) Use (a) to show that $E[X^2] = np \cdot \sum_{k=0}^{n-1} (k+1)\binom{n-1}{k} p^k (1-p)^{(n-1)-k}$.

(c) Use (b) to explain why $E[X^2] = np \cdot E[Y+1]$ where $Y \sim \text{Binomial}(n-1, p)$.

(d) Use (c) together with Theorem 4.2.5 to prove that $Var[X] = np(1-p)$.

4.3 STANDARD UNITS

When there is no confusion about what random variable is being discussed, it is usual to use the Greek letter μ in place of $E[X]$ and σ in place of $SD[X]$. When more than one variable is involved the same letters can be used with subscripts (μ_X and σ_X) to indicate which variable is being described.

In statistics one frequently measures results in terms of "standard units" – the number of standard deviations a result is from its expected value. For instance if $\mu = 12$ and $\sigma = 5$, then a result of $X = 20$ would be 1.6 standard units because $20 = \mu + 1.6\sigma$. That is, 20 is 1.6 standard deviations above expected value. Similarly a result of $X = 10$ would be -0.4 standard units because $10 = \mu - 0.4\sigma$.

Since the standard deviation measures a typical distance from average, results that are within one standard deviation from average (between -1 and $+1$ standard units) will tend to be fairly common, while results that are more than two standard deviations from average (less than -2 or greater than $+2$ in standard units) will usually be relatively rare. The likelihoods of some such events will be calculated in the next two examples. Notice that the event $(|X - \mu| \leq k\sigma)$ describes those outcomes of X that are within k standard deviations from average.

Example 4.3.1. Let Y represent the sum of two dice. How likely is it that Y will be within one standard deviation of its average? How likely is it that Y will be more than two standard deviations from its average?

We can use our previous calculations that $\mu = 7$ and $\sigma = \sqrt{\frac{35}{6}} \approx 2.42$. The achievable values that are within one standard deviation of average are $5, 6, 7, 8$, and 9. So the probability that the sum of two dice will be within one standard deviation of average is

$$\begin{aligned} P(|Y - \mu| \leq \sigma) &= P(Y \in \{5,6,7,8,9\}) \\ &= \frac{4}{36} + \frac{5}{36} + \frac{6}{36} + \frac{5}{36} + \frac{4}{36} \\ &= \frac{2}{3}. \end{aligned}$$

There is about a 66.7% chance that a pair of dice will fall within one standard deviation of their expected value.

Two standard deviations is $2\sqrt{\frac{35}{6}} \approx 4.83$. Only the results 2 and 12 further than this distance from the expected value, so the probability that X will be more than two standard deviations from average is

$$\begin{aligned} P(|Y - \mu| > 2\sigma) &= P(Y \in \{2, 12\}) \\ &= \frac{2}{36} \approx 0.056. \end{aligned}$$

There is only about a 5.6% chance that a pair of dice will be more than two standard deviations from expected value. ■

Example 4.3.2. If $X \sim Uniform\{(1, 2, \ldots, 100)\}$, what is the probability that X will be within one standard deviation of expected value? What is the probability it will be more than two standard deviations from expected value?

Again, based on earlier calculations we know that $\mu = \frac{101}{2} = 50.5$ and that $\sigma = \sqrt{\frac{9999}{12}} \approx 28.9$. Of the possible values that X can achieve, only the numbers $22, 23, \ldots, 79$ fall within one standard deviation of average. So the desired probability is

$$\begin{aligned} P(|X - \mu| \leq \sigma) &= P(X \in \{22, 23, \ldots, 79\}) \\ &= \frac{58}{100}. \end{aligned}$$

There is a 58% chance that this random variable will be within one standard deviation of expected value.

Similarly we can calculate that two standard deviations is $2\sqrt{\frac{9999}{12}} \approx 57.7$. Since $\mu = 50.5$ and since the minimal and maximal values of X are 1 and 100 respectively, results that are more than two or more standard deviations from average cannot happen at all for this random variable. In other words $P(|X - \mu| > 2\sigma) = 0$. ■

4.3.1 *Markov and Chebychev Inequalities*

The examples of the previous section show that the exact probabilities a random variable will fall within a certain number of standard deviations of its expected value depend on the distribution of the random variable. However, there are some general results that apply to all random variables. To prove these results we will need to investigate some inequalities.

Theorem 4.3.3. (Markov's Inequality) *Let X be a discrete random variable which takes on only non-negative values and suppose that X has a finite expected value. Then for any $c > 0$,*

$$P(X \geq c) \leq \frac{\mu}{c}.$$

Proof - Let T be the range of X, so T is a countable subset of the positive real numbers. By dividing T into those numbers smaller than c and those numbers that are at least as large as c we have

$$\begin{aligned} \mu &= \sum_{t \in T} t \cdot P(X = t) \\ &= \sum_{t \in T, t < c} t \cdot P(X = t) + \sum_{t \in T, t \geq c} t \cdot P(X = t). \end{aligned}$$

The first sum must be non-negative, since we assumed that T consisted of only non-negative numbers, so we only make the quantity smaller by deleting it. Likewise, for each term in the second sum, $t \geq c$ so we only make the quantity smaller by replacing t by c. This gives us

$$\begin{aligned} \mu &= \sum_{t \in T, t < c} t \cdot P(X = t) + \sum_{t \in T, t \geq c} t \cdot P(X = t) \\ &\geq \sum_{t \in T, t \geq c} c \cdot P(X = t) \\ &= c \cdot \sum_{t \in T, t \geq c} P(X = t). \end{aligned}$$

The events $(X = t)$ indexed over all values $t \in T$ for which $t \geq c$ are a countable collection of disjoint sets whose union is $(X \geq c)$. So,

$$\begin{aligned} \mu &\geq c \cdot \sum_{t \in T, t \geq c} P(X = t) \\ &= cP(X \geq c). \end{aligned}$$

Dividing by c gives the desired result.

Markov's theorem can be useful in its own right for producing an upper bound on the liklihood of certain events, but for now we will use it simply as a lemma to prove our next result.

Theorem 4.3.4. (Chebychev's Inequality) *Let X be a discrete random variable with finite, non-zero variance. Then for any $k > 0$,*

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Proof - The event $(|X - \mu| \geq k\sigma)$ is the same as the event $((X - \mu)^2 \geq k^2\sigma^2)$. The random variable $(X - \mu)^2$ is certainly non-negative and its expected value is the variance of X which we have assumed to be finite. Therefore we may apply Markov's inequality to $(X - \mu)^2$ to get

$$\begin{aligned} P(|X - \mu| \geq k\sigma) &= P((X - \mu)^2 \geq k^2\sigma^2) \\ &\leq \frac{E[(X - \mu)^2]}{k^2\sigma^2} \\ &= \frac{Var[X]}{k^2\sigma^2} \\ &= \frac{\sigma^2}{k^2\sigma^2} \\ &= \frac{1}{k^2}. \end{aligned}$$

Though the theorem is true for all $k > 0$, it doesn't give any useful information unless $k > 1$.

Example 4.3.5. Let X be a discrete random variable. Find an upper bound on the likelihood that X will be more than two standard deviations from its expected value.

For the question to make sense we need to assume that X has finite variance to begin with. In which case we may apply Chebychev's inequality with $k = 2$ to find that

$$P(|X - \mu| > 2\sigma) \leq P(|X - \mu| \geq 2\sigma) \leq \frac{1}{4}.$$

There is at most a 25% chance that a random variable will be more than two standard deviations from its expected value. ■

EXERCISES

Ex. 4.3.1. Let $X \sim Binomial(10, \frac{1}{2})$.

(a) Calculate μ and σ.

(b) Calculate $P(|X - \mu| \leq \sigma)$, the probability that X will be within one standard deviation of average. Approximate your answer to the nearest tenth of a percent.

(c) Calculate $P(|X - \mu| > 2\sigma)$, the probability that X will be more than two standard deviations from average. Approximate your answer to the nearest tenth of a percent.

Ex. 4.3.2. Let $X \sim Geometric(\frac{1}{4})$.

(a) Calculate μ and σ.

(b) Calculate $P(|X - \mu| \leq \sigma)$, the probability that X will be within one standard deviation of average. Approximate your answer to the nearest tenth of a percent.

(c) Calculate $P(|X - \mu| > 2\sigma)$, the probability that X will be more than two standard deviations from average. Approximate your answer to the nearest tenth of a percent.

Ex. 4.3.3. Let $X \sim Poisson(3)$.

(a) Calculate μ and σ.

(b) Calculate $P(|X - \mu| \leq \sigma)$, the probability that X will be within one standard deviation of average. Approximate your answer to the nearest tenth of a percent.

(c) Calculate $P(|X - \mu| > 2\sigma)$, the probability that X will be more than two standard deviations from average. Approximate your answer to the nearest tenth of a percent.

Ex. 4.3.4. Let $X \sim Binomial(n, \frac{1}{2})$. Determine the smallest value of n for which $P(|X - \mu| > 4\sigma) > 0$. That is, what is the smallest n for which there is a positive probability that X will be more than four standard deviations from average.

Ex. 4.3.5. For $k \geq 1$ there are distributions for which Chebychev's inequality is an equality.

(a) Let X be a random variable with probability mass function $P(X = 1) = P(X = -1) = \frac{1}{2}$. Prove that Chebychev's inequality is an equality for this random variable when $k = 1$.

(b) Let X be a random variable with probability mass function $P(X = 1) = P(X = -1) = p$ and $P(X = 0) = 1 - 2p$. For any given value of $k > 1$, show that it is possible to select a value of p for which Chebychev's inequality is an equality when applied to this random variable.

Ex. 4.3.6. Let X be a discrete random variable with finite expected value μ and finite variance σ^2.

(a) Explain why $P(|X - \mu| > \sigma) = P((X - \mu)^2 > \sigma^2)$.

(b) Let T be the range of the random variable $(X - \mu)^2$.

Explain why $\sum_{t \in T} P((X - \mu)^2 = t) = 1$.

(c) Explain why $Var[X] = \sum_{t \in T} t \cdot P((X-\mu)^2 = t)$.

(d) Prove that if $P(|X-\mu| > \sigma) = 1$, then

$Var[X] > \sum_{t \in T} \sigma^2 \cdot P((X-\mu)^2 = t)$. (Hint: Use (a) to explain why replacing t by σ^2 in the sum from (c) will only make the quantity smaller).

(e) Use parts (b) and (d) to derive a contradiction. Note that this proves that the assumption that was made in part (d), namely that $P(|X-\mu| > \sigma) = 1$, cannot be true for any discrete random variable where μ and σ are finite quantities. In other words, no random variable can produce only values that are more than one standard deviation from average.

Ex. 4.3.7. Let X be a discrete random variable with finite expected value and finite variance.

(a) Prove $P(|X-\mu| \geq \sigma) = 1 \iff P(|X-\mu| = \sigma) = 1$. (A random variable that assumes only values one or more standard deviations from average must only produce values that are exactly one standard deviation from average).

(b) Prove that if $P(|X-\mu| > \sigma) > 0$ then $P(|X-\mu| < \sigma) > 0$. (If a random variable is able to produce values more one standard deviation from average, it must also be able to produce values that are less than one standard deviation from average).

4.4 CONDITIONAL EXPECTATION AND CONDITIONAL VARIANCE

In previous chapters we saw that information that a particular event had occurred could substantially change the probability associated with another event. That realization led us to the notion of conditional probability. It is also reasonable to ask how such information might affect the expected value or variance of a random variable.

Definition 4.4.1. *Let $X : S \to T$ be a discrete random variable and let $A \subset S$ be an event for which $P(A) > 0$. The "conditional expected value" is defined from conditional probabilities in the same way the (ordinary) expected value is defined from (ordinary) probabilities. Likewise the "conditional variance" is described in terms of the conditional expected value in the same way the (ordinary) variance is described in terms of the (ordinary) expected value. Specificially, the "conditional expected value" of X given A is*

$$E[X|A] = \sum_{t \in T} t \cdot P(X = t|A),$$

and the "conditional variance" of X given A is

$$Var[X|A] = E[(X - E[X|A])^2|A].$$

EXAMPLE 4.4.2. A die is rolled. What are the expected value and variance of the result given that the roll was even?

Let X be the die roll. Then $X \sim \text{Uniform}(\{1, 2, 3, 4, 5, 6\})$, but conditioned on the event A that the roll was even, this changes so that

$$P(X = 1|A) = P(X = 3|A) = P(X = 5|A) = 0 \quad \text{while}$$

$$P(X = 2|A) = P(X = 4|A) = P(X = 6|A) = \frac{1}{3}.$$

Therefore,

$$E[X|A] = 2(\frac{1}{3}) + 4(\frac{1}{3}) + 6(\frac{1}{3}) = 4.$$

Note that the (unconditioned) expected value of a die roll is $E[X] = 3.5$, so the knowledge of event A slightly increases the expected value of the die roll.

The conditional variance is

$$Var[X|A] = (2-4)^2(\frac{1}{3}) + (4-4)^2(\frac{1}{3}) + (6-4)^2(\frac{1}{3}) = \frac{8}{3}.$$

This result is slightly less than $\frac{35}{12}$, the (unconditional) variance of a die roll. This means that knowledge of event A sligthly decreased the typical spread of the die roll results. ■

In many cases the event A on which an expected value is conditioned will be described in terms of another random variable. For instance $E[X|Y = y]$ is the conditional expectation of X given that variable Y has taken on the value y.

Example 4.4.3. Cards are drawn from an ordinary deck of 52, one at a time, randomly and with replacement. Let X and Y denote the number of draws until the first ace and first king are drawn, respectively. We are interested in say, $E[X|Y=3]$. When $Y=3$ an ace was seen of draw 3, but not on draws 1 or 2. Hence

$$P(\text{king on draw } n|Y=3) = \begin{cases} \frac{4}{48} & \text{if } n=1 \text{ or } 2 \\ 0 & \text{if } n=3 \\ \frac{4}{52} & \text{if } n>3 \end{cases}$$

so that

$$P(X=n|Y=5) = \begin{cases} \left(\frac{44}{48}\right)^{n-1}\frac{4}{48} & \text{if } n=1 \text{ or } 2 \\ 0 & \text{if } n=3 \\ \left(\frac{44}{48}\right)^{2}\left(\frac{48}{52}\right)^{n-4}\frac{4}{52} & \text{if } n>3 \end{cases}$$

For example, when $n>3$, in order to have $X=n$ a non-king must have been seen on draws 1 and 2 (each with probability $\frac{44}{48}$), a non-king must have resulted on draw 3 (which is automatic, since an ace was drawn), a non-king must have been seen on each of draws 4 through $n-1$ (each with probability $\frac{48}{52}$), and finally a king was produced on draw n (with probability $\frac{4}{52}$). Hence,

$$\begin{aligned} E[X|Y=3] &= \sum_{n=1}^{2} n\Big(\frac{44}{48}\Big)^{n-1}\frac{4}{48} + \sum_{n=4}^{\infty} n\Big(\frac{44}{48}\Big)^{2}\Big(\frac{48}{52}\Big)^{n-4}\frac{4}{52} \\ &= \sum_{n=1}^{2} n\Big(\frac{44}{48}\Big)^{n-1}\frac{4}{48} + \sum_{m=0}^{\infty} (m+4)\Big(\frac{44}{48}\Big)^{2}\Big(\frac{48}{52}\Big)^{m}\frac{4}{52}. \end{aligned}$$

But

$$\begin{aligned} \sum_{m=0}^{\infty}(m+4)r^m &= \sum_{m=0}^{\infty}\Big(3r^m + \frac{d}{dr}r^{m+1}\Big) \\ &= \frac{3}{1-r} + \frac{d}{dr}\Big(\frac{r}{1-r}\Big) \\ &= \frac{3}{1-r} + \frac{1}{(1-r)^2}, \end{aligned}$$

so

$$\begin{aligned} E[X|Y=3] &= \frac{4}{48} + 2\Big(\frac{44}{48}\Big)\Big(\frac{4}{48}\Big) + \Big(\frac{44}{48}\Big)^2\Big(\frac{4}{52}\Big)\Big(\frac{3}{1-(48/52)} + \frac{1}{(1-(48/52))^2}\Big) \\ &= \frac{4}{48} + 2\Big(\frac{44}{48}\Big)\Big(\frac{4}{48}\Big) + \Big(\frac{44}{48}\Big)^2\Big(\frac{4}{52}\Big)\Big(\frac{3\times 52}{4} + \frac{52^2}{4^2}\Big) \\ &= \frac{1}{12} + 2\Big(\frac{11}{12}\Big)\Big(\frac{1}{12}\Big) + 3\Big(\frac{11}{12}\Big)^2 + \frac{52}{4}\Big(\frac{11}{12}\Big)^2 \\ &= \frac{985}{72} \approx 13.68. \end{aligned}$$

Given that the first ace appeared on draw 3, it takes an average of between 13 and 14 draws until the first king appears. Compare this to the unconditional $E[X]$. Since $X \sim$ Geometric$(\frac{4}{52})$ we know $E[X] = \frac{52}{4} = 13$. In other words, on average it takes 13 draws to observe the first king. But given that the first ace appeared on draw three, we should expect to need about 0.68 draws more (on average) to see the first king. ■

Recall how Theorem 1.3.2 described a way in which a non-conditional probability could be calculated in terms of conditional probabilities. There is an analogous theorem for expected value.

Theorem 4.4.4. *Let $X : S \to T$ be a discrete random variable and let $\{B_i : i \geq 1\}$ be a disjoint collection of events for which $P(B_i) > 0$ for all i and such that $\bigcup_{i=1}^{\infty} B_i = S$. Suppose $P(B_i)$ and $E[X|B_i]$ are known. Then $E[X]$ may be computed as*

$$E[X] = \sum_{i=1}^{\infty} E[X|B_i]P(B_i).$$

Proof - Using Theorem 1.3.2 and the definition of conditional expectation,

$$\begin{aligned} \sum_{i=1}^{\infty} E[X|B_i]P(B_i) &= \sum_{i=1}^{\infty}\sum_{t\in T} t\cdot P(X=t|B_i)P(B_i) \\ &= \sum_{t\in T}\sum_{i=1}^{\infty} t\cdot P(X=t|B_i)P(B_i) \\ &= \sum_{t\in T} t\cdot P(X=t) = E[X]. \end{aligned}$$

■

Example 4.4.5. A venture capitalist estimates that regardless of whether the economy strengthens, weakens, or remains the same in the next fiscal quarter, a particular investment could either gain or lose money. However, he figures that if the economy strengthens,

the investment should, on average, earn 3 million dollars. If the economy remains the same, he figures the expected gain on the investment will be 1 million dollars, while if the economy weakens, the investment will, on average, lose 1 million dollars. He also trusts economic forcasts which predict a 50% chance of a weaker economy, a 40% chance of a stagnant economy, and a 10% chance of a stronger economy. What should he calculate is the expected return on the investment?

Let X be the return on investment and let A, B, and C represent the events that the economy will be stronger, the same, and weaker in the next quarter, respectively. Then the estimates on return give the following information in millions:

$$E[X|A] = 3; \;\; E[X|B] = 1; \text{ and } E[X|C] = -1.$$

Therefore,

$$\begin{aligned} E[X] &= E[X|A]P(A) + E[X|B]P(B) + E[X|C]P(C) \\ &= 3(0.1) + 1(0.4) + (-1)(0.5) = 0.2 \end{aligned}$$

The expected return on investment is \$200,000. ■

When the conditioning event is described in terms of outcomes of a random variable, Theorem 4.4.4 can be written in another useful way.

Theorem 4.4.6. *Let X and Y be two discrete random variables on a sample space S with $Y : S \to T$. Let $g : T \to \mathbb{R}$ be defined as $g(y) = E[X|Y = y]$. Then*

$$E[g(Y)] = E[X].$$

It is common to use $E[X|Y]$ to denote $g(Y)$ after which the theorem may be expressed as $E[E[X|Y]] = E[X]$. This can be slightly confusing notation, but one must keep in mind that the exterior expected value in the expression $E[E[X|Y]]$ refers to the averge of $E[X|Y]$ viewed as a function of Y.

Proof - As y ranges over T, the events $(Y = y)$ are disjoint and cover all of S. Therefore, by Theorem 4.4.4,

$$\begin{aligned} E[g(Y)] &= \sum_{y \in T} g(y)P(Y = y) \\ &= \sum_{y \in T} E[X|Y = y]P(Y = y) \\ &= E[X]. \end{aligned}$$

■

Example 4.4.7. Let $Y \sim \text{Uniform}(\{1, 2, \ldots, n\})$ and let X be the number of heads on Y flips of a coin. What is the expected value of X?

Without Theorem 4.4.6 this problem would require computing many complicated probabilities. However, it is made much simpler by noting that the distribution of X is given conditionally by $(X|Y = j) \sim \text{Binomial}(j, \frac{1}{2})$. Therefore we know $E[X|Y = j] = \frac{j}{2}$. Using the notation above, this may be written as $E[X|Y] = \frac{Y}{2}$ after which

$$E[X] = E[E[X|Y]] = E[\frac{Y}{2}] = \frac{1}{2}\frac{n+1}{2} = \frac{n+1}{4}.$$

■

Though it requires a somewhat more complicated formula, the variance of a random variable can be computed from conditional information.

Theorem 4.4.8. *Let $X : S \to T$ be a discrete random variable and let $\{B_i : i \geq 1\}$ be a disjoint collection of events for which $P(B_i) > 0$ for all i and such that $\bigcup_{i=1}^{\infty} B_i = S$. Suppose $E[X|B_i]$ and $Var[X|B_i]$ are known. Then $Var[X]$ may be computed as*

$$Var[X] = \Big(\sum_{i=1}^{\infty}(Var[X|B_i] + (E[X|B_i])^2)P(B_i)\Big) - (E[X])^2.$$

Proof- First note that $Var[X|B_i] = E[X^2|B_i] - (E[X|B_i])^2$, and so

$$Var[X|B_i] + (E[X|B_i])^2 = E[X^2|B_i].$$

Therefore,

$$\sum_{i=1}^{\infty}(Var[X|B_i] + (E[X|B_i])^2)P(B_i) = \sum_{i=1}^{\infty} E[X^2|B_i]P(B_i),$$

but the right hand side of this equation is $E[X^2]$ from Theorem 4.4.4. The fact that $Var[X] = E[X^2] - (E[X])^2$ completes the proof of the theorem. ■

As with expected value, this formula may be rewritten in a different form if the conditioning events describe the outcomes of a random variable.

Theorem 4.4.9. *Let X and $Y : S \to T$ be two discrete random variables on a sample space S. As in Theorem 4.4.6 let $g(y) = E[X|Y = y]$. Let $h(y) = Var[X|Y = y]$. Denoting $g(Y)$ by $E[X|Y]$ and denoting $h(Y)$ by $Var[X|Y]$, then*

$$Var[X] = E[Var[X|Y]] + Var[E[X|Y]].$$

Proof - First consdier the following three facts:

(1) $\sum_{t\in T} Var[X|Y = t]P(Y = t) = E[Var[X|Y]]$;

(2) $\sum_{t\in T} (E[X|Y = t])^2 P(Y = t) = E[(E[X|Y])^2]$; and

(3) $Var[E[X|Y]] = E[(E[X|Y])^2] - (E[E[X|Y]])^2 = E[(E[X|Y])^2] - (E[X])^2$.

Then from Theorem 4.4.8,

$$\begin{aligned} Var[\ \ X\ \ ] &= \sum_{t\in T} (Var[X|Y = t] + (E[X|Y = t])^2)P(Y = t) - (E[X])^2 \\ &= \sum_{t\in T} Var[X|Y = t]P(Y = t) + \sum_{t\in T} (E[X|Y = t])^2 P(Y = t) - (E[X])^2 \\ &= E[Var[X|Y]] + E[(E[X|Y])^2] - (E[X])^2 \\ &= E[Var[X|Y]] + Var[E[X|Y]]. \end{aligned}$$

■

Example 4.4.10. The number of eggs N found in nests of a certain species of turtles has a Poisson distribution with mean λ. Each egg has probability p of being viable and this event is independent from egg to egg. Find the mean and variance of the number of viable eggs per nest.

Let N be the total number of eggs in a nest and X the number of viable ones. Then if $N = n$, X has a Binomial distribution with number of trials n and probability p of success for each trial. Thus, if $N = n$, X has mean np and variance $np(1-p)$. That is,

$$E[X|N = n] = np; \qquad Var[X|N = n] = np(1-p)$$

or

$$E[X|N] = pN; \qquad Var[X|N] = p(1-p)N.$$

Hence

$$E[X] = E[E[X|N]] = E[pN] = pE[N] = p\lambda$$

and

$$\begin{aligned} Var[X] &= E[Var[X|N]] + Var[E[X|N]] \\ &= E[p(1-p)N] + Var[pN] = p(1-p)E[N] + p^2 Var[N]. \end{aligned}$$

Since N is Poisson we know that $E[N] = Var[N] = \lambda$, so that

$$E[X] = p\lambda \qquad \text{and} \qquad Var[X] = p(1-p)\lambda + p^2\lambda = p\lambda.$$

■

EXERCISES

Ex. 4.4.1. Let $X \sim$ Geometric(p) and let A be event $(X \leq 3)$. Calculate $E[X|A]$ and $Var[X|A]$.

Ex. 4.4.2. Calculate the variance of the quantity X from Example 4.4.7.

Ex. 4.4.3. Return to Example 4.4.5. Suppose that, in addition to the estimates on average return, the investor had estimates on the standard deviations. If the economy strengthens or weakens, the estimated standard deviation is 3 million dollars, but if the economy stays the same, the estimated standard deviation is 2 million dollars. So, in millions of dollars,

$$SD[X|A] = 3; \;\; SD[X|B] = 2; \text{ and } SD[X|C] = 3.$$

Use this information, together with the conditional expectations from Example 4.4.5 to calculate $Var[X]$.

Ex. 4.4.4. A standard light bulb has an average lifetime of four years with a standard deviation of one year. A Super D-Lux lightbulb has an average lifetime of eight years with a standard devaition of three years. A box contains many bulbs – 90% of which are standard bulbs and 10% of which are Super D-Lux bulbs. A bulb is selected at random from the box. What are the average and standard deviation of the lifetime of the selected bulb?

Ex. 4.4.5. Let X and Y be described by the joint distribution

	$X=-1$	$X=0$	$X=1$
$Y=-1$	1/15	2/15	2/15
$Y=0$	2/15	1/15	2/15
$Y=1$	2/15	2/15	1/15

and answer the following questions.

(a) Calculate $E[X|Y = -1]$.

(b) Calculate $Var[X|Y = -1]$.

(c) Describe the distribution of $E[X|Y]$.

(d) Describe the distribution of $Var[X|Y]$.

Ex. 4.4.6. Let X and Y be discrete random variables. Let x be in the range of X and let y be in the range of Y.

(a) Suppose X and Y are independent. Show that $E[X|Y = y] = E[X]$ (and so $E[X|Y] = E[X]$).

(b) Show that $E[X|X = x] = x$ (and so $E[X|X] = X$). (From results in this section we know $E[X|Y]$ is always a random variable with expected value equal to $E[X]$. The results above in some sense show two extremes. When X and Y are independent, $E[X|Y]$ is a constant random variable $E[X]$. When X and Y are equal, $E[X|X]$ is just X itself).

Ex. 4.4.7. Let $X \sim$ Uniform $\{1, 2, \ldots, n\}$ be independent of $Y \sim$ Uniform $\{1, 2, \ldots, n\}$. Let $Z = \max(X, Y)$ and $W = \min(X, Y)$.

(a) Find the joint distribution of (Z, W).

(b) Fine $E[Z \mid W]$.

4.5 COVARIANCE AND CORRELATION

When faced with two different random variables, we are frequently interested in how the two different quantities relate to each other. Often the purpose of this is to predict something about one variable knowing information about the other. For instance, if rainfall amounts in July affect the quantity of corn harvested in August, then a farmer, or anyone else keenly interested in the supply and demand of the agriculture industry, would like to be able to use the July information to help make predictions about August costs.

4.5.1 *Covariance*

Just as we developed the concepts of expected value and standard deviation to summarize a single random variable, we would like to develop a number that describes something about how two different random variables X and Y relate to each other.

Definition 4.5.1. (Covariance of X and Y) *Let X and Y be two discrete random variables on a sample space S. Then the "covariance of X and Y" is defined as*

$$Cov[X,Y] = E[(X - E[X])(Y - E[Y])]. \tag{4.5.1}$$

Since it is defined in terms of an expected value, there is the possibility that the covariance may be infinite or not defined at all because the sum describing the expectation is divergent.

Notice that if X is larger than its average at the same time that Y is larger than its average (or if X is smaller than its average at the same time Y is smaller than its average) then $(X - E[X])(Y - E[Y])$ will contribute a positive result to the expected value describing the covariance. Conversely, if X is smaller than $E[X]$ while Y is larger than $E[Y]$ or vica versa, a negative result will be contributed toward the covariance. This means that when two variables tend to be both above average or both below average simultaneously, the covariance will typically be positive (and the variables are said to be positively correlated), but when one variable tends to be above average when the other is below average, the covariance will typically be negative (and the variables are said to be negatively correlated). When $Cov[X,Y] = 0$ the variables X and Y are said to be "uncorrelated".

For example, suppose X and Y are the height and weight, respectively, of an individual randomly selected from a large population. We might expect that $Cov[X,Y] > 0$ since people who are taller than average also tend to be heavier than average and people who are shorter than average tend to be lighter. Conversely suppose X and Y represent elevation and air density at a randomly selected point on Earth. We might expect $Cov[X,Y] < 0$ since locations at a higher elevation tend to have thinner air.

Example 4.5.2. Consider a pair of random variables X and Y with joint distribution

	$X=-1$	$X=0$	$X=1$
$Y=-1$	1/15	2/15	2/15
$Y=0$	2/15	1/15	2/15
$Y=1$	2/15	2/15	1/15

By a routine calculation of the marginal distributions it can be shown that $X, Y \sim \text{Uniform}(\{-1,0,1\})$ and therefore that $E[X] = E[Y] = 0$. However, it is clear from the joint distribution that when $X = -1$, then Y is more likely to be above average than below,

while when $X = 1$, then Y is more likely to be below average than above. This suggests the two random variables should have a negative correlation. In fact, we can calculate

$$E[XY] = (-1)(\frac{4}{15}) + 0(\frac{9}{15}) + 1(\frac{2}{15}) = -\frac{2}{15},$$

and therefore $Cov[X,Y] = E[XY] - E[X]E[Y] = -\frac{2}{15}$. ■

As its name suggests, the covariance is closely related to the variance.

Theorem 4.5.3. *Let X be a discrete random variable. Then*

$$Cov[X,X] = Var[X].$$

Proof - $Cov[X,X] = E[(X - E[X])(X - E[X])) = E[(X - E[X])^2) = Var[X]$. ■

With Theorem 4.2.5 it was shown that $Var[X] = E[X^2] - (E[X])^2$, which provided an alternate formula for the variance. There is an analogous alternate formula for the covariance.

Theorem 4.5.4. *Let X and Y be discrete random variables with finite mean for which $E[XY]$ is also finite. Then*

$$Cov[X,Y] = E[XY] - E[X]E[Y].$$

Proof - Using the linearity properties of expected value,

$$\begin{aligned} Cov[X,Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - XE[Y] - E[X]Y + E[X]E[Y]] \\ &= E[XY] - E[XE[Y]] - E[E[X]Y] + E[E[X]E[Y]] \\ &= E[XY] - E[Y]E[X] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y]. \end{aligned}$$

■

As with the expected value, the covariance is a linear quantity. It is also related to the concept of independence.

Theorem 4.5.5. *Let X, Y, and Z be discrete random variables, and let $a, b \in \mathbb{R}$. Then,*

(a) $Cov[X, Y] = Cov[Y, X]$;

(b) $Cov[X, aY + bZ] = a \cdot Cov[X, Y] + b \cdot Cov[X, Z]$;

(c) $Cov[aX + bY, Z] = a \cdot Cov[X, Z] + b \cdot Cov[Y, Z]$; *and*

(d) If X and Y are independent with a finite covariance, then $Cov[X, Y] = 0$.

Proof of (1) - This follows immediately from the definition.

$$\begin{aligned} Cov[X, Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[(Y - E[Y])(X - E[X])] = Cov[Y, X]. \end{aligned}$$

Therefore, reversing the roles of X and Y does not change the correlation.

Proof of (2) - This follows from linearity properties of expected value. Using Theorem 4.5.4

$$\begin{aligned} Cov[X, aY + bZ] &= E[X(aY + bZ)] - E[X]E[aY + bZ] \\ &= a \cdot E[XY] + b \cdot E[XZ] - a \cdot E[X]E[Y] - b \cdot E[X]E[Z] \\ &= a \cdot (E[XY] - E[X]E[Y]) + b \cdot (E[XZ] - E[X]E[Z]) \\ &= a \cdot Cov[X, Y] + b \cdot Cov[X, Z] \end{aligned}$$

Proof of (3) - This proof is essentially the same as that of (2) and is left as an exercise.

Poof of (4) - We have previously seen that if X and Y are independent, then $E[XY] = E[X]E[Y]$. Using Theorem 4.5.4 it follows that

$$Cov[X, Y] = E[XY] - E[X]E[Y] = 0.$$

Though independence of X and Y guarantees that they are uncorrelated, the converse is not true. It is possible that $Cov[X, Y] = 0$ and yet that X and Y are dependent, as the next example shows.

Example 4.5.6. Let X, Y be two discrete random variables taking values $\{-1, 1\}$. Suppose their joint distribution $P(X = x, Y = y)$ is given by the table

	x=-1	x=1
y=-1	0.3	0.2
y=1	0.3	0.2

By summing the columns and rows respectively,

$$P(X = 1) = 0.4 \text{ and } P(X = -1) = 0.6, \text{ while}$$

$$P(Y = 1) = 0.5 \text{ and } P(Y = -1) = 0.5.$$

Moreover,

$$\begin{aligned} E[XY] &= (1)(-1)P(X = 1, Y = -1) + (-1)(1)P(X = -1, Y = 1) \\ &\quad +(1)(1)P(X = 1, Y = 1) + (-1)(-1)P(X = -1, Y = -1) \\ &= -0.3 - 0.2 + 0.2 + 0.3 = 0, \\ E[X] &= (1)0.4 + (-1)0.6 = -0.2, \\ E[Y] &= (1)0.5 + (-1)0.5 = 0, \end{aligned}$$

implying that $\text{Cov}[X, Y] = 0$. As

$$P(X = 1, Y = 1) = 0.2 \neq 0.1 = P(X = 1)P(Y = 1),$$

they are not independent random variables. ■

4.5.2 *Correlation*

The possible size of $Cov[X, Y]$ has upper and lower bounds based on the standard deviations of the two variables.

Theorem 4.5.7. *Let X and Y be two discrete random variables both with finite variance. Then*

$$-\sigma_X \sigma_Y \leq Cov[X, Y] \leq \sigma_X \sigma_Y,$$

and therefore $-1 \leq \frac{Cov[X,Y]}{\sigma_X \sigma_Y} \leq 1$.

Proof - Standardize both variables and consider the expected value of their sum squared. Since this is the expected value of a non-negative quantity,

$$\begin{aligned} 0 &\leq E[(\frac{X-\mu_X}{\sigma_X}+\frac{Y-\mu_Y}{\sigma_Y})^2] \\ &= E[\frac{(X-\mu_X)^2}{\sigma_X^2}+2\frac{(X-\mu_X)(Y-\mu_Y)}{\sigma_X\sigma_Y}+\frac{(Y-\mu_Y)^2}{\sigma_Y^2}] \\ &= \frac{E[(X-\mu_X)^2]}{\sigma_X^2}+\frac{2E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y}+\frac{E[(Y-\mu_Y)^2]}{\sigma_Y^2} \\ &= 1+2\frac{Cov[X,Y]}{\sigma_X\sigma_Y}+1. \end{aligned}$$

Sovling the inequality for the covariance yields

$$Cov[X,Y] \geq -\sigma_X\sigma_Y.$$

A similar computation (see Exercises) for the expected value of the squared difference of the standardized variables shows

$$Cov[X,Y] \leq \sigma_X\sigma_Y.$$

Putting both inequalities together proves the theorem. ■

Definition 4.5.8. *The quantity* $\frac{Cov[X,Y]}{\sigma_X\sigma_Y}$ *from Theorem 4.5.7 is known as the"correlation" of* X *and* Y *and is often denoted as* $\rho[X,Y]$*. Thinking in terms of dimensional analysis, both the numerator and denominator include the units of* X *and the units of* Y*. The correlation, therefore, has no units associated with it. It is thus a dimensionless rescaling of the covariance and is frequently used as an absolute measure of trends between the two variables.*

EXERCISES

Ex. 4.5.1. Consider the experiment of flipping two coins. Let X be the number of heads among the coins and let Y be the number of tails among the coins.

(a) Should you expect X and Y to be posivitely correlated, negatively correlated, or uncorrelated? Why?

(b) Calculate $Cov[X,Y]$ to confirm your answer to (a).

Ex. 4.5.2. Let $X \sim \text{Uniform}(\{0,1,2\})$ and let Y be the number of heads in X flips of a coin.

(a) Should you expect X and Y to be positively correlated, negatively correlated, or uncorrelated? Why?

(b) Calculate $Cov[X,Y]$ to confirm your answer to (a).

Ex. 4.5.3. Prove part (3) of Theorem 4.5.5.

Ex. 4.5.4. Prove the missing inequality from the proof of Theorem 4.5.7. Specifically, use the inequality

$$0 \leq E[(\frac{X-\mu_X}{\sigma_X} - \frac{Y-\mu_Y}{\sigma_Y})^2]$$

to prove that $Cov[X,Y] \leq \sigma_X \sigma_Y$.

Ex. 4.5.5. Prove that the inequality of Theorem 4.5.7 is an equality if and only if there are $a, b \in \mathbb{R}$ with $a \neq 0$ for which $P(Y = aX + b) = 1$. (Put another way, the correlation of X and Y is ± 1 exactly when Y can be expressed as a non-trivial linear function of X).

Ex. 4.5.6. In previous sections it was shown that if X and Y are independent, then $Var[X+Y] = Var[X] + Var[Y]$. If X and Y are dependent, the result is typically not true, but the covariance provides a way relate the variances of X and Y to the variance of their sum.

(a) Show that for any discrete random variables X and Y,

$$Var[X+Y] = Var[X] + Var[Y] + 2Cov[X,Y].$$

(b) Use (a) to conclude that when X and Y are positively correlated, then $Var[X+Y] > Var[X] + Var[Y]$, while when X and Y are negatively correlated, $Var[X+Y] < Var[X] + Var[Y]$.

(c) Suppose X_i $1 \leq i \leq n$ are discrete random variables with finite variance and covariances. Use induction and (a) to conclude that

$$Var[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} Var[X_i] + 2 \sum_{1 \leq i < j \leq n} Cov[X_i, X_j].$$

4.6 EXCHANGEABLE RANDOM VARIABLES

We conclude this section with a discussion on exchangeable random variables. In brief we say that a collection of random variables is exchangeable if the joint probability mass

function of $(X_1, X_2, \ldots, X_n)$ is a symmetric function. In other words, the distribution of $(X_1, X_2, \ldots, X_n)$ is independent of the order in which the X_i's appear. In particular any collection of mutually independent random variables is exchangeable.

Definition 4.6.1. *Let $n \geq 2$ and $\sigma : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ be a bijection. We say that a subset T of $\mathbb{R}^n$ is symmetric if*

$$(x_1, x_2, \ldots, x_n) \in T \iff (x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(n)}) \in T$$

for all $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. For any symmetric set T, a function $f : T \to \mathbb{R}$ is symmetric if

$$f(x_1, x_2, \ldots, x_n) = f(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(n)})$$

for all $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$.

A bijection $\sigma : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ is often referred to as a permutation of $\{1, 2, \ldots, n\}$. When $n = 2$ the function f would be symmetric if $f(x, y) = f(y, x)$ for all $x, y \in \mathbb{R}$.

Definition 4.6.2. *Let $n \geq 1$ and $X_1, X_2, \ldots, X_n$ be discrete random variables. We say that $X_1, X_2, \ldots, X_n$ is a collection of exchangeable random variables if the joint probability mass function given by*

$$f(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, \ldots X_n = x_n)$$

is a symmetric function.

In particular, $X_1, X_2, \ldots, X_n$ are exchangeable then for any one of the possible $n!$ permutations, σ, of $\{1, 2, \ldots, n\}$, $X_1, X_2, \ldots, X_n$ and $X_{\sigma(1)}, X_{\sigma(2)}, \ldots, X_{\sigma(n)}$ have the same distribution.

Example 4.6.3. Suppose we have an urn of m distinct objects labelled $\{1, 2, \ldots, m\}$. Objects are drawn at random from the urn without replacements till the urn is empty. Let X_i be the label of the i-th object that is drawn. Then $X_1, X_2, \ldots, X_m$ is a particular ordering of the objects in the urn. Since each ordering is equally likely and there are $m!$ possible orderings we have that the joint probability mass function

$$f(x_1, x_2, \ldots, x_m) = P(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m) = \frac{1}{m!},$$

whenever $x_i \in \{1, 2, \dots m\}$ with $x_i \neq x_j$. As the function is a constant function on the symmetric set $\{1, 2, \dots, m\}$, it is clearly symmetric. So the random variables $X_1, X_2, \dots, X_m$ are exchangeable. ■

Theorem 4.6.4. *Let $X_1, X_2, \dots, X_n$ be a collection of exchangeable random variables on a sample space S. Then for any $i, j \in \{1, 2, \dots, n\}$, X_i and X_j have the same marginal distribution.*

Proof - The random variables $(X_1, X_2, \dots, X_n)$ are exchangeable. Then we have for any permutation σ and $x_i \in \text{Range}(X_i)$

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = P(X_{\sigma(1)} = x_1, X_{\sigma(2)} = x_2, \dots, X_{\sigma(n)} = x_n).$$

As this is true for all permutations σ all the random variables must have same range. Otherwise if any two of them differ the we could get a contradiction by choosing an appropriate permutation.

Let T denote the common range. Let $i \in \{2, \dots, n\}, a, b \in T$. Let

$$A = \{x_j \in T : 1 \leq j \neq 1, i \leq n\}$$

By using the exchangeable property with the permutation σ that is given by $\sigma(i) = 1, \sigma(1) = i$ and $\sigma(j) = j$ for all $j \neq 1, i$. We have that for any $x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n \in A$

$$\begin{aligned} &P(X_1 = a, X_2 = x_2, \dots, X_{i-1} = x_{i-1}, X_i = b, X_{i+1} = x_{i+1}, \dots X_n = x_n) \\ &= P(X_1 = b, X_2 = x_2, \dots, X_{i-1} = x_{i-1}, X_i = a, X_{i+1} = x_{i+1}, \dots X_n = x_n). \end{aligned}$$

Therefore,

$$\begin{aligned}
P(X_1 = a) &= P(\bigcup_{b\in T} X_1 = a, X_i = b) \\
&= \sum_{b\in T} P(X_1 = a, X_i = b) \\
&= \sum_{b\in T} P(\bigcup_{x_j\in A} X_1 = a, X_2 = x_2, \dots X_{i-1} = x_{i-1}, X_i = b, X_{i+1} = x_{i+1}, \dots X_n = x_n) \\
&= \sum_{b\in T}\sum_{x_j\in A} P(X_1 = a, X_2 = x_2, \dots, X_i = b, \dots X_n = x_n) \\
&= \sum_{b\in T}\sum_{x_j\in A} P(X_1 = b, X_2 = x_2, \dots, X_i = a, \dots X_n = x_n) \\
&= \sum_{b\in T} P(\bigcup_{x_j\in A} X_1 = b, X_2 = x_2, \dots X_{i-1} = x_{i-1}, X_i = a, X_{i+1} = x_{i+1}, \dots X_n = x_n) \\
&= \sum_{b\in T} P(X_1 = b, X_i = a) \\
&= P(\bigcup_{b\in T} X_1 = b, X_i = a) \\
&= P(X_i = a)
\end{aligned}$$

So the distribution of X_i is the same as the distribution of X_1 and hence all of them have the same distribution. ■

Example 4.6.5. **(Sampling without Replacement)** An urn contains b black balls and r red balls. A ball is drawn at random and its colour noted. This procedure is repeated n times. Assume that $n \leq b + r$. Let $\max 0, n - r \leq k \leq \min(n, b)$. In this example we examine the random variables X_i given by

$$X_i = \begin{cases} 1 & \text{if } i\text{-th ball drawn is black} \\ 0 & \text{otherwise} \end{cases}$$

We have already seen that (See Theorem 2.3.2 and Example 2.3.1)

$$P(k \text{ black balls are drawn in } n \text{ draws}) = \binom{n}{k} \frac{\prod_{i=0}^{k-1}(b-i)\prod_{i=0}^{m-k-1}(r-i)}{\prod_{i=0}^{m-1}(r+b-i)}.$$

Using the same proof we see that the joint probability mass function of $(X_1, X_2, \dots, X_n)$ is given by

$$f(x_1, x_2, \dots, x_n) = P(X_1 = x_1, X_2 = x_2 \dots X_n = x_n) = \frac{\prod_{i=0}^{\sum_{i=1}^{n} x_i - 1}(b-i)\prod_{i=0}^{\sum_{i=1}^{n} x_i - k - 1}(r-i)}{\prod_{i=0}^{\sum_{i=1}^{n} x_i - 1}(r+b-i)},$$

where $x_i \in \{0,1\}$. It is clear from the right hand side of the above that the function f depends only on the $\sum_{i=1}^{n} x_i$. Hence any permutation of the x_i's will not change the value of f. So f is a symmetric function and the random variables are exchangeable. Therefore, by Theorem 4.6.4 we know that for any $1 \leq i \leq n$,

$$P(X_i = 1) = P(X_1 = 1) = \frac{b}{b+r}.$$

So we can conclude that they are all identically distributed as Bernoulli ($\frac{b}{b+r}$) and the probability of choosing a black ball in the i-th draw is $\frac{b}{b+r}$ (See Exercise 4.6.4 for a similar result). Further for any i, j

$$\begin{aligned} Cov[X_i, X_j] &= E[X_i X_j] - E[X_i]E[X_j] \\ &= E[X_1 X_2] - \left(\frac{b}{b+r}\right)^2 \\ &= \frac{b(b-1)}{(b+r)(b+r-1)} - \left(\frac{b}{b+r}\right)^2 \\ &= \frac{-br}{(b+r)^2(b+r-1)} \end{aligned}$$

Finally, we observe that $Y = \sum_{i=1}^{n} X_i$ is a Hypergeometric $(b+r, b, m)$. Exchangeability thus provides another alternative way to compute the mean and variance of Y. Using the linearity of expectation provided by Theorem 4.1.7, we have

$$E[Y] = E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] = n\frac{b}{b+r}.$$

and by Exercise 4.5.6,

$$\begin{aligned} Var[Y] &= Var[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} Var[X_i] + 2\sum_{1 \leq i < j \leq n}^{n} Cov[X_i, X_j] \\ &= nVar[X_1] + n(n-1)Cov[X_1, X_2] \\ &= n\frac{br}{(b+r)^2} + n(n-1)(\frac{-br}{(b+r)^2(b+r-1)}) \\ &= n\frac{br}{(b+r)^2}\frac{b+r-n}{b+r-1}. \end{aligned}$$

■

EXERCISES

Ex. 4.6.1. Suppose $X_1, X_2, \ldots, X_n$ are exchangeable random variables. For any $2 \leq m < n$, show that $X_1, X_2, \ldots, X_m$ are also a collection of exchangeable random variables.

Ex. 4.6.2. Suppose $X_1, X_2, \ldots, X_n$ are exchangeable random variables. Let T denote their common range. Suppose $b : T \to \mathbb{R}$. Show that $b(X_1), b(X_2), \ldots, b(X_n)$ is also a collection of exchangeable random variables.

Ex. 4.6.3. Suppose n cards are drawn from a standard pack of 52 cards without replacement (so we will assume $n \leq 52$). For $1 \leq i \leq n$, let X_i be random variables given by

$$X_i = \begin{cases} 1 & \text{if } i\text{-th card drawn is black in colour} \\ 0 & \text{otherwise} \end{cases}$$

(a) Suppose $n = 52$. Using Example 4.6.3 and the Exercise 4.6.2 show that $(X_1, X_2, X_3, \ldots X_n)$ are exchangeable.

(b) Show that $(X_1, X_2, X_3, \ldots, X_n)$ are exchangeable for any $2 \leq n \leq 52$. *Hint: If $n < 52$ extend the sample to exhause the deck of cards. Use (a) and Exercise 4.6.1*

(c) Find the probability that the second and fourth card drawn have the same colour.

Ex. 4.6.4. **(Polya Urn Scheme)** An urn contains b black balls and r red balls. A ball is drawn at random and its colour noted. Then it is replaced along with $c \geq 0$ balls of the same colour. This procedure is repeated n times.

(a) Let $1 \leq k \leq m \leq n$. Show that

$$P(k \text{ black balls are drawn in } m \text{ draws}) = \binom{m}{k} \frac{\prod_{i=0}^{k-1}(b+ci)\prod_{i=0}^{m-k-1}(r+ci)}{\prod_{i=0}^{m-1}(r+b+ci)}$$

(b) Let $1 \leq i \leq n$ and random variables X_i be given by

$$X_i = \begin{cases} 1 & \text{if } i\text{-th ball drawn is black} \\ 0 & \text{otherwise} \end{cases}$$

Show that the collection of random variables is exchangeable.

(c) Let $1 \leq m \leq n$. Let B_m be the event that the m-th ball drawn is black. Show that

$$P(B_m) = \frac{b}{b+r}.$$

5 CONTINUOUS PROBABILITIES AND RANDOM VARIABLES

We have thus far restricted discussion to discrete spaces and discrete random variables—those consisting of at most a countably infinite number of outcomes. This is not because it is not possible, interesting, or useful to consider probabilities on an uncountably infinite set such as the real line or the interval $(0, 1)$. Instead, there are a few technicalities that arise when discussing such probabilities that are best avoided until they are needed. That time is now.

5.1 UNCOUNTABLE SAMPLE SPACES AND DENSITIES

Suppose we want to randomly select a number on the interval $(0, 1)$ in some uniform way. In the discrete setting we would have said that "uniform" meant that every outcome in our sample space $S = (0, 1)$ was equally likely. Suppose we took that same approach here and declared that there was some value p for which $P(\{x\}) = p$ for every $x \in (0, 1)$. Then if we let E be the event $E = \{\frac{1}{n} : n = 2, 3, 4, \dots\} \subset S$, we find that

$$\begin{aligned} P(E) &= P(\left\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots\right\}) \\ &= P(\frac{1}{2}) + P(\frac{1}{3}) + P(\frac{1}{4}) + \dots \\ &= p + p + p + \dots \end{aligned}$$

If $p > 0$ this sum diverges to infinity, which cannot be since it describes a probability. Therefore it must be that $p = 0$. If every individual outcome in $S = (0, 1)$ is equally likely, then each outcome must have a probability of zero. After several chapters considering only discrete probabilities many readers may suspect that this, in and of itself, is a contradiction. How is it possible for $P(S) = 1$ when every single element of S has probability zero? Could not one then show

$$\begin{aligned} P(S) &= P(\bigcup_{s \in S} \{s\}) \\ &= \sum_{s \in S} P(\{s\}) \\ &= \sum_{s \in S} 0 \\ &= 0 \end{aligned}$$

using the probability axioms? The answer to that question is "no". The probability space axiom that allows us to write the probability of a disjoint union as the sum of separate probabilities only applies to countable collections of events. But the events $\{s\}$ that combine to create $(0,1)$ are an uncountable collection. If S is uncountable, we could still have $P(S) = 1$ even if every individual element of $s \in S$ has probability zero.

However, all of that does not yet explain how to define a uniform probability on $(0,1)$. Knowing that each individual outcome has probability zero does not tell us how to calculate $P([\frac{1}{4}, \frac{3}{4}])$, for example, since we cannot simply add up the probabilities of each of the constituent outcomes individually. Instead we need to reinterpret what we mean by "uniform" in this situation. It would make sense to suggest that the event $[\frac{1}{4}, \frac{3}{4}]$ should have a probability of $\frac{1}{2}$ since its length is exactly half of the length of $(0,1)$. Indeed it is tempting (and essentially correct) to declare that $P(A)$ should be the length of the set A. The complication with making such a statement is that, although length is easy to define if A is an interval or even a countable collection of disjoint intervals, it is not even possible to consistently define a length for every single subset of $(0,1)$. Because of this unfortunate fact, we will need to reconsider which subsets of S are actually events which will be assigned a probability.

At a minimum we will want events to include any interval. The axioms and basic properties of probability spaces also require that for any collection of events we must be able to consider complements and countable unions of these events. Further, the entire sample space S should also be considered a legitimate event. Consequently we make the following definition.

Definition 5.1.1. **(σ-field)** *If S is a sample space, then a σ-field $\mathcal{F}$ is a collection of subsets of S such that*

(1) $S \in \mathcal{F}$

(2) If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$

(3) If $A_1, A_2, \ldots$ is a countable collection of sets in $\mathcal{F}$ then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$

We shall refer to an element of the σ-field as an event.

If S happens to be the set of real numbers there is a smallest σ-field that contains all intervals, and this collection of subsets of $\mathbb{R}$ is known as the Borel sets. It happens that the concept of the "length" of a set can be consistantly described for such sets. Because of this we will modify our definition of probability space slightly at this point.

Definition 5.1.2. (Probability Space Axioms) *Let S be a sample space and let $\mathcal{F}$ be a σ-field of S. A "***probability***" is a function $P : \mathcal{F} \to [0, 1]$ such that*

(1) $P(S) = 1$;

(2) If $E_1, E_2, \ldots$ are a countable collection of disjoint events in $\mathcal{F}$, then

$$P(\bigcup_{j=1}^{\infty} E_j) = \sum_{j=1}^{\infty} P(E_j).$$

The triplet $(S, \mathcal{F}, P)$ is referred to as a probability space.

Our old definition is simply a special case where the σ-field was the collection of all subset of S, so all results we have previously seen in the discrete setting are still legitimate in this new framework. There are many technicalities that arise due to the fact that not every set may be viewed as an event, but these issues would be distracting from the primary goal of this text. Thus we give the definitions above only to provide the modern definition of probability space.

Throughout the remainder of the sections on continuous probability spaces we will restrict our attention to the sample space being $\mathbb{R}$. Whenever we state or prove anything for an event A (a Borel set) we shall restrict ourselves to the case the event A is a finite or countable unions of intervals. This will enable us to use standard results from calculus and thereby avoid technicalities. A thorough study of the Borel sets and the related theory of integration is beyond the scope of this text (the interested reader may see [AS09] in the bibliography for additional information).

5.1.1 *Probability Densities on* $\mathbb{R}$

The primary way we will define continuous probabilities on $\mathbb{R}$ is through a "density function". We begin by providing an example of what should be meant by a uniform distribution on $(0, 1)$.

Example 5.1.3. Let $f : \mathbb{R} \to \mathbb{R}$ be a function defined by

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

For an event A define $P(A) = \int_A f(x)\,dx$. Note that for an interval $A = [a, b] \subset (0, 1)$ it happens that $P(A)$ is just the length of the interval.

$$\begin{aligned} P(A) &= \int_A f(x)\,dx \\ &= \int_a^b 1\,dx \\ &= b - a \end{aligned}$$

For disjoint unions of intervals, the lengths simply add. For instance if $A = [\frac{1}{5}, \frac{2}{5}] \cup [\frac{3}{5}, \frac{4}{5}]$, then

$$\begin{aligned} P(A) &= \int_{[\frac{1}{5},\frac{2}{5}]\cup[\frac{3}{5},\frac{4}{5}]} f(x)\,dx \\ &= \int_{\frac{1}{5}}^{\frac{2}{5}} 1\,dx + \int_{\frac{3}{5}}^{\frac{4}{5}} 1\,dx \\ &= \frac{1}{5} + \frac{1}{5} = \frac{2}{5} \end{aligned}$$

which is the sum of the lengths of the two component intervals. In particular note that $P((0, 1)) = 1$ while $P(\{c\}) = 0$ for any c since a single point has no length. Similarly, if $A = [a, b]$ is an interval that is disjoint from $(0, 1)$, then

$$\begin{aligned} P(A) &= \int_A f(x)\,dx \\ &= \int_a^b 0\,dx \\ &= 0 \end{aligned}$$

We will soon see that P defines a probability on $\mathbb{R}$. From the computation above this probability gives equal likelihood to all equal-width intervals within $(0, 1)$ and assigns zero probability to any interval outside of $(0, 1)$. Therefore it is consistant with the properties a uniform probability on $(0, 1)$ should have. ■

The function f from the example above is known as a density. What properties must be required of such a function in order for it to define a probability? The fact that probabilities cannot be negative suggests we will need to require $f(x)$ to be non-negative for all real numbers x. The fact that $P(S) = 1$ means that $\int_{-\infty}^{\infty} f(x)\,dx$ has to be 1. It turns out that these two requirements are essentially all that are needed. The only other assumption we will make is that a density funciton be piecewise continuous. Though this final requirement is more restrictive than necessary, the assumption will help avoid technicalities and will

include all densities of interest to us in the remainder of the text. We give a precise definition.

Definition 5.1.4. *Let $f : \mathbb{R} \to \mathbb{R}$ is called a density function if f satisfies the following:*

(i) $f(x) \geq 0$,

(ii) f *is piecewise-continuous, and*

(iii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

We proceed to state and prove a result that will help us construct probabilities on $\mathbb{R}$ with the help of density functions. This will also ensure that in Example 5.1.3 we indeed constructed a probability on $\mathbb{R}$.

Theorem 5.1.5. *Let $f(x)$ be a density function. Define*

$$P(A) = \int_A f(x)\,dx, \tag{5.1.1}$$

for any event $A \subset \mathbb{R}$. Then P defines a probability on $\mathbb{R}$. The function f is called the "density function" for the probability P.

Proof - First note

$$\begin{aligned} P(\mathbb{R}) &= \int_{\mathbb{R}} f(x)\,dx \\ &= \int_{-\infty}^{\infty} f(x)\,dx = 1 \end{aligned}$$

by assumption, so the entire sample space has probability 1. Now let A be a Borel subset of $\mathbb{R}$. Since $f(x)$ is non-negative,

$$\begin{aligned} P(A) &= \int_A f(x)\,dx \geq 0, \qquad \text{and} \\ P(A) &= \int_A f(x)\,dx \leq \int_{\mathbb{R}} f(x)\,dx = 1, \end{aligned}$$

so $P(A) \in [0,1]$. Finally, if $E_1, E_2, \ldots$ are a countable collection of disjoint events, then

$$\begin{aligned} P(\bigcup_{n=1}^{\infty} E_n) &= \int_{\bigcup_{n=1}^{\infty} E_n} f(x)\,dx \\ &= \sum_{n=1}^{\infty} \int_{E_n} f(x)\,dx \\ &= \sum_{n=1}^{\infty} P(E_n). \end{aligned}$$

Therefore P satisfies the conditions of a probability space on $\mathbb{R}$. ■

Example 5.1.6. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} 3x^2 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

f is piecewise continuous, $f(x)$ is non-negative for all x and

$$\int_{\mathbb{R}} f(x)\,dx = \int_0^1 3x^2\,dx = x^3\Big|_0^1 = 1.$$

As it satisfies $(i) - (iii)$ in Definition 5.1.4, f is a density function. Let P be as defined in (5.1.1). As with the uniform example, $f(x)$ is zero outside of $(0,1)$, so events lying outside this interval will have zero probability. However, note that

$$P([\frac{1}{5}, \frac{2}{5}]) = \int_{\frac{1}{5}}^{\frac{2}{5}} 3x^2\,dx = \frac{7}{125}$$

while

$$P([\frac{3}{5}, \frac{4}{5}]) = \int_{\frac{3}{5}}^{\frac{4}{5}} 3x^2\,dx = \frac{37}{125}.$$

In other words, intervals of the same length do not have equal probabilities; this probability is not uniform on $(0,1)$.

The probability of individual points is still zero, so $P(\{\frac{1}{5}\}) = P(\{\frac{2}{5}\}) = 0$, but in terms of the density function, $f(\frac{2}{5})$ is four times as large as $f(\frac{1}{5})$. What does this mean in practical terms?

Let ϵ be a small positive quantity (certianly less than $\frac{1}{5}$). Then

$$P([\frac{1}{5} - \epsilon, \frac{1}{5} + \epsilon]) = \frac{2}{25}\epsilon + 2\epsilon^3 \approx \frac{2}{25}\epsilon \qquad \text{while}$$

$$P([\frac{2}{5} - \epsilon, \frac{2}{5} + \epsilon]) = \frac{8}{25}\epsilon + 2\epsilon^3 \approx \frac{8}{25}\epsilon.$$

The fact that $f(\frac{2}{5})$ is four times as large as $f(\frac{1}{5})$ essentially means that a tiny interval around $\frac{2}{5}$ has approximately four times the probability of a similarly sized interval around $\frac{1}{5}$. ■

EXERCISES

Ex. 5.1.1. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} 2x & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Show that f is a probability density function.

(b) Use f to calculate $P((0, \frac{1}{2}))$.

Ex. 5.1.2. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} x & \text{if } 0 < x < 1 \\ 2 - x & \text{if } 1 \leq x < 2 \\ 0 & \text{otherwise} \end{cases}$$

(a) Sketch a graph of the function f.

(b) Show that f is a probability density function.

(c) Use f to calculate : $P((0, \frac{1}{4}))$, $P((\frac{3}{2}, 2))$, $P((-3, -2))$ and $P((\frac{1}{2}, \frac{3}{2}))$.

Ex. 5.1.3. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} k & \text{if } 0 < x < \frac{1}{4} \\ 2k & \text{if } \frac{1}{4} \leq x < \frac{3}{4} \\ 3k & \text{if } \frac{3}{4} \leq x < 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find k that makes f a probability density function.

(b) Sketch a graph of the function f.

(c) Use f to calculate : $P((0, \frac{1}{4}))$, $P((\frac{1}{4}, \frac{3}{4}))$, $P((\frac{3}{4}, 1))$.

Ex. 5.1.4. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} k \cdot \sin(x) & \text{if } 0 < x < \pi \\ 0 & \text{otherwise} \end{cases}$$

(a) Determine the value of k that makes f a probability density function.

(b) Calculate $P((0, \frac{1}{2}))$ and $P((\frac{1}{2}, 1))$.

(c) Which will be larger, $P((0, \frac{1}{4}))$ or $P((\frac{1}{4}, \frac{1}{2}))$? Explain how you can answer this question without actually calculating either probability.

(d) A game is played in the following way. A random variable X is selected with a density described by f above. You must select a number r and you win the game if the random variable results in an outcome in the interval $(r - 0.01, r + 0.01)$. Explain how you should choose r to maximize your chance of winning the game. (A formal proof requires only basic calculus, but it should take very little computation to determine the correct answer).

Ex. 5.1.5. Let $\lambda > 0$ and $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } 0 < x \\ 0 & \text{otherwise} \end{cases}$$

(a) Show that f is a probability density function.

(b) Let $a > 0$. Find $P((a, \infty))$.

Ex. 5.1.6. Let $a, b \in \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

(a) Show that f is a probability density function.

(b) Show that if $I, J \subset [a, b]$ are two intervals that have the same length, then $P(I) = P(J)$.

Ex. 5.1.7. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} \frac{1}{x^2} & \text{if } 1 < x \\ 0 & \text{otherwise} \end{cases}$$

(a) Show that f is a probability density function.

(b) Let $a > 1$. Find $P((a, \infty))$.

Ex. 5.1.8. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} \frac{1}{6}x^2 e^{-x} & \text{if } 0 < x \\ 0 & \text{otherwise} \end{cases}$$

Show that f is a probability density function.

Ex. 5.1.9. For any $x \in \mathbb{R}$, the hyperbolic secant is defined as $\operatorname{sech} x = \frac{2}{(e^x + e^{-x})}$. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \frac{1}{2}\operatorname{sech}(\frac{\pi}{2}x), x \in \mathbb{R}$$

Show that f is a probability density function.

Ex. 5.1.10. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad x \in \mathbb{R}$$

Follow the steps below to show that the function f is a density function.

(a) Let $I = \int_{-\infty}^{\infty} e^{-x^2/2}\, dx$ and then explain why

$$I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\, dx\, dy$$

(Hint: Write I^2 as a product of two integrals each over a different variable and explain why the resulting expression may be written as the double integral above).

(b) Explain why

$$I^2 = \int_0^{2\pi} \int_0^{\infty} r \cdot e^{-r^2/2}\, dr\, d\theta$$

after switching from rectangular to polar coordinates. (Hint: Use the fact from multivariate calculus that after the change of variables $(dx\, dy)$ becomes $(r\, dr\, d\theta)$ and explain the new limits of integration based on the region being described in the plane).

(c) Compute the integral from (b) to find the value of I.

(d) Use (c) to show that $\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = 1$. (Hint: Use a u-substitution $u = \frac{x-\mu}{\sigma}$).

5.2 CONTINUOUS RANDOM VARIABLES

Just as the move from discrete to continuous spaces required a slight change in the definition of probability space, so it also requires a slight change in the definition of random variable.

In the discrete setting we frequently needed to consider the preimage $X^{-1}(A)$ of a set. Now we need to make sure that such a preimage is a legitimate event.

Definition 5.2.1. *Let $(S, \mathcal{F}, P)$ be a probability space and let $X : S \to \mathbb{R}$ be a function. Then X is a random variable provided that whenever B is an event in $\mathbb{R}$ (i.e. a Borel set), $X^{-1}(B)$ is also an event in $\mathcal{F}$.*

Note that in the discrete setting this extra condition was met trivially as every subset of S was an event. Therefore the discrete setting is simply a special case of this new definition. As with the introduction of σ-fields, we include this definition for completeness. We will only consider functions which meet this criterion. In this section we shall consider only continuous random variables. These are defined next.

Definition 5.2.2. *Let $(S, \mathcal{F}, P)$ be a probability space. A random variable $X : S \to \mathbb{R}$ is called a continuous random variable if there exists a density function $f_X : \mathbb{R} \to \mathbb{R}$ such that for any event A in $\mathbb{R}$,*

$$P(X \in A) = \int_A f_X(x)\, dx. \tag{5.2.1}$$

f_X is called the probability density function of X.

The following lemma demonstrates an elementary property of continous random variables that distinguishes them from discrete random variables.

Lemma 5.2.3. *Let X be a continuous random variable. Then for any $a \in \mathbb{R}$,*

$$P(X = a) = 0 \tag{5.2.2}$$

Proof- Let $a \in \mathbb{R}$, then $P(X = a) = \int_a^a f(x)dx = 0$. ■

Random variables may also be described using a "distribution function" (also commonly known as a "cumulative distribution function").

Definition 5.2.4. *If X is a random variable then its distribution funciton $F : \mathbb{R} \to [0, 1]$ is defined by*

$$F(x) = P(X \leq x). \tag{5.2.3}$$

When it must be emphasized that a distribution function belongs to a particular random variable X the notation $F_X(x)$ will be used to indicate the random variable.

Though a distribution function is defined for any real-valued random variable, there is a special relationship between $f_X(x)$ and $F_X(x)$ when the random variable has a density.

Theorem 5.2.5. *Let X be a random variable with a piecewise continuous density function $f(x)$. If $F(x)$ denotes the distriubtion function of X then*

$$F(x) = \int_{-\infty}^{x} f(x)\,dx. \tag{5.2.4}$$

Moreover, where $f(x)$ is continuous, $F(x)$ is differentiable and $F'(x) = f(x)$.

Proof - By definition $F(x) = P(X \leq x) = P(X \in (-\infty, x])$, but this probability is described in terms of an integral over the density of X, so $F(x) = \int_{-\infty}^{x} f(x)\,dx$.

The result that $F'(x) = f(x)$ then follows from the fundamental theorem of calculus after taking derivatives of both sides of the equation (when such a derivative exists). Note, in particular, that since densities are assumed to be piecewise continuous, their corresponding distribution functions are piecewise differentiable. ■

This theorem will be useful for computation, but it also shows that the distribution of a continuous random variable X is completely determined by its distribution function F_X. That is, if we know $F_X(x)$ and want to calculate $P(X \in A)$ for some set A we could do so by differentiating $F_X(x)$ to find $f_X(x)$ and then integrating this density over the set A. In fact $F_X(x)$ always completely determines the distribution of X (regardless of whether or not X is a continuous random variable), but a proof of that fact is beyond the scope of the course and will not be needed for subsequent results.

5.2.1 *Common Distributions*

In the literature random variables whose distributions satisfy (5.2.1) are called absolutely continuous random variables and those that satisfy (5.2.2)are referred to as continous random variables. Since we shall only consider continuous random variables that satisfy (5.2.1) we refer to them as continous random variables.

There are many continuous distributions that commonly arise. Some of these are continuous analogs of discrete random variables we have already studied. We will define these in the context of continuous random variables having the corresponding distributions. We begin with the already discussed uniform distribution but on an arbitrary interval.

Definition 5.2.6. *$X \sim$ **Uniform**(a,b): Let (a,b) be an open interval. If X is a random variable with its probabilty density function given by*

$$f(x) = \begin{cases} \frac{1}{(b-a)} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

then X is said to be uniformly distributed on (a,b). Note that this is consistant with the example at the beginning of the section since the density of a Uniform$(0,1)$ is one on the interval $(0,1)$ and zero elsewhere. Further, recall that in Exercise 5.1.6 we have shown that f is indeed a probability density function.
Since X only takes values on (a,b) if $x < a$ then $P(X \leq x) = 0$ while if $x > b$ then $P(X \leq x) = 1$. So let $a \leq x \leq b$. Then,

$$P(X \leq x) = \int_{-\infty}^{x} f_X(y)\,dy = \int_{-\infty}^{a} 0\,dy + \int_{a}^{x} \frac{1}{b-a}\,dy = \frac{x-a}{b-a}.$$

Therefore the distribution function for X is

$$F_X(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } x > b \end{cases}$$

Exponential Random Variable

The next continuous distribution we introduce is called the exponential distribution. It is well known from physical experiments that the radioactive isotopes decay to its stable form. Suppose there were $N(0)$ atoms of a certain radioctive material at time 0 then one is interested in the fraction of radioactive material that have not decayed at time $t > 0$. It is observed from experiments that if $N(t)$ is the number of atoms of radioactive material that has not decayed by time t then the fraction

$$\frac{N(t)}{N(0)} \approx e^{-\lambda t},$$

for some $\lambda > 0$. One can introduce a probability model for the above experiment in the following manner. Suppose X represented the time taken by a randomly chosen radioactive

atom to decay to its stable form. The distribution of the random variable X needs to satisfy

$$P(X \geq t) = e^{-\lambda t}, \tag{5.2.5}$$

for $t > 0$. It is possible to define such a random variable.

Definition 5.2.7. $X \sim \mathbf{Exp}(\lambda)$*: Suppose* $\lambda > 0$*. If* X *is a random variable with its probabilty density function given by*

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

it is said to be distributed exponentially with parameter λ*. Recall that in Exercise 5.1.5 we have shown that* f *is indeed a probability density function. Since* X *only takes values on* $(0, \infty)$ *if* $x < 0$ *then* $P(X \leq x) = 0$*. So let* $x \geq 0$*. Then,*

$$P(X \leq x) = \int_{-\infty}^{x} f_X(x)\, dx = \int_{-\infty}^{0} 0\, dx + \int_{0}^{x} \lambda e^{-\lambda y} dy = -e^{-\lambda y} \,|_0^x = 1 - e^{-\lambda x}.$$

Therefore the distribution function for X *is*

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\lambda x} & \text{if } 0 \leq x \end{cases}$$

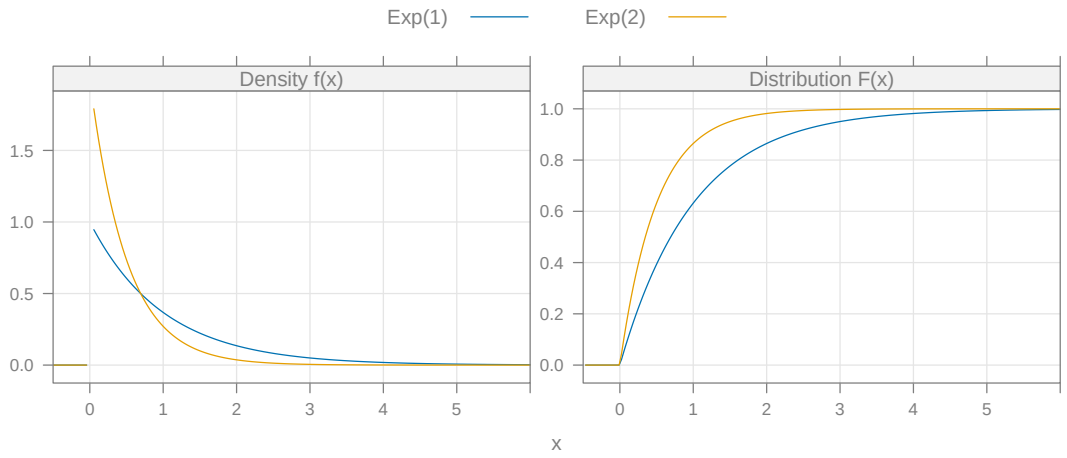


Figure 5.1: The shape of typical Exponential density and cumulative distribution functions.

We have previously seen that geometric random variables have the memoryless property (See (3.2.2)). It turns out that the exponential random variable also possess the *memoryless property* in continuous time. Clearly if $X \sim \text{Exp}(\lambda)$ then $P(X \geq 0) = 1$ and

$$P(X \geq t) = P(X \in [t, \infty)) = \int_t^\infty \lambda e^{-\lambda x}\, dx = -e^{-\lambda x} \,|_t^\infty = e^{-\lambda t},$$

for $t > 0$. Further if $s, t > 0$, $X > s + t$ imples $X > s$. So

$$\begin{aligned} P(X > s + t | X > s) &= \frac{P((X > s + t) \cap (X > s))}{P(X > s)} \\ &= \frac{P(X > s + t)}{P(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t}. \end{aligned}$$

Therefore for all $s, t > 0$

$$P(X > s + t | X > s) = P(X > t) \tag{5.2.6}$$

Thinking of the variables s and t in terms of time, this says that if an exponential random variable has not yet occurred by time s, then its distribution from that time onward continues to be distributed like an exponential random variable with the same parameter. Situations that involve waiting times such as the lifetime of a light bulb or the time spent in a queue at a service counter are often modelled with the exponential distribution. It is a fact (see Exercise 5.2.12) that if a positive continuous random variable has the memoryless property then it necessarily is an exponential random variable.

Example 5.2.8. Let $X \sim \text{Exp}(2)$. Calculate the probability that X produces a value larger than 4.

The density of X is

$$f(x) = \begin{cases} 2e^{-2x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

So, $P(X > 4)$ may be calculated via an integral.

$$\begin{aligned} P(X > 4) &= \int_4^\infty 2e^{-2x}\, dx \\ &= -e^{-2x} \,|_4^\infty = 0 - (-e^{-8}) = e^{-8} \approx 0.000335 \end{aligned}$$

So there is only about a 0.0335% chance of such a result. ■

Normal Random Variable

Of all continuous distributions, The normal distribution (also sometimes called a "Gaussian distribution") is the most fundamental for applictions of statistics as it frequently arises as a limiting distribution of sampling procedures.

Definition 5.2.9. $X \sim \textbf{Normal}(\mu, \sigma^2)$*: Let* $\mu \in \mathbb{R}$ *and let* $\sigma > 0$*. Then* X *is said to be normally distributed with parameters* μ *and* σ^2 *if it has the density*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5.2.7}$$

for all $x \in \mathbb{R}$*. We will prove that* μ *and* σ *are, respectively, the mean and standard deviation of such a random variable (See Definiton 6.1.1, Definition 6.1.9, Example 6.1.11). Recall that in Exercise 5.1.10 we have seen that* f *is a probability density function.*

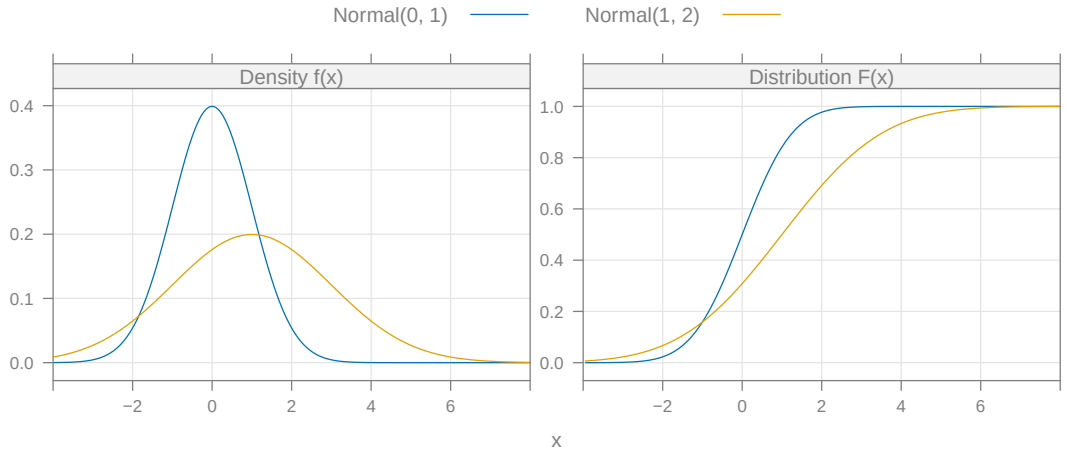


Figure 5.2: The shape of typical Normal density and cumulative distribution functions.

It is observed during statistical experiments that if X were to denote the number of leaves in an apple tree or the height of adult men in a population then X would be close to a normal random variable with approrpriate parameters μ and σ^2. It also arises as a limiting distribution. We shall discuss this aspect in general in Chapter 8, but here we will briefly mention one such limit that appears as an approximation for Binomial probabilities.

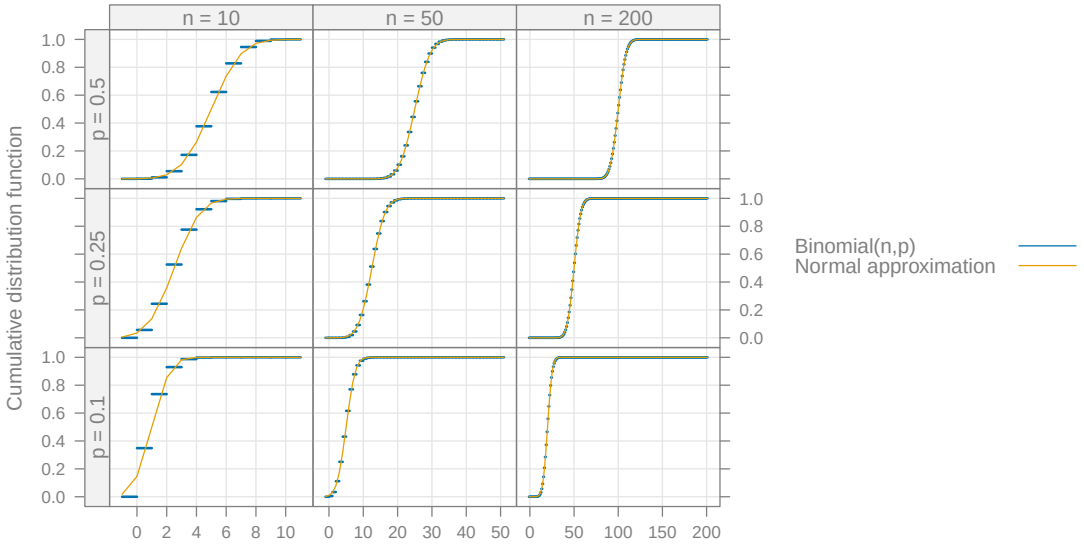


Figure 5.3: The Normal approximation to Binomial.

Suppose we have $X_1, X_2, \ldots, X_n$ are i.i.d Bernoulli (p) random variables. Then we know that $S_n = \sum_{i=1}^{n} X_i$ is a Binomial (n, p) random variable. In Theorem 2.2.2 we saw that for $\lambda > 0$, $k \geq 1$, $0 \leq p = \frac{\lambda}{n} < 1$,

$$\lim_{n\to\infty} P(S_n = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Such an approximation was useful when p was decreasing to zero while n grew to infinity with np remaining constant. The De Moivre-Laplace Central Limit Theorem allows us to consider another form of limit where p remains fixed, but n increases.

Theorem 5.2.10. (De Moivre-Laplace Central Limit Theorem) *Suppose $S_n \sim$ Binomial (n, p), where $0 < p < 1$. Then for any $a < b$*

$$\lim_{n\to\infty} P(a < \frac{S_n - np}{\sqrt{np(1-p)}} \leq b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx \tag{5.2.8}$$

We shall omit the proof of the above Theorem for now. We prove it in a more general setting in Chapter 8. For the students well versed with Real Analysis the proof is sketched in Exercise 5.2.16. We refer the reader to [Ram97] for a detailed discussion of the Theorem 5.2.10.

Calculating Normal Probabilities and Necessity of Normal Tables

In a standard introduction to integral calculus one learns many different techniques for calculating integrals. But there are some functions whose indefinite integral has no closed-form solution in terms of simple functions. The density of a normal random variable is one such function. Because of this if $X \sim \text{Normal}(0,1)$ the probability

$$P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

cannot be expressed exactly in terms of standard functions. Many scientific calculators will have a feature that allows this expression to be evaluated. For example, in R, the command `pnorm(x)` evaluates the integral above. Another common solution in statistical texts is to provide a table of values.

	0.00	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18
0.0	0.500	0.508	0.516	0.524	0.532	0.540	0.548	0.556	0.564	0.571
0.2	0.579	0.587	0.595	0.603	0.610	0.618	0.626	0.633	0.641	0.648
0.4	0.655	0.663	0.670	0.677	0.684	0.691	0.698	0.705	0.712	0.719
0.6	0.726	0.732	0.739	0.745	0.752	0.758	0.764	0.770	0.776	0.782
0.8	0.788	0.794	0.800	0.805	0.811	0.816	0.821	0.826	0.831	0.836
1.0	0.841	0.846	0.851	0.855	0.860	0.864	0.869	0.873	0.877	0.881
1.2	0.885	0.889	0.893	0.896	0.900	0.903	0.907	0.910	0.913	0.916
1.4	0.919	0.922	0.925	0.928	0.931	0.933	0.936	0.938	0.941	0.943
1.6	0.945	0.947	0.949	0.952	0.954	0.955	0.957	0.959	0.961	0.962
1.8	0.964	0.966	0.967	0.969	0.970	0.971	0.973	0.974	0.975	0.976
2.0	0.977	0.978	0.979	0.980	0.981	0.982	0.983	0.984	0.985	0.985

Table 5.1: Table of $\text{Normal}(0,1)$ probabilities. For $X \sim \text{Normal}(0,1)$, the table gives values of $P(X \leq z)$ for various values of z between 0 and 2.18 upto three digits. The value of z for each entry is obtained by adding the corresponding row and column labels.

Table 5.1 gives values only for positive values of z because for negative z, $P(X \leq z)$ can be easily computed using the symmetry of the $\text{Normal}(0,1)$ distribution as (see Figure 5.4)

$$P(X \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \int_{-z}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - P(X \leq -z). \tag{5.2.9}$$

A more complete version of this table is given in the Appendix. A similar computation can be made for other normally distributed random variables by normalizing them. Suppose

Figure 5.4: Computation of Normal$(0,1)$ probabilities as area under the normal density curve. For Normal$(0,1)$ and in fact for any symmetric distribution in general, it is enough to know the distribution function for positive values (see Exercise 5.2.8).

$Y \sim$ Normal (μ, σ^2) and we were interested in finding the distribution function of Y. Observe that

$$P(Y \leq y) \quad = \quad \int_{-\infty}^{y} \frac{1}{\sigma\sqrt{2\pi}} e^{-(z-\mu)^2/2\sigma^2}\, dz.$$

Now perform a change of variables $u = \frac{z-\mu}{\sigma}$ so that $du = \frac{1}{\sigma}\, dz$. This integral then becomes

$$P(Y \leq y) = \int_{-\infty}^{\frac{y-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du = P(X \leq \frac{y-\mu}{\sigma}), \tag{5.2.10}$$

where $X \sim$ Normal $(0,1)$. Now we may use Table 5.1 to compute the distribution function of Y. We conclude this section with two examples.

Example 5.2.11. If $X \sim$ Normal$(0,1)$, how likely is it that X will be within one standard deviation of its expected value?

In this case the expected value of the random variable is zero and the standard deviation is one. Therefore the answer is given by

$$\begin{aligned} P(-1 \leq X \leq 1) &= \int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \\ &= \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx - \int_{-\infty}^{-1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \\ &= P(X \leq 1) - P(X \leq -1) \end{aligned}$$

R tells us that

```
pnorm(1)
```

```
[1] 0.8413447
```

```
pnorm(-1)
```

```
[1] 0.1586553
```

```
pnorm(1) - pnorm(-1)
```

```
[1] 0.6826895
```

Alternatively, using Table 5.1, we see that $P(X \leq 1) = 0.841$ (upto three decimal places), and by symmetry $P(X \leq -1) = P(X \geq 1) = 1 - P(X \leq 1) = 1 - 0.841 = 0.159$. Therefore, $P(-1 \leq X \leq 1) \approx 0.841 - 0.159 = 0.682$. In other words, there is roughly a 68% chance that a standardized normal random variable will produce a value within one standard deviation of expected value. ■

Example 5.2.12. A machine fills bags with cashews. The intended weight of cashews in the bag is 200 grams. Assume the machine has a tolerance such that the actual weight of the cashews is a normally distributed random variables with an expected value of 200 grams and a standard deviation of 4 grams. How likely is it that a bag filled by this machine will have fewer than 195 grams of cashews in it?

We know $Y \sim \text{Normal}(200, 4^2)$ and we want the probability $P(Y < 195)$. By above computation, (5.2.10),

$$P(Y < 195) = P(X < \frac{195 - 200}{4}) = P(X < -\frac{5}{4})$$

where $X \sim \text{Normal}(0, 1)$. If we were to use Table 5.1, we would first obtain

$$P(X < -\frac{5}{4}) = 1 - P(X < \frac{5}{4}) = 1 - P(X < 1.25) = 1 - 0.896 = 0.104$$

Using the R command `pnorm(-5/4)`, we obtain the value 0.1056498. That is, there is slightly more than a 10% chance of a bag this light being produced by the machine. ■

5.2.2 *A Word About Individual Outcomes*

We began this section by noting that continuous random variables must necessarily give probability zero to any single outcome. It is an awkward consequence of this that two different densities may give rise to exactly the same probabilities. For instance, the functions

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$g(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

are different because they assign different values to the points $x = 0$ and $x = 1$. However, these individual points cannot affect the computation of probabilities so both $f(x)$ and $g(x)$ give rise to the same probability distribution. The same thing would occur even if $f(x)$ and $g(x)$ differed in a countably infinite number of points, since these will still have probability zero when taken collectively.

Because of this we will describe $f(x)$ and $g(x)$ as the same density (and sometimes even write $f(x) = g(x)$) when the two densities produce the same probabilities. We do this even when f and g may technically be different functions. Though it is a more restirctive assumption than is necessary, we have required densities to be piecewise continuous. As a consquence of the explanation above, altering the values of the function at the endpoints of intervals of continuity will not change the resulting probabilities and will result in the same density.

EXERCISES

Ex. 5.2.1. Suppose X was continuous random variable with distribution function F. Express the following probabilities in terms of F:

(a) $P(a < X \leq b)$, where $-\infty < a < b < \infty$

(b) $P(a < X < \infty)$ where $a \in \mathbb{R}$.

(c) $P(\mid X - a \mid \geq b)$ where $a, b \in \mathbb{R}$ and $b > 0$.

Ex. 5.2.2. Let $R > 0$ and $X \sim$ Uniform $[0, R]$. Let $Y = \min(X, \frac{R}{10})$. Find the distribution function of Y.

Ex. 5.2.3. Let X be a random variable with distribution function given by

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 < x < \frac{1}{4} \\ \frac{x}{2} + \frac{1}{8} & \text{if } \frac{1}{4} \leq x < \frac{3}{4} \\ 2x - 1 & \text{if } \frac{3}{4} \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

(a) Sketch a graph of the function F.

(b) Use F to calculate : $P([0, \frac{1}{4}))$, $P([\frac{1}{8}, \frac{3}{2}])$, $P((\frac{3}{4}, \frac{7}{8}])$.

(c) Find the probabilty density function of X.

Ex. 5.2.4. Let X be a continuous random variable with distribution function $F : \mathbb{R} \to [0, 1]$. Then $G : \mathbb{R} \to [0, 1]$ given by

$$G(x) = 1 - F(x)$$

is called the reliability function of X or the right tail distribution function of X. Suppose $T \sim$ Exponential(λ) for some $\lambda > 0$, then find the reliability function of T.

Ex. 5.2.5. Let X be a random variable whose probability density function $f : \mathbb{R} \to [0, 1]$ is given by

$$f(x) = \begin{cases} kx^{k-1}e^{-x^k} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the distribution function of X for $k = 2$.

(b) Find the distribution function of X for general k.

The distribution of X is called the Weibull distribution. Figure 5.5 plots the Weibull distribution for selected values of k.

Ex. 5.2.6. Let X be a random variable whose probability density function $f : \mathbb{R} \to [0, 1]$ is given by

$$f(x) = \begin{cases} \frac{2}{\pi R^2}\sqrt{R^2 - x^2} & \text{if } -R < x < R \\ 0 & \text{otherwise} \end{cases}$$

Find the distribution function of X. The distribution of X is called the semicircular distribution (see Figure 5.6).

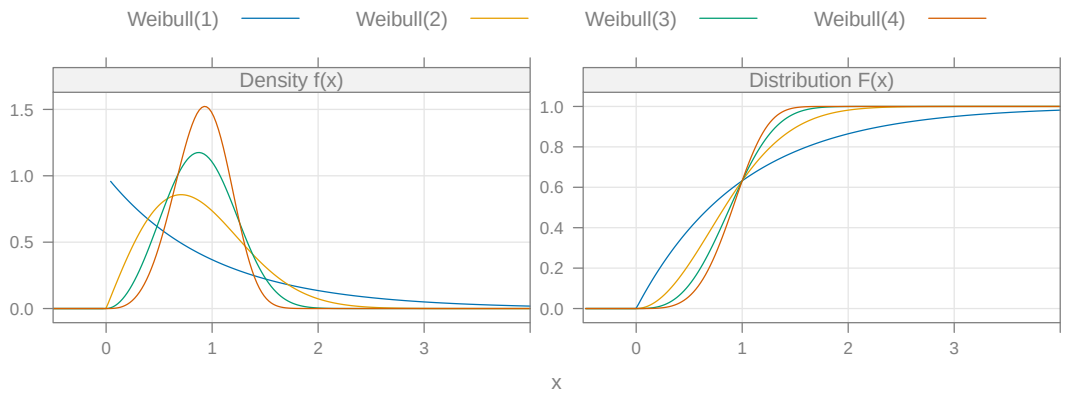


Figure 5.5: The shape of typical Weibull density and cumulative distribution functions.

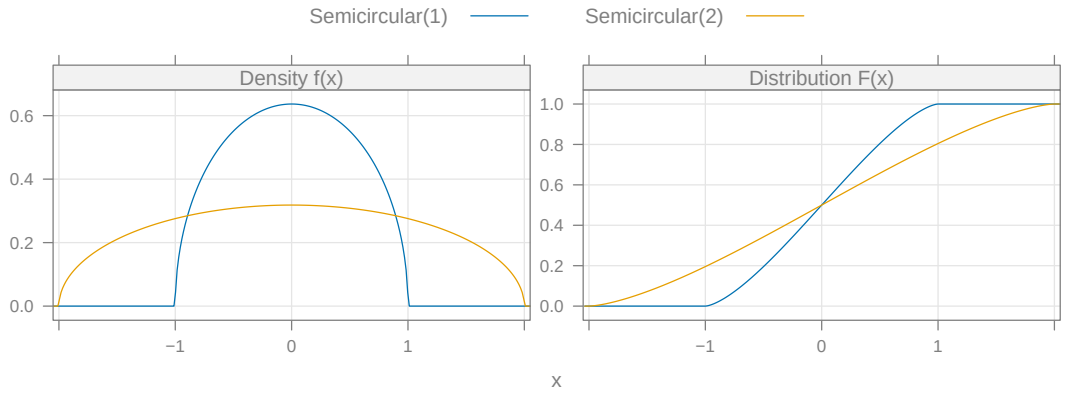


Figure 5.6: The shape of the semicircular density and cumulative distribution functions.

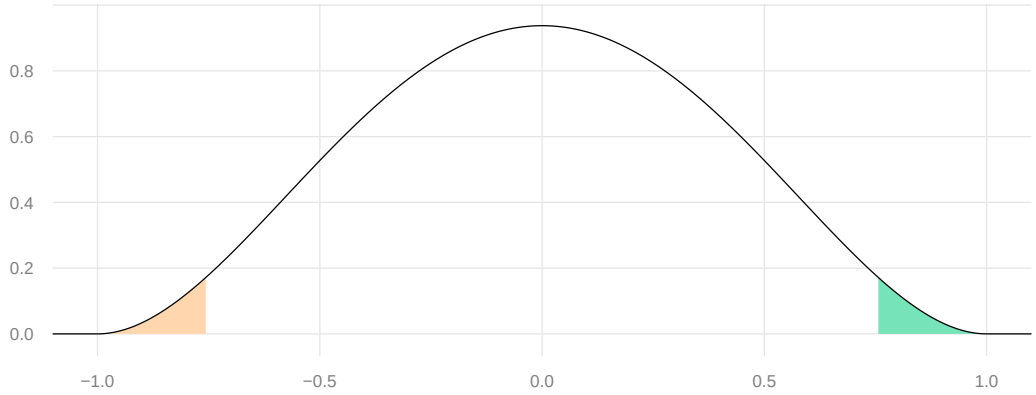


Figure 5.7: Computation of probabilities as area under the density curve. For symmetric distributions, it is enough to know the (cumulative) distribution function for positive values.

Ex. 5.2.7. Let X be a random variable whose distribution function $F : \mathbb{R} \to [0, 1]$ is given by

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \frac{2}{\pi} \arcsin(\sqrt{x}) & \text{if } 0 < x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

Find the probability density function of X. The distribution of X is called the standard arcsine law.

Ex. 5.2.8. Let X be a continuous random variable with probability density function f and distribution function F. Suppose f is a symmetric function, i.e. $f(x) = f(-x)$ for all $x \in \mathbb{R}$. Then show that

(a) $P(X \leq 0) = P(X \geq 0) = \frac{1}{2}$,

(b) for $x \geq 0$, $F(x) = \frac{1}{2} + P(0 \leq X \leq x)$,

(c) for $x \leq 0$, $F(x) = P(X \geq -x) = \frac{1}{2} + P(0 \leq X \leq -x)$.

We have observed this fact for the normal distribution earlier (see Figure 5.7).

Ex. 5.2.9. Let $X \sim \text{Exp}(\lambda)$. The "90th percentile" is a value a such that X is larger than a 90% of the time. Find the 90th percentile of X by determining the value of a for which $P(X < a) = 0.9$.

Ex. 5.2.10. Let X be a continuous random variable such that its distribution function F is strictly increasing on the set $\{x \in \mathbb{R} : 0 < F(x) < 1\}$. The "median" of X is the value of x for which $P(X > x) = P(X < x) = \frac{1}{2}$.

(a) If $X \sim \text{Uniform}(a, b)$ calculate the median of X.

(b) If $Y \sim \text{Exp}(\lambda)$ calcluate the median of Y.

(c) Let $Z \sim \text{Normal}(\mu, \sigma^2)$. Show that the median of Z is μ.

Ex. 5.2.11. Let $X \sim \text{Normal}(\mu, \sigma^2)$. Show that $P(|X - \mu| < k\sigma)$ does not depend on the values of μ or σ. (Hint: Use a change of variables for the appropriate integral).

Ex. 5.2.12. Above we saw that exponential random variables satisfied the memoryless property, (5.2.6). It can be shown that any positive, continuous random variable with the memoryless property must be exponential. Follow the steps below to prove a slightly weakened version of this result. For all parts, suppose X is a positive, continuous random variable with the memoryless property for which the distribution function $F_X(t)$ has a continuous derivative for $t > 0$. Suppose further that $lim_{t\to 0^+} F'(t)$ exists and call this quantity α. Let $G(t) = 1 - F_X(t) = P(X > t)$ and do the following.

(a) Use the memoryless property to show that $G(s + t) = G(s) \cdot G(t)$ for all postiive s and t.

(b) Use part (a) to conclude that $G'(t) = -\alpha G(t)$. (Hint: Take a derivative with respect to s and then take an appropriate limit).

(c) It is a fact (which you may take as granted) that the differential equation from (b) has solutions of the form $G(t) = Ce^{-\alpha t}$. Use the fact that X is positive to explain why it must be that $C = 1$.

(d) Use part (c) to calculate $F_X(t)$ and then differentiate to find $f_X(t)$.

(e) Conclude that X must be exponentially distributed and determine the associated parameter in terms of α.

Ex. 5.2.13. Let X be a random variable with density $f(x) = 2x$ for $0 < x < 1$ (and $f(x) = 0$ otherwise). Calculate the distribution function of X.

Ex. 5.2.14. Let $X \sim \text{Uniform}(\{1, 2, 3, 4, 5, 6\})$. Despite the fact this is a discrete random variable without a density, the distribution function $F_X(x)$ is still defined. Find a piecewise defined expression for $F_X(x)$ (see Figure 5.8 for a plot).

Ex. 5.2.15. Suppose $F : \mathbb{R} \to [0, 1]$ is given by (5.2.3). Then show that

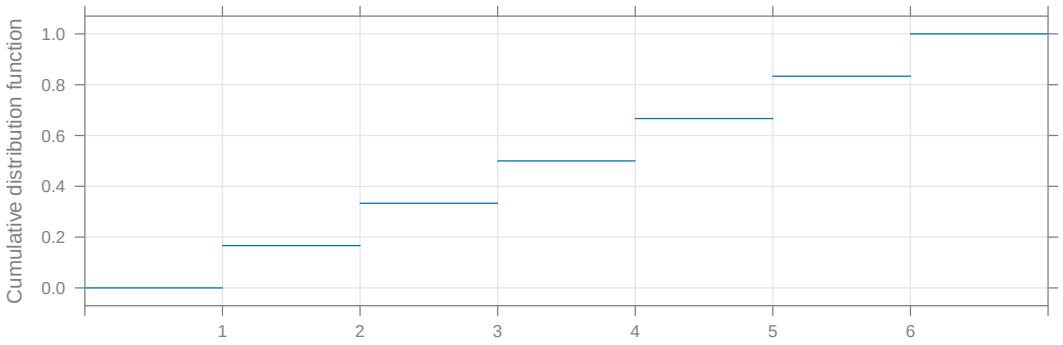


Figure 5.8: The cumulative distribution function for Exercise 5.2.14.

1. F is a monotonically increasing function.

2. $\lim_{x\to\infty} F(x) = 1$.

3. $\lim_{x\to-\infty} F(x) = 0$.

4. if, in addition, F is given by (5.2.4) then F is continuous.

Ex. 5.2.16. We use the notation as in Theorem 5.2.10.

(a) Let

$$A_n = \left\{k : 0 \le k \le n, np + a\sqrt{np(1-p)} \le k \le np + a\sqrt{np(1-p)}\right\}.$$

Show that

$$P(a \le \frac{S_n - np}{\sqrt{np(1-p)}} \le b) = \sum_{k\in A_n} P(S_n = k).$$

(b) Let

$$\xi_{k,n} = \frac{k - np}{\sqrt{np(1-p)}}.$$

Using the definition of the Riemann integral show that

$$\lim_{n\to\infty} \sum_{k\in A_n} \frac{e^{-\frac{\xi_{k,n}^2}{2}}}{\sqrt{2\pi np(1-p)}} = \int_a^b \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$$

(c) Using Stirling's approximation show that

$$\lim_{n\to\infty} \sup_{k\in A_n} \frac{\binom{n}{k}p^k(1-p)^{n-k}}{\sqrt{2\pi np(1-p)}e^{-\frac{\xi_{k,n}^2}{2}}} = 1$$

(d) Prove Theorem 5.2.10 by observing

$$P(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b) =$$
$$\sum_{k\in A_n} \frac{e^{-\frac{\xi_{k,n}^2}{2}}}{\sqrt{2\pi np(1-p)}} + \sum_{k\in A_n} \frac{e^{-\frac{\xi_{k,n}^2}{2}}}{\sqrt{2\pi np(1-p)}} \left(\frac{\binom{n}{k}p^k(1-p)^{n-k}}{\frac{1}{\sqrt{2\pi np(1-p)}}e^{-\frac{\xi_{k,n}^2}{2}}} - 1 \right)$$

5.3 TRANSFORMATIONS OF CONTINUOUS RANDOM VARIABLES

In Section 3.3 we have discussed functions of discrete random variables and how to find their distributions. Suppose $g : \mathbb{R} \to \mathbb{R}$ and $Y = g(X)$, to find the distribution of Y we converted events associated with Y with events of X by inverting the function g. In the setting of continuous random variables distribution functions are used for calculating probabilities associated with functions of a known random variable. We next present a simple example for which $g(x) = x^2$ followed by a result that covers situations when $g(x)$ is any linear function.

Example 5.3.1. Let $X \sim \text{Uniform}(0, 1)$ and let $Y = X^2$. What is the density for Y?

Since X takes values on $(0, 1)$ and since $Y = X^2$, it will also be the case that Y takes values on $(0, 1)$. However, though X is uniform on the interval, there should be no expectation that Y will also be uniform. In fact, since squaring a positive number less than one results in a smaller number than the original, it should seem intuitive that results of Y will be more likely to be near to zero than they are to be near to one.

It is not easy to see how to calculate the density of Y directly from the density of X. However, it is a much easier task to compute the distribution of Y from the distribution of X. Therefore we will use the following plan in the calculation below – integrate $f_X(x)$ to find $F_X(x)$; use $F_X(x)$ to determine $F_Y(y)$; then differentiate $F_Y(y)$ to calculate $f_Y(y)$.

For the first step, note

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\,dx = \begin{cases} 0 & \text{if } 0 < x \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}$$

Next, since Y takes values in $(0,1)$, if $y \leq 0$ then $F_Y(y) = P(Y \leq y) = 0$. But if $y > 0$ then

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}).$$

Since X is always positive, the event $(X < -\sqrt{y})$ has zero probability we may connect this to the distribution of X by writing

$$\begin{aligned} F_Y(y) &= P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= P(X < -\sqrt{y}) + P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= P((X < -\sqrt{y}) \cup (-\sqrt{y} \leq X \leq \sqrt{y})) \\ &= P(X \leq \sqrt{y}) = F_X(\sqrt{y}). \end{aligned}$$

Therefore,

$$F_Y(y) = \begin{cases} 0 & \text{if } 0 \leq y \\ \sqrt{y} & \text{if } 0 < y < 1 \\ 1 & \text{if } y \geq 1 \end{cases}$$

and finally by using the fact that $F'(y) = f(y)$ we can determine that

$$f_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}} & \text{if } 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

As noted in the beginning of this example, this distribution is far from uniform and gives much more weight to intervals close to zero than it does intervals close to one. ■

Lemma 5.3.2. *Let $a \neq 0$ and $b \in \mathbb{R}$. Suppose X is a continuous random variable with probability density function f_X. Let $g(x) = ax + b$ be any non-constant linear function (so $a \neq 0$) and let $Y = g(X)$ then Y is also a continuous random variable whose density function f_Y is given by*

$$f_Y(y) = \frac{1}{\mid a \mid} f_X(\frac{y-b}{a}), \tag{5.3.1}$$

for all $y \in \mathbb{R}$.

Proof- Let $y \in \mathbb{R}$. Assume first that $a > 0$. Then

$$P(Y \leq y) = P(aX + b \leq y) = P(X \leq \frac{y-b}{a}) = \int_{-\infty}^{\frac{y-b}{a}} f_X(z)dz$$

By a simple change of variable $z = \frac{u-b}{a}$ we obtain that

$$P(Y \leq y) = \int_{-\infty}^{y} \frac{1}{a} f_X(\frac{u-b}{a})du. \tag{5.3.2}$$

If $a < 0$ then

$$P(Y \leq y) \quad = \quad P(aX + b \leq y) = P(X \geq \frac{y-b}{a}) = \int_{\frac{y-b}{a}}^{\infty} f_X(z)dz$$

Again a simple change of variable $z = \frac{u-b}{a}$, with $a < 0$, we obtain that

$$P(Y \leq y) = \int_{-\infty}^{y} \frac{1}{-a} f_X(\frac{u-b}{a})du. \tag{5.3.3}$$

Using (5.3.2) and (5.3.3) we have that Y is a continuous random varable with density as in (5.3.1). ■

Lemma 5.3.2 provides a method to standardize the normal random variable.

Corollary 5.3.3. *(a) Let $X \sim Normal(0,1)$ and let $Y = aX + b$ with $a, b \in \mathbb{R}, a \neq 0$. Then, $Y \sim Normal(b, a^2)$.*

(b) Let $X \sim Normal(\mu, \sigma^2)$ and let $Z = \frac{X-\mu}{\sigma}$. Then $Z \sim Normal(0,1)$.

Proof - X has a probability density function given by (5.2.7).
(a)By Lemma 5.3.2, we have that the density of Y is given by

$$f_Y(y) = \frac{1}{\mid a \mid} f_X\left(\frac{y-b}{a}\right) = \frac{1}{\sqrt{2\pi} \mid a \mid} e^{-\frac{(z-b)^2}{2a^2}},$$

for all $y \in \mathbb{R}$. Hence $Y \sim$ Normal (b, a^2).

(b) By Lemma 5.3.2, with $a = \frac{1}{\sigma}$ and $b = -\frac{\mu}{\sigma}$ we have that the density of Z is given by

$$f_Z(z) = \sigma f_X\left(\sigma(z + \frac{\mu}{\sigma})\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}},$$

for all $z \in \mathbb{R}$. Hence $Z \sim$ Normal $(0,1)$. ■

Example 5.3.4. Consider the two parallel lines in $\mathbb{R}^2$, given by $y = 0$ and $y = 1$. Piku is standing at the origin in the plane. She chooses an angle θ uniformly in $(0, \pi)$ and she draws a line segment between the lines $y = 0$ and $y = 1$ at an angle θ from the origin in $\mathbb{R}^2$. Suppose the line segment meets the line $y = 1$ at the point $(X, 1)$. Find the probability density function of X.

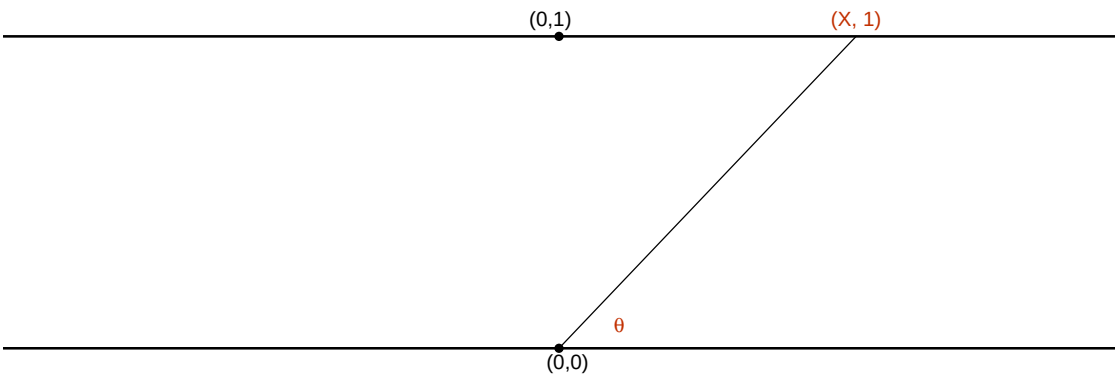


Figure 5.9: Illustration of Example 5.3.4.

First observe that $X = \tan(\frac{\pi}{2} - \theta)$. We shall first find the distribution function of X. Let $x \in \mathbb{R}$. Observe that $\tan(x)$ is a strictly increasing function in the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$ and has an inverse denoted by $\arctan(x)$. So

$$\begin{aligned} P(X \leq x) &= P(\tan(\frac{\pi}{2} - \theta) \leq x) \\ &= P((\frac{\pi}{2} - \theta) \leq \arctan(x)) \\ &= P(\theta \geq \frac{\pi}{2} - \arctan(x)) \\ &= 1 - P(\theta \leq \frac{\pi}{2} - \arctan(x)) \end{aligned}$$

For any $x \in \mathbb{R}$, $\frac{\pi}{2} - \arctan(x) \in (0, \pi)$. As θ has Uniform $(0, \pi)$ distribution, the above is

$$\begin{aligned} &= 1 - \frac{1}{\pi}(\frac{\pi}{2} - \arctan(x)) \\ &= \frac{1}{2} + \frac{1}{\pi}\arctan(x) \end{aligned}$$

Hence the distribution function of X is differentiable and therefore the probability density function of X is given by

$$f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2},$$

for all $x \in \mathbb{R}$. Such a random variable is an example of a Cauchy distribution which we define more generally next. ■

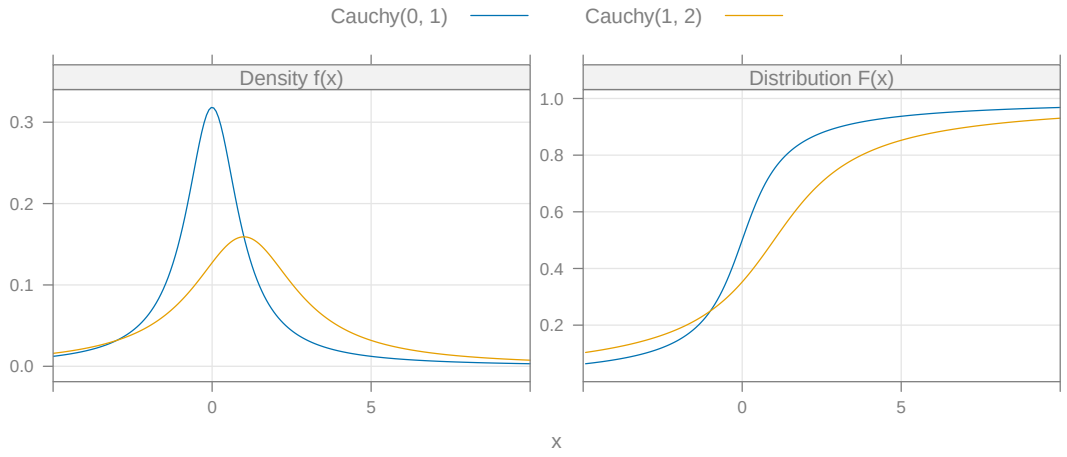


Figure 5.10: The shape of Cauchy density and cumulative distribution functions for selected parameter values.

Definition 5.3.5. *$X \sim$ **Cauchy**(θ, α^2): Let $\theta \in \mathbb{R}$ and let $\alpha > 0$. Then X is said to have a Cauchy distribution with parameters θ and α^2 if it has the density*

$$f(x) = \frac{1}{\pi} \frac{\alpha}{\alpha^2 + (x-\theta)^2} \tag{5.3.4}$$

for all $x \in \mathbb{R}$. Here θ is referred to as the location parameter and α is referred to as the scale parameter. The distribution function of X is given by

$$F(x) = \frac{1}{\pi} \arctan\left(\frac{x-\theta}{\alpha}\right) \tag{5.3.5}$$

Figure 5.10 gives plots of the Cauchy density and distribution functions.

Similar computations as above are useful for simulations. Most computer progamming languages and spreadsheets have a "Random" function designed to approximate a Uniform$(0,1)$ random variable. How could one use such a feature to simulate random variables with other densities? We start with an example.

Example 5.3.6. If $X \sim$ Uniform$(0,1)$, our goal is to find a function $g : (0,1) \to \mathbb{R}$ for which $Y = g(X) \sim$ Exponential (λ). We will try to find such a $g : (0,1) \to \mathbb{R}$ which is strictly increasing so that it has an inverse. This will be important when it comes to relating the distributions of X and Y.

We require Y to Exponential(λ). So the distribution function of Y is

$$F_Y(y) = \begin{cases} 0 & \text{if } y \leq 0 \\ 1 - e^{-\lambda y} & \text{if } y > 0 \end{cases}$$

But

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y))$$

where the final equality comes from our decree that the function g should be strictly increasing. Therefore,

$$F_Y(y) = F_X(g^{-1}(y)).$$

But the distribution function of a uniform random variable has previously been computed. Hence,

$$F_X(g^{-1}(y)) = \begin{cases} 0 & \text{if } g^{-1}(y) \leq 0 \\ g^{-1}(y) & \text{if } 0 < g^{-1}(y) < 1 \\ 1 & \text{if } g^{-1}(y) \geq 1 \end{cases}$$

Thus we are forced to have

$$g^{-1}(y) = 1 - e^{-\lambda y}$$

for $y > 0$. So inverting the above formula, we get $g : (0,1) \to (0,\infty)$ is given by

$$g(x) = -\frac{1}{\lambda}\log(1-x),$$

for $x \in (0,1)$. Hence,

$$X \sim \text{ Uniform}(0,1) \implies -\frac{1}{\lambda}\log(1-X) \sim \text{ Exponential}(\lambda).$$

In conclusion one could view g as the inverse of F_Y, on $(0,\infty)$. It turns out that this is a general result. We state a special case of this in the lemma below. ■

Lemma 5.3.7. *Let $U \sim$ Uniform $(0,1)$ random variable. Let X be a continuous random variable such that its distribution function, F_X, is a strictly increasing continous function. Then*

(a) $Y = F_X^{-1}(U)$ has the same distribution as X.

(b) $Z = F_X(X)$ has the same distribution as U.

Proof- We observe that as F is strictly increasing continuous distribution function $F : \mathbb{R} \to (0,1)$ and Range $(F) = (0,1)$.

(a) We shall verify that Y and X have the same distribution function. Let $y \in \mathbb{R}$, then

$$F_Y(y) = P(Y \leq y) = P(F_X^{-1}(U) \leq y) = P(U \leq F_X(y)) = F_X(y)$$

Hence X and Y have the same distribution.

(b) We shall verify that Z and U have the same distribution function. Let $z \in \mathbb{R}$. If $z \leq 0$ then

$$P(Z \leq z) \quad = \quad P(F(X) \leq z) = 0$$

as $F : \mathbb{R} \to (0,1)$. If $z \geq 1$ then

$$P(Z \leq z) \quad = \quad P(F(X) \leq z) = 1$$

as $F : \mathbb{R} \to (0,1)$. If $0 < z < 1$ then $F^{-1}(z)$ is well defined as Range $(F) = (0,1)$ and

$$P(Z \leq z) \quad = \quad P(F(X) \leq z) = P(X \leq F^{-1}(z)) = F(F^{-1}(z)) = z.$$

Hence Z and U have the same distribution. ■

The previous lemma may be generalized even to the case when F is not strictly increasing. It requires a concept called the generalized inverse. The interested reader will find it discussed in Exercise 5.3.12.

EXERCISES

Ex. 5.3.1. Let $X \sim \text{Uniform}(0,1)$ and let $Y = \sqrt{X}$. Determine the density of Y.

Ex. 5.3.2. Let $X \sim \text{Uniform}(0,1)$ and let $Z = \frac{1}{X}$. Determine the density of Z.

Ex. 5.3.3. Let $X \sim \text{Uniform}(0,1)$. Let $r > 0$ and define $Y = rX$. Show that Y is uniformly distributed on $(0,r)$.

Ex. 5.3.4. Let $X \sim \text{Uniform}(0,1)$. Let $Y = 1 - X$. Show that $Y \sim \text{Uniform}(0,1)$ as well.

Ex. 5.3.5. Let $X \sim \text{Uniform}(0,1)$. Let a and b be real numbers with $a < b$ and let $Y = (b-a)X + a$. Show that $Y \sim \text{Uniform}(a,b)$.

Ex. 5.3.6. Let $X \sim \text{Uniform}(0,1)$. Find a function $g(x)$ (which is strictly increasing) such that the random variable $Y = g(X)$ has density $f_Y(y) = 3y^2$ for $0 < y < 1$ (and $f_Y(y) = 0$ otherwise).

Ex. 5.3.7. Let $X \sim Normal(\mu, \sigma^2)$. Let $g : (-\infty, \infty) \to \mathbb{R}$ be given by $g(x) = x^2$. Find the probability density function of $Y = g(X)$.

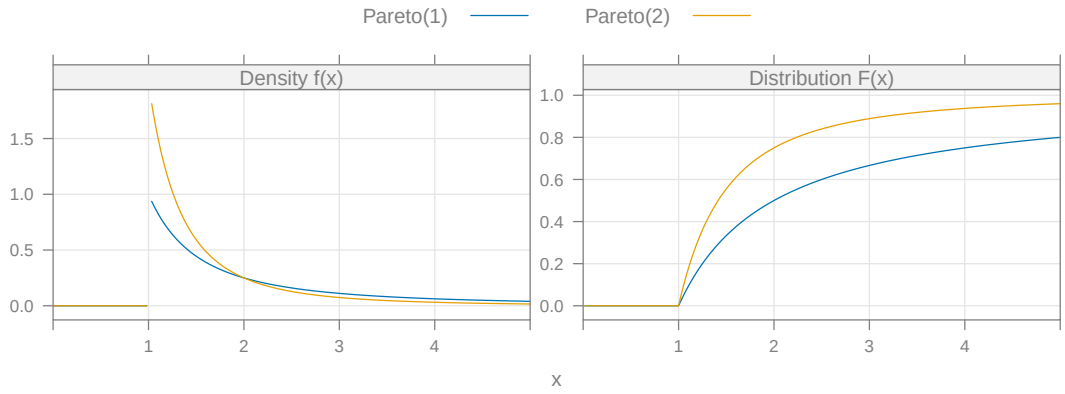


Figure 5.11: The shape of the pareto density and cumulative distribution functions.

Ex. 5.3.8. Let $\alpha > 0$ and X be a random variable with the p.d.f given by

$$f(x) = \begin{cases} \frac{\alpha}{x^{\alpha+1}} & 1 \leq x < \infty \\ 0 & \text{otherwise} \end{cases}$$

The random variable X is said to have Pareto (α) distribution (see Figure 5.11).

(a) Find the distribution of $X_1 = X^2$

(b) Find the distribution of $X_2 = \frac{1}{X}$

(c) Find the distribution of $X_3 = \log(X)$

In the above exercises we assume that the transformation function is defined as above when the p.d.f of X is positive and zero otherwise.

Ex. 5.3.9. Let X be a continuous random variable with probability density function $f_X : \mathbb{R} \to \mathbb{R}$. Let $a > 0$, $b \in \mathbb{R}$ $Y = \frac{1}{a}(X - b)^2$. Show that Y is also a continuous random variable with probability density function $f_Y : \mathbb{R} \to \mathbb{R}$ given by

$$f_Y(y) = \frac{\sqrt{a}}{2\sqrt{y}} \left[f_X(\sqrt{ay} + b) + f_X(-\sqrt{ay} + b) \right]$$

for $y > 0$.

Ex. 5.3.10. Let $-\infty \leq a < b \leq \infty$ and $I = (a, b)$ and $g : I \to \mathbb{R}$. Let X be a continuous random variable whose density f_X is zero on the complement of I. Set $Y = g(X)$.

(a) Let g be a differentiable strictly increasing function.

(i) Show that inverse of g exists and g^{-1} is strictly increasing on $g(I)$.

(ii) For any $y \in \mathbb{R}$, show that $P(Y \leq y) = P(X \leq g^{-1}(y))$

(iii) Show that Y has a density $f_Y(\cdot)$ given by

$$f_Y(y) = f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y).$$

(b) Let g be a differentiable strictly decreasing function.

(i) Show that inverse of g exists and g^{-1} is strictly decreasing on $g(I)$.

(ii) For any $y \in \mathbb{R}$, show that $P(Y \leq y) = 1 - P(X \leq g^{-1}(y))$

(iii) Show that Y has a density $f_Y(\cdot)$ given by

$$f_Y(y) = f_X(g^{-1}(y))\left(-\frac{d}{dy}g^{-1}(y)\right).$$

Ex. 5.3.11. Let X be a random variable having an exponential density. Let $g : [0, \infty) \to \mathbb{R}$ be given by $g(x) = x^{\frac{1}{\beta}}$, for some $\beta \neq 0$. Find the probability density function of $Y = g(X)$.

Ex. 5.3.12. Let $U \sim$ Uniform $(0, 1)$. Let X be a continuous random variable with a distribution function F. Extend $F : \mathbb{R} \to \mathbb{R}$ to $F : \mathbb{R} \cup \{-\infty\} \cup \{\infty\} \to \mathbb{R}$ by setting $F(\infty) = 1$ and $F(-\infty) = 0$. Define the generalised inverse of F, $G : [0, 1] \to \mathbb{R} \cup \{-\infty\} \cup \{\infty\}$ by

$$G(y) = \inf\{x \in \mathbb{R} : F(x) \geq y\}.$$

Show that

(a) Show that for all $y \in [0, 1]$, $F(G(y)) = y$.

(b) Show that for all $x \in \mathbb{R}$ and $y \in [0, 1]$

$$F(x) \geq y \iff x \geq G(y).$$

(c) $Y = G(U)$ has the same distribution as X.

(d) $Z = F(X)$ has the same distribution as U.

5.4 MULTIPLE CONTINUOUS RANDOM VARIABLES

When analyzing multiple random variables at once, one may consider a "joint density" analogous to the joint distribution of the discrete variable case. In this section we will restrict considerations to only two random variables, but we shall see in Chapter 8 that the definitions and results all generalize to any finite collection of variables.

Theorem 5.4.1. *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a non-negative function, piecewise-continuous in each variable for which*

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1.$$

For a Borel set $A \subset \mathbb{R}^2$ define

$$P(A) = \int_A f(x,y)\,dx\,dy.$$

Then P is a probability on $\mathbb{R}^2$ and f is called the density for P.

Proof- The proof of the theorem is essentially the same as in the one-variable version of Theorem 5.1.5. We will not reproduce it here. As in the discrete case we will typically associate such densities with random variables. ■

Definition 5.4.2. *A pair of random variables (X,Y) is said to have a joint density $f(x,y)$ if for every Borel set $A \subset \mathbb{R}^2$*

$$P((X,Y) \in A) = \int_A f(x,y)\,dx\,dy.$$

As in the one-variable case we describe this in terms of "Borel sets" to be precise, but in practice we will only consider sets A which are simple regions in the plane. In fact regions such as $(-\infty, a] \times (-\infty, b]$, for all real numbers a, b are enough to characterise the joint distribution. As in the one variable case we can define a "joint distribution function" of (X,Y) as

$$F_{(X,Y)}(a,b) = P((X \le a) \cap (Y \le b)) = \int_{-\infty}^{a}\int_{-\infty}^{b} f(z,w)dwdz \tag{5.4.1}$$

for all $a, b \in \mathbb{R}$. We will usually denote the joint distribution function by F omiting the subscripts unless it is particularly needed. One can state and prove a similar type of result as Theorem 5.2.5 for $F(a,b)$ when (X,Y) have a joint density. In particular, we can conclude that since the joint densities are assumed to be piecewise continuous, the corresponding distribution functions are piecewise differentiable. Further, the joint distribution of two continuous random variables (X,Y) are completely determined by their joint distribution function F. That is, if we know the value of $F(a,b)$ for all $a, b \in \mathbb{R}$, we could use multivariable calculus to differentiate $F(a,b)$ to find $f(a,b)$. Then $P((X,Y) \in A)$

for any event A is obtained by integrating the joint density f over the event A. We illustrate this with a couple of examples.

Example 5.4.3. Consider the open rectangle in $\mathbb{R}^2$ given by $R = (0,1) \times (3,5)$ and $| R | = 2$ denote its area. Let (X,Y) have a joint density $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x,y) = \begin{cases} \frac{1}{2} & \text{if } (x,y) \in R \\ 0 & \text{otherwise.} \end{cases}$$

The above is clearly a density function. So for any recntangle $A = (a,b) \times (c,d) \subset R$,

$$P((X,Y) \in A) = \int_c^d \int_a^b f(x,y)dxdy = \frac{(b-a)(d-c)}{2} = \frac{| A |}{| R |}.$$

■

In general one can use the following definition to define a uniform random variable on the plane.

Definition 5.4.4. *Let $D \subset \mathbb{R}^2$ be non-empty and with positive area (assume D is a Borel set or in particular f or any simple region whose area is well defined). Then $(X,Y) \sim$Uniform (D) if it has a joint probability density function given by $f : \mathbb{R}^2 \to \mathbb{R}$ given by*

$$f(x,y) = \begin{cases} \frac{1}{|D|} & \text{if } (x,y) \in D \\ 0 & \text{otherwise,} \end{cases}$$

where $| D |$ denotes the area of D.

When $(X,Y) \sim$ Uniform (D) then the probability that (X,Y) lies in a region $A \subset D$ is proportional to the area of A.

Example 5.4.5. Let (X,Y) have a joint density $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x,y) = \begin{cases} x+y & \text{if } 0 < x < 1, 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

We note that this really does describe a density. The function $f(x,y)$ is non-negative and

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy &= \int_0^1\int_0^1 x+y\,dx\,dy\\ &= \int_0^1 (\frac{1}{2}x^2+xy)\,|_{x=0}^{x=1}\,dy\\ &= \int_0^1 \frac{1}{2}+y\,dy\\ &= \frac{1}{2}y+\frac{1}{2}y^2\,|_{y=0}^{y=1}=1.\end{aligned}$$

Calculating a probability such as $P((X<\frac{1}{2})\cap(Y<\frac{1}{2}))$ requires integrating over the appropriate region.

$$\begin{aligned}P((X<\frac{1}{2})\cap(Y<\frac{1}{2})) &= \int_{-\infty}^{1/2}\int_{-\infty}^{1/2} f(x,y)\,dx\,dy\\ &= \int_0^{1/2}\int_0^{1/2} x+y\,dx\,dy\\ &= \int_0^{1/2}\frac{1}{8}+\frac{1}{2}y\,dy\\ &= \frac{1}{8}.\end{aligned}$$

A probability only involving one variable may still be calculated from the joint density. For instance $P(X<\frac{1}{2})$ does not appear to involve Y, but this simply means that Y is unrestircted and the corresponding integral should range over all possible values of Y. Therefore,

$$\begin{aligned}P(X<\frac{1}{2}) &= \int_{-\infty}^{\infty}\int_{-\infty}^{1/2} f(x,y)\,dx\,dy\\ &= \int_0^1\int_0^{1/2} x+y\,dx\,dy=\frac{3}{8}.\end{aligned}$$

It is just as easy to compute that $P(Y<\frac{1}{2})=\frac{3}{8}$. Note that these computations also demonstrate that X and Y are not independent since

$$P(X<\frac{1}{2})\cdot P(Y<\frac{1}{2})=\frac{9}{64}\neq P((X<\frac{1}{2})\cap(Y<\frac{1}{2})).$$

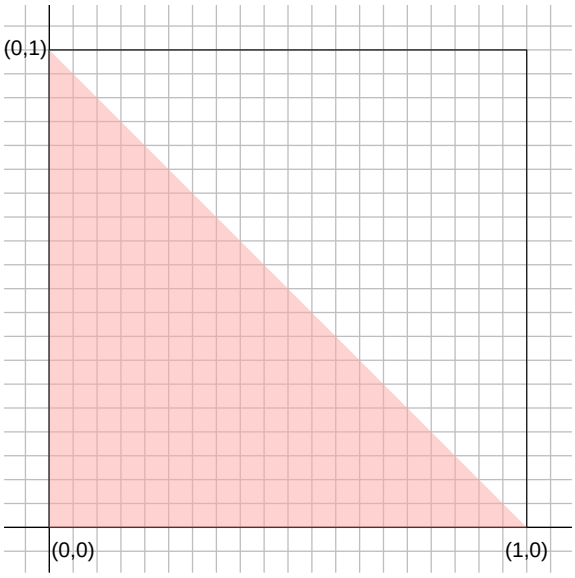


Figure 5.12: The subset A of the unit square that represents the region $x + y < 1$.

A probability such as $P(X + Y < 1)$ can be found by integrating over a non-rectangular region in the plane, as shown in Figure 5.12. Let $A = \{(x, y)|x + y < 1\}$. Then

$$\begin{aligned} P(X+Y<1) &= \int_A f(x,y)\,dx\,dy \\ &= \int_0^1 \int_0^{1-y} x+y\,dx\,dy \\ &= \int_0^1 \frac{1}{2}x^2 + xy\,|_0^{1-y}\;dy \\ &= \int_0^1 \frac{1}{2}(1-y)^2 + (1-y)y\,dy \\ &= \int_0^1 \frac{1}{2} - \frac{1}{2}y^2\,dy \\ &= \frac{1}{3}. \end{aligned}$$

■

5.4.1 *Marginal Distributions*

As in the discrete case, when we begin with the joint density of many random variables, but want to speak of the distribution of an individual variable we will frequently refer to it as a "marginal distribution" .

Suppose (X, Y) are random variables and have a joint probability density function $f : \mathbb{R}^2 \to \mathbb{R}$. Then we obseve that

$$P(X \leq x) = P(X \leq x, -\infty < Y < \infty) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(u, y) dy du.$$

If $g : \mathbb{R} \to \mathbb{R}$ is given by

$$g(u) = \int_{-\infty}^{\infty} f(u, y) dy$$

then

$$P(X \leq x) \int_{-\infty}^{x} g(u) du.$$

Using Theorem 5.2.5, by the continuity assumptions on f, we find that the random variable X is also a continuous random variable with probability density function of X given by

$$f_X(x) = g(x) = \int_{-\infty}^{\infty} f(x, y) dy. \tag{5.4.2}$$

As it was derived from a joint probability density function, the density of X is referred to as the marginal density of X. Similarly one can show that Y is also a continuous random variable and its marginal density is given by

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx. \tag{5.4.3}$$

Example 5.4.6. (Example 5.4.3 contd.) Going back to Example 5.4.3, we can compute the marginal density of X and Y. The marginal density of X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy = \begin{cases} \int_3^5 \frac{1}{2} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases} = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

The marginal density of Y is given by

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx = \begin{cases} \int_0^1 \frac{1}{2} & \text{if } 3 < y < 5 \\ 0 & \text{otherwise.} \end{cases} = \begin{cases} \frac{1}{2} & \text{if } 3 < y < 5 \\ 0 & \text{otherwise.} \end{cases}$$

So we observe that $X \sim$ Uniform $(0, 1)$ and $Y \sim$ Uniform $(3, 5)$. ■

While it is routine to find the marginal densities from the joint density there is no standard way to get to the joint from the marginals. Part of the reason for this difficulty is that the marginal desnitieis offer no information about how the varaibles relate to each other, which is critical information for determining how they behave jointly. However, in the case that the random variables happen to be independent there is a convenient relationship between the joint and marginal densities.

5.4.2 *Independence*

Theorem 5.4.7. *Let f be the joint density of random variables X and Y and let f_X and f_Y be the respective marginal densities. Then*

$$f(x,y) = f_X(x)f_Y(y)$$

if and only if X and Y are independent.

Proof - First suppose X and Y are independent and consider the quantity $P((X \leq x) \cap (Y \leq y))$. On one hand independnece gives

$$P((X \leq x) \cap (Y \leq y)) = P(X \leq x)P(Y \leq y) = F_X(x)F_Y(y) \tag{5.4.4}$$

On the other hand, integrating the joint density yields

$$P((X \leq x) \cap (Y \leq y)) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x,y)\,dx\,dy. \tag{5.4.5}$$

Since equations 5.4.4 and 5.4.5 are equal we may differentiate both with respect to each of the variables x and y and they remain equal. However, differentiating the former gives $f_X(x)f_Y(y)$ because of the relationship between the distribution and the density, while differentiating the latter yields $f(x,y)$ by a two-fold application of the fundamental theorem of calculus.

To prove the opposite direction, suppose $f(x,y) = f_X(x)f_Y(y)$. Let A and B be Borel sets in $\mathbb{R}$. Then

$$\begin{aligned} P((X \in A) \cap (Y \in B)) &= \int_B\int_A f(x,y)\,dx\,dy \\ &= \int_B\int_A f_X(x)f_Y(y)\,dx\,dy \\ &= \left(\int_A f_X(x)\,dx\right)\left(\int_B f_Y(y)\,dy\right) \\ &= P(X \in A)P(Y \in B) \end{aligned}$$

Since this is true for all sets such sets A and B, the variables X and Y are independent. ■

Example 5.4.8. (Example 5.4.3 contd.) We had observed that if $(X,Y) \sim$ Uniform (R) then $X \sim$ Uniform $(0,1)$ and $Y \sim$ Uniform $(3,5)$. Note further that

$$f(x,y) = f_X(x)f_Y(y)$$

for all $x, y \in R$. Consequently X, Y are independent as well. ■

It is tempting to generalise and say that $(X, Y) \sim$ Uniform (D) for a region D with non-trivial area then X and Y would be independent. This is not the case, we illustrate in the example below.

Example 5.4.9. Consider the open disk in $\mathbb{R}^2$ given by $C = \{(x, y) : x^2 + y^2 < 25\}$ and $| C | = 25\pi$ denote its area. Let (X, Y) have a joint density $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x, y) = \begin{cases} \frac{1}{|C|} & \text{if } (x, y) \in C \\ 0 & \text{otherwise.} \end{cases}$$

As before for any Borel $A \subset C$,

$$P((X, Y) \in A) = \frac{| A |}{| C |},$$

and the probability that (X, Y) lies in A is proportional to the area of A. However the marginal density calculation is a little different. The marginal density of X is given by

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x, y)dy = \begin{cases} \int_{-\sqrt{25-x^2}}^{\sqrt{25-x^2}} \frac{1}{| C |} dy & \text{if } -5 < x < 5 \\ 0 & \text{otherwise.} \end{cases} \\ &= \begin{cases} \frac{2}{25\pi}\sqrt{25 - x^2} & \text{if } -5 < x < 5 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The distribution of X is the Semi-circular law described in Exercise 5.2.6. As the joint density f is symmetric in x and y (i.e $f(x, y) = f(y, x)$) the marginal density of Y is the same as that of X (why ?). It is easy to see

$$\frac{1}{25\pi} = f(0, 0) \neq f_X(0)f_Y(0) = \frac{4}{25\pi^2}$$

Consequently X, Y are not independent. This fact should make intuitive sense as well, for if X happens to take a value near 5 or -5 the range of possible values of Y is much more restricted than if X takes a value near 0. ■

We shall see the utility of independence when computing distributions of various functions of independent random variables (see Section 5.5). Independence of random variables also makes it easier to compute their joint density and hence probabilites. For instance, consider the following example.

Example 5.4.10. Suppose $X \sim$ Exponential(λ_1), $Y \sim$ Exponential(λ_2) are independent random variables. Find $P(X - Y < 0)$.

The joint density of (X, Y) is given by

$$f(x,y) = f_X(x)f_Y(y) = \begin{cases} \lambda_1\lambda_2 e^{-(\lambda_1 x+\lambda_2 y)} & \text{if } x > 0 \text{ and } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

Therefore

$$\begin{aligned} P(X-Y<0) &= \int_0^\infty \int_0^y \lambda_1\lambda_2 e^{-(\lambda_1 x+\lambda_2 y)} dxdy = \lambda_1\lambda_2 \int_0^\infty e^{-\lambda_2 y}[\int_0^y e^{-\lambda_1 x} dx]dy \\ &= \lambda_1\lambda_2 \int_0^\infty e^{-\lambda_2 y} \frac{1}{\lambda_1}[1-e^{-\lambda_1 y}]dy \\ &= \lambda_2 \left[\int_0^\infty e^{-\lambda_2 y} - e^{-(\lambda_1+\lambda_2)y} dy\right] \\ &= \lambda_2 \left[\frac{-1}{\lambda_2}(e^{-\lambda_2 y}\ |_0^\infty) + \frac{1}{\lambda_1+\lambda_2}(e^{-(\lambda_1+\lambda_2)y}\ |_0^\infty)\right] \\ &= \lambda_2 \left[\frac{1}{\lambda_2} - \frac{1}{\lambda_1+\lambda_2}\right] \\ &= \frac{\lambda_1}{\lambda_1+\lambda_2}. \end{aligned}$$

Similarly one can also compute $P(Y-X<0) = \frac{\lambda_2}{\lambda_1+\lambda_2}$. This fact is quite useful when using exponential random variables to model waiting times, for $P(X-Y<0) = P(X<Y)$, so we have determined the probability that one waiting time will be shorter than another. ■

5.4.3 *Conditional Density*

In Section 3.2.2 we have seen the notion of conditional distributions for discrete random variables and in Section 4.4 we have seen the notions of conditional expectation and variance for discrete random variables. Suppose X measures the parts per million of a particulate matter less than 10 microns in the air and Y is the incidence rate of asthma in the population. It is clear that X and Y ought to be related; for the distribution of one affects the other. Towards this, in this section we shall discuss conditional distributions for two continuous random variables having a joint probability density function. We recall from Definition 3.2.5 that if X is a random variable on a sample space S and $A \subset S$ be an event such that $P(A) > 0$, then the probability Q described by

$$Q(B) = P(X \in B|A)$$

is called the conditional distribution of X given the event A.

Suppose X and Y have a joint probability density function f. Given our discussion for discrete random variables it is natural to characterise the conditional distribution of X given some information on Y. In the discrete setting we typically considered an event $A = \{Y = b\}$ for some real number b in the range of Y. In the continuous setting such an event A would have zero probability, so the usual way of conditioning on an event would not be possible. However, there is a way to make such a conditioning meaningful and precise provided $f_Y(b) > 0$, where f_Y is the marginal density of Y.

Suppose we wish to find the following :

$$P(X \in [3,4] \mid Y = b).$$

We shall argue heuristically and arrive at an expression for the above probability. Suppose the marginal density of X is $f_X(\cdot)$, and that of Y is $f_Y(\cdot)$. Assume first that f_Y is piecewise continuous and $f_Y(b) > 0$. Then it is a standard fact from real analysis to see that

$$P(Y \in [b, b + \frac{1}{n})) > 0,$$

for all $n \geq 1$. One can then view the conditional probability as before, that is

$$\begin{aligned} P(X \in [3,4] \mid X \in [b, b+\frac{1}{n})) &= \frac{P(X \in [3,4] \cap X \in [b, b+\frac{1}{n}))}{P(X \in [b, b+\frac{1}{n}))} \\ &= \frac{\int_3^4 \left(\int_b^{b+\frac{1}{n}} f(u,v)du \right) dv}{\int_b^{b+\frac{1}{n}} f_X(u)du} \\ &= \frac{\int_3^4 \left(n \int_b^{b+\frac{1}{n}} f(u,v)du \right) dv}{n \int_b^{b+\frac{1}{n}} f_X(u)du} \end{aligned}$$

From facts in real analysis (under some mild assumptions on f) the following can be established,

$$\lim_{n\to\infty} n \int_b^{b+\frac{1}{n}} f(u,v)du = f(b,v),$$

for all real numbers v and

$$\lim_{n\to\infty} n \int_b^{b+\frac{1}{n}} f_X(u)du = f_X(b).$$

We have seen earlier (see Exercise 1.1.13 (b))

$$\lim_{n\to\infty} P(Y \in [b, b+\frac{1}{n})) = P(Y = b).$$

Hence it would be reasonable to argue that $P(X \in [3,4] \mid Y = b)$ ought to be defined as

$$P(X \in [3,4] \mid Y = b) = \frac{\int_3^4 f(b,v)dv}{f_Y(b)}.$$

With the above motivation we are now ready to define conditional densities for two random variables.

Definition 5.4.11. *Let (X,Y) be random variables having joint density f. Let the marginal density of Y be $f_Y(\cdot)$. Suppose b is a real number such that $f_Y(b) > 0$ and is continuous at b then conditional density of X given $Y = b$ is given by*

$$f_{X|Y=b}(x) = \frac{f(x,b)}{f_Y(b)} \tag{5.4.6}$$

for all real numbers x. Similarly, let the marginal density of X be $f_X(\cdot)$. Suppose a is a real number such that $f_X(a) > 0$ and is continuous at a then conditional density of Y given $X = a$ is given by

$$f_{Y|X=a}(y) = \frac{f(a,y)}{f_X(a)}$$

for all real numbers y.

This definition genuinely defines a probability density function, for $f_{X|Y=b}(x) \geq 0$ since it is the ratio of a non-negative quantity and a positive quantity. Moreover,

$$\begin{aligned}\int_{-\infty}^{\infty} f_{X|Y=b}(x)dx &= \int_{-\infty}^{\infty} \frac{f(x,b)}{f_Y(b)}dx \\ &= \frac{1}{f_Y(b)}\int_{-\infty}^{\infty} f(x,b)dx = \frac{1}{f_Y(b)} f_Y(b) = 1\end{aligned}$$

Note that if X and Y are independent then

$$f_{X|Y=b}(x) = \frac{f(x,b)}{f_Y(b)} = \frac{f_X(x)f_Y(b)}{f_Y(b)} = f_X(x).$$

One can use the conditional density to compute the conditional probabilities, namely if (X, Y) are random variables having joint density f and b is a real number such that its marginal density has the property $f_Y(b) > 0$ then

$$P(X \in A \mid Y = b) = \int_A f_{X|Y=b}(x)dx = \int_A \frac{f(x,b)}{f_Y(b)}dx.$$

We conclude this section with two examples where we compute conditional densities. In both the examples the dependencies between the random variables imply that the conditional distributions are different from the marginal distributions.

Example 5.4.12. Let (X, Y) have joint probability density function f given by

$$f(x,y) = \frac{\sqrt{3}}{4\pi} e^{-\frac{1}{2}(x^2-xy+y^2)} \quad -\infty < x, y < \infty.$$

Let $x \in \mathbb{R}$, then the marginal density of X at x is given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_{-\infty}^{\infty} \frac{\sqrt{3}}{4\pi} e^{-\frac{1}{2}(x^2-xy+y^2)}dy$$

By a standard completing the square computation, $\frac{1}{2}(x^2 - xy + y^2) = \frac{3x^2}{8} + \frac{1}{2}(y - \frac{x}{2})^2$. Therefore,

$$f_X(x) = \frac{\sqrt{3}}{4\pi} e^{-\frac{3x^2}{8}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(y-\frac{x}{2})^2}dy$$

Observing that $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\frac{x}{2})^2}dy = 1$ (why ?), we have

$$f_X(x) = \frac{\sqrt{3}}{4\pi} e^{-\frac{3x^2}{8}} \sqrt{2\pi} = \sqrt{\frac{3}{4}} \frac{1}{\sqrt{2\pi}} e^{-\frac{3x^2}{8}}$$

Hence X is a Normal random variable with mean 0 and variance $\frac{4}{3}$. By symmetry (or calculating similarly as above) we can also show that Y is a Normal random variable with mean 0 and variance $\frac{4}{3}$. Also, we can easily see that

$$f_X(x)f_Y(y) = \frac{3}{8\pi} e^{-\frac{3}{8}(x^2+y^2)} \neq \frac{\sqrt{3}}{4\pi} e^{-\frac{1}{2}(x^2-xy+y^2)} = f(x,y)$$

for many $x, y \in \mathbb{R}$. Hence X and Y are not independent. Note that $f_X(x) \neq 0$ for all real numbers x and is continuous at all $x \in \mathbb{R}$. Fix $x \in \mathbb{R}$, the conditional density of Y given $X = x$ is given by

$$f_{Y|X=x}(y) = \frac{f(x,y)}{f_X(x)} = \frac{\frac{\sqrt{3}}{4\pi} e^{-\frac{1}{2}(x^2 - xy + y^2)}}{\sqrt{\frac{3}{4}} \frac{1}{\sqrt{2\pi}} e^{-\frac{3x^2}{8}}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y - \frac{x}{2})^2} \quad \forall\ y \in \mathbb{R}.$$

Hence though the marginal distribution of Y is Normal$(0, \frac{4}{3}$, the)the conditional distribution of Y given $X = x$ is Normal with mean $\frac{x}{2}$ and variance 1. Put another way, if we are given that $X = x$ the mean of Y changes from 0 to x and the variance reduces from $\frac{4}{3}$ to 1.

Such a pair (X, Y) is an example of a bivariate normal random variable and will be discussed in detail in Section 6.4. ■

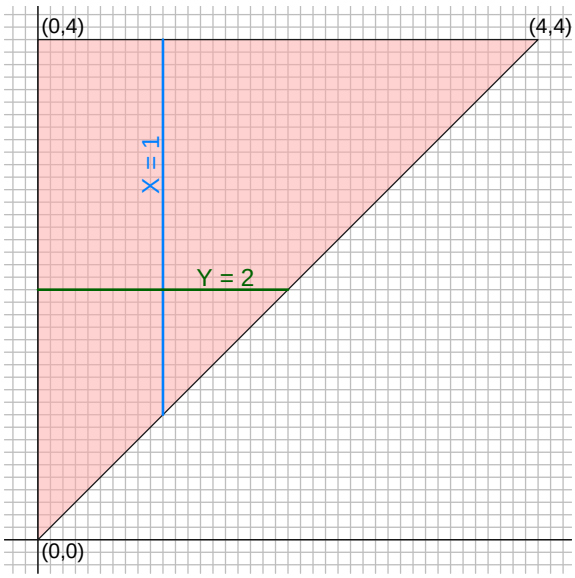


Figure 5.13: The region $T = \{(x, y) \mid 0 < x < y < 4\}$ from Example 5.4.13.

Example 5.4.13. Suppose $T = \{(x, y) \mid 0 < x < y < 4\}$ and let $(X, Y) \sim$ Uniform (T). Therefore its joint density is given by (see Figure 5.13)

$$f(x,y) = \begin{cases} \frac{1}{8} & \text{if } (x,y) \in T \\ 0 & \text{otherwise.} \end{cases}$$

The marginal density of X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \begin{cases} \int_x^4 \frac{1}{8} dy & \text{if } 0 < x < 4 \\ 0 & \text{otherwise.} \end{cases} = \begin{cases} \frac{4-x}{8} & \text{if } 0 < x < 4 \\ 0 & \text{otherwise.} \end{cases}$$

The marginal density of Y is given by

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \begin{cases} \int_0^y \frac{1}{8}dy & \text{if } 0 < y < 4 \\ 0 & \text{otherwise.} \end{cases} = \begin{cases} \frac{y}{8} & \text{if } 0 < y < 4 \\ 0 & \text{otherwise.} \end{cases}$$

Let us fix $0 < b < 4$. So $f_Y(\cdot)$ is non-zero at b and is continuous at b. The conditional density of $(X \mid Y = b)$ is given by

$$f_{X|Y=b}(x) = \frac{f(x,b)}{f_Y(b)} = \begin{cases} \frac{1/8}{b/8} & \text{if } 0 < x < b \\ 0 & \text{otherwise.} \end{cases} = \begin{cases} \frac{1}{b} & \text{if } 0 < x < b \\ 0 & \text{otherwise.} \end{cases}$$

Therefore $(X \mid Y = b) \sim$ Uniform $(0, b)$. Similarly if we fix $0 < a < 4$, we observe $f_X(\cdot)$ is non-zero at a and is continuous at a. The conditional density of $(Y \mid X = a)$ is given by

$$f_{Y|X=a}(y) = \frac{f(a,y)}{f_X(a)} = \begin{cases} \frac{1/8}{(4-a)/8} & \text{if } a < y < 4 \\ 0 & \text{otherwise.} \end{cases} = \begin{cases} \frac{1}{4-a} & \text{if } a < y < 4 \\ 0 & \text{otherwise.} \end{cases}$$

Therefore $(Y \mid X = a) \sim$ Uniform $(a, 4)$.

Clearly X and Y are continuous random variables with distributions that are not uniform, but the conditional distributions turn out to be uniform. ■

EXERCISES

Ex. 5.4.1. Let (X, Y) be random variables whose probability density function is given by $f : \mathbb{R}^2 \to \mathbb{R}$. Find the probability density function of X and probability density function of Y in each of the following cases:-

(a) $f(x,y) = (x+y)$ if $0 \le x \le 1, 0 \le y \le 1$ and 0 otherwise

(b) $f(x,y) = 2(x+y)$ if $0 \le x \le y \le 1$ and 0 otherwise

(c) $f(x,y) = 6x^2y$ if $0 \le x \le 1, 0 \le y \le 1$ and 0 otherwise

(d) $f(x,y) = 15x^2y$ if $0 \le x \le y \le 1$ and 0 otherwise

Ex. 5.4.2. Let $c > 0$. Suppose that X and Y are random variables with joint probability density

$$f(x,y) = \begin{cases} c(xy+1) & \text{if } 0 \le x \le 1 \text{ and } 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find c.

(b) Compute the marginal densities $f_X(\cdot)$ and $f_Y(\cdot)$ and the conditional density $f_{X|Y=b}(\cdot)$

Ex. 5.4.3. Let $A = \{(x, y) \in \mathbb{R}^2 : x > 0, y > 0, x + y < 1\}$ and let X and Y be random variables defined by the joint density $f(x, y) = 24xy$ if $(x, y) \in A$ (and $f(x, y) = 0$ otherwise).

(a) Verify the claim that $f(x, y)$ is a density.

(b) Show that X and Y are dependent random variables.

(c) Explain why (b) doesn't violate Theorem 5.4.7 despite the fact that $24xy$ is a product of a function of x with a function of y.

Ex. 5.4.4. Consider the set $D = [-1, 1] \times [-1, 1]$. Let

$$L = \{(x, y) \in D : x = 0 \text{ or } \text{ or } x = -1 \text{ or } x = 1 \text{ or } y = 0 \text{ or } y = 1 \text{ or } y = -1\}$$

be the lines that create a tiling of D. Suppose we drop a coin of radius R at a uniformly chosen point in D what is the probability that it will intersect the set L ?

Ex. 5.4.5. Let X and Y be two independent uniform $(0, 1)$ random variables. Let $U = \max(X, Y)$ and $V = \min(X, Y)$.

(a) Find the joint distribution of U, V.

(b) Find the conditional distribution of $(V \mid U = 0.5)$

Ex. 5.4.6. Suppose X is a random variable with density

$$f(x) = \begin{cases} cx^2(1 - x) & \text{for } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find:

(a) the value of c.

(b) the distribution function of X.

(c) the conditional probability $P(X > 0.2 \mid X < 0.5)$.

Ex. 5.4.7. Suppose $g : \mathbb{R} \to \mathbb{R}$ be a continuous probability density function, such that $g(x) = 0$ when $x \notin [0, 1]$. Let $D \subset \mathbb{R}^2$ be given by

$$D = \{(x, y) : x \in \mathbb{R} \text{ and } 0 \le y \le g(x)\}$$

Let (X, Y) be uniformly distributed on D. Find the probability density function of X.

Ex. 5.4.8. Continuous random variables X and Y have a joint density

$$f(x,y) = \begin{cases} \frac{1}{24}, & \text{for } 0 < x < 6, 0 < y < 4 \\ 0, & \text{elsewhere.} \end{cases}$$

(a) Find $P(2Y > X)$.

(b) Are X and Y independent?

Ex. 5.4.9. Let

$$f(x,y) = \begin{cases} \eta(y-x)^{\gamma} & \text{if } 0 \le x < y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) For what values of γ can η be chosen so that f be a joint probability density function of X, Y.

(b) Given a γ from part (a), what is the value of η ?

(c) Given a γ and η from parts (a) and (b), find the marginal densities of X and Y.

Ex. 5.4.10. Let $D = \{(x,y) : x^3 \le y \le x\}$. A point (X,Y) is chosen uniformly from D. Find the joint probability density function of X and Y.

Ex. 5.4.11. Let X and Y be two random variables with the joint p.d.f given by

$$f(x,y) = \begin{cases} ae^{-by} & 0 \le x \le y \\ \\ 0 & \text{otherwise} \end{cases}$$

Find a conditions on a and b that make this a joint probability density function.

Ex. 5.4.12. Suppandi and Meera plan to meet at Gopalan Arcade between 7pm and 8pm. Each will arrive at a time (independent of each other) uniformly between 7pm and 8pm and will wait for 15 minutes for the other person before leaving. Find the probability that they will meet ?

5.5 FUNCTIONS OF INDEPENDENT RANDOM VARIABLES

In Section 5.3 we have seen how to compute the distribution of $Y = g(X)$ from the distribution of X for various $g : \mathbb{R}^2 \to \mathbb{R}$. Suppose (X,Y) are random variables having a joint probability density function $f : \mathbb{R}^2 \to \mathbb{R}$. Let $h : \mathbb{R}^2 \to \mathbb{R}$. A natural follow up objective is then to determine the distribution of

$$Z = h(X,Y).$$

In Section 3.3 we discussed an approach to this question when the random variables where discrete.

One could prove a result as attained in Exercise 5.3.10 for functions of two variables but this will require knowledge of Linear Algebra and multivariable calculus. Here we limit our objective and shall focus on two specific functions namely the sum and the product.

5.5.1 *Distributions of Sums of Independent Random variables*

Let X and Y be two independent continous random variables with densities f_X and f_Y. In this section we shall see how to compute the distribution of $Z = X + Y$. We first prove a proposition that describes the probability density function of Z.

Proposition 5.5.1. *(Sum of two independent random variables) Let X and Y be two independent random variables with marginal densities given by $f_X : \mathbb{R} \to \mathbb{R}$ and $f_Y : \mathbb{R} \to \mathbb{R}$. Then $Z = X + Y$ has a probability density function $f_Z : \mathbb{R} \to \mathbb{R}$ given by*

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx. \tag{5.5.1}$$

Proof- Let us first find an expression for the distribution function of Z.

$$\begin{aligned} F(z) &= P(Z \leq z) \\ &= P(X + Y \leq z) \\ &= \int\int_{\{(x,y):x+y\leq z\}} f_X(x) f_Y(y) dy dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_X(x) f_Y(y) dy dx \\ &= \int_{-\infty}^{z} [\int_{-\infty}^{\infty} f_X(x) f_Y(u - x) dx] du. \end{aligned}$$

As $f_X(\cdot)$ and $f_Y(\cdot)$ are densities, it can be shown that the integrand is a piecewise continuous function. Hence F is of the form (5.2.4) and Theorem 5.2.5 implies that the probability density function of Z is given by (5.5.1). ■

The integral expression on the right hand side of (5.5.1) is referred to as the convolution of f_X and f_Y and is denoted by $f_X \star f_Y(z)$. It is a property of convolutions that $f_X \star f_Y(z) = f_Y \star f_X(z)$ for all $z \in \mathbb{R}$. Thus if we view the sum of X and Y as $Z = X + Y$ or $Z = Y + X$ the distribution will be the same (See Exercise 5.5.8).

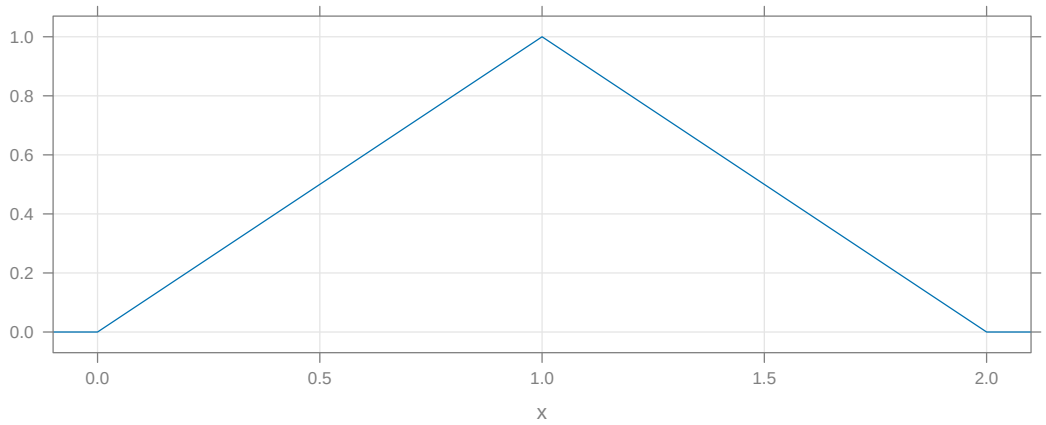


Figure 5.14: The region $T = \{(x, y) \mid 0 < x < y < 4\}$ from Example 5.5.2.

Example 5.5.2. (Sum of Uniforms) Let X and Y be two independent Uniform $(0, 1)$ random variables. Let $Z = X + Y$. From the above proposition that Z has a density given by (5.5.1). Note that

$$f_X(x)f_Y(z-x) = \begin{cases} 1 & \text{if } 0 < x < 1, 0 < z - x < 1 \text{ and } 0 < z < 2 \\ 0 & \text{otherwise} \end{cases}$$

Therefore $f_X(x)f_Y(z-x)$ is non-zero if and only if $\max\{0, z-1\} < x < \min\{1, z\}, 0 < z < 2$. So for $0 < z < 2$,

$$f_Z(z) = \int_{\max\{0,z-1\}}^{\min\{1,z\}} f_X(x)f_Y(z-x)dx = \int_{\max\{0,z-1\}}^{\min\{1,z\}} 1dx = \min\{1, z\} - \max\{0, z-1\}.$$

Therefore,

$$f_Z(z) = \begin{cases} \min\{1, z\} - \max\{0, z-1\} & \text{if } 0 < z < 2 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} z & \text{if } 0 < z \leq 1 \\ 2 - z & \text{if } 1 < z < 2 \\ 0 & \text{otherwise} \end{cases}$$

A graph of this density is displayed in Figure 5.14. ■

Our next example will deal with sum of two independent exponential random variables. This will lead us to the Gamma distribution which is of significant interest in statistics.

Example 5.5.3. (Sum of Exponentials) Let $\lambda > 0$, X and Y be two independent Exponential (λ) random variables. Let $Z = X + Y$. Then we know and Z has a density given by (5.5.1). Further,

$$f_X(x)f_Y(z-x) = \begin{cases} \lambda^2 e^{-\lambda x}e^{-\lambda(z-x)} & \text{if } x \geq 0, z-x \geq 0 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} \lambda^2 e^{-\lambda z} & \text{if } x \geq 0, x \leq z, z \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Hence $f_X(x)f_Y(z-x)$ is non-zero if and only if $0 \leq x \leq z$. So

$$f_Z(z) \quad = \quad \int_0^z f_X(x)f_Y(z-x)dx = \lambda^2 e^{-\lambda z}\int_0^z 1dx = \lambda^2 z e^{-\lambda z},$$

for $z \geq 0$ and $f_Z(z) = 0$ otherwise. This is known as Gamma $(2, \lambda)$ distribution. ■

Before we define the Gamma distribution more generally we prove a lemma in real analysis, the proof of which can be skipped upon first reading.

Lemma 5.5.4. *For $n \geq 1$, and $\lambda > 0$,*

$$\int_0^\infty x^{n-1}e^{-\lambda x} = \frac{(n-1)!}{\lambda^n} \tag{5.5.2}$$

Proof. For all $n \geq 1, \lambda > 0, a > 0$ define $u : [0, a] \to \mathbb{R}$ and $v : [0, a] \to \mathbb{R}$ by

$$u(x) = x^{n-1} \text{ and } v(x) = e^{-\lambda x}.$$

As u, v are continuous functions, clearly $I^a_{n,\lambda}$ given by

$$I^a_{n,\lambda} = \int_0^a x^{n-1}e^{-\lambda x}.$$

is well defined finite positive number. As $x^\alpha e^{-\beta x} \to 0$ as $x \to \infty$ for any $\alpha, \beta > 0$ there is a $K > 0$ such that

$$0 \leq x^{n-1}e^{-\lambda x} < e^{-\frac{\lambda x}{2}},$$

for all $K > 0$. Therefore $b > a > k$ we have

$$| I^a_{n,\lambda} - I^b_{n,\lambda} | = \int_a^b x^{n-1}e^{-\lambda x} \leq \int_a^b e^{-\frac{\lambda x}{2}}dx = 2(e^{-\frac{\lambda b}{2}} - e^{-\frac{\lambda a}{2}}).$$

From this it is standard to note that

$$I_{n,\lambda} := \int_0^\infty x^{n-1}e^{-\lambda x} = \lim_{a\to\infty} I^a_{n,\lambda}$$

is a well defined finite positive number. Now, as u, v are differentiable we have by the integration by parts formula

$$\int_0^a u(x)v'(x)dx = u(a)v(a) - u(0)v(0) - \int_0^a u'(x)v(x)dx.$$

Substituting for u, v above we get

$$-\lambda I^a_{n,\lambda} = a^{n-1}e^{-\lambda a} - (n-1)I^a_{n-1,\lambda}.$$

Taking limits as $a \to \infty$ we have

$$\lambda I_{n,\lambda} = (n-1)I_{n-1,\lambda}.$$

Applying the above inductively we have

$$I_{n,\lambda} = \prod_{i=1}^{n-1} \frac{(n-i)}{\lambda} I_{1,\lambda} = \frac{(n-1)!}{\lambda^{n-1}} I_{1,\lambda}.$$

Using the fact that $I_1 = \frac{1}{\lambda}$ we have the result. ■

Definition 5.5.5. *$X \sim$ **Gamma**(n, λ): Let $\lambda > 0$ and $n \in \mathbb{N}$. Then X is said to be Gamma distributed with parameters n and λ if it has the density*

$$f(x) = \frac{\lambda^n}{(n-1)!} x^{n-1} e^{-\lambda x}, \tag{5.5.3}$$

where $x \geq 0$. The parameter n is referred to as the shape parameter and λ as the rate parameter. By (5.5.2) we know that f given by (5.5.3) is a density function.

We saw in Example 5.5.3 that sum of two exponential distributions resulted in a gamma distribution. If $X \sim$ Exponential (λ) then it can also be viewed as a Gamma$(1, \lambda)$ distribution. The result in Example 5.5.3 could be rephrased as follows: the sum of two gamma random variables with shape parameter 1 and rate parameter λ is distributed as a gamma random variable with shape parameter 2 and rate parameter λ. This holds more generally as we show in the next example.

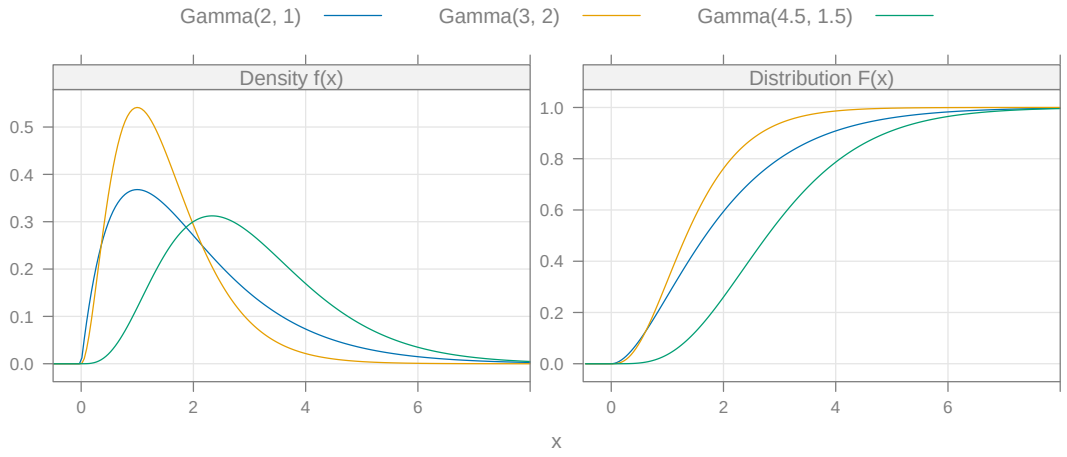


Figure 5.15: The Gamma density and cumulative distribution functions for various shape and rate parameters.

Example 5.5.6. (Sum of Gammas) Let $n \in \mathbb{N}, m \in \mathbb{N}, \lambda > 0$, X and Y be two independent Gamma(n, λ) and Gamma(m, λ) random variables respectively. Let $Z = X + Y$. Then we know that Z has a density given by (5.5.1). Further,

$$\begin{aligned} f_X(x) f_Y(z-x) &= \begin{cases} \frac{\lambda^n}{(n-1)!} x^{n-1} e^{-\lambda x} \frac{\lambda^m}{(m-1)!} (z-x)^{m-1} e^{-\lambda(z-x)} & \text{if } x \geq 0, z-x \geq 0 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{e^{-\lambda z} \lambda^{n+m}}{(n-1)!(m-1)!} x^{n-1} (z-x)^{m-1} & \text{if } x \geq 0, x \leq z, z \geq 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

For $z \geq 0$, we have

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_{X_1}(x) f_{X_2}(z-x) dx = \int_0^z f_{X_1}(x) f_{X_2}(z-x) dx \\ &= \frac{e^{-\lambda z} \lambda^{n+m}}{(n-1)!(m-1)!} \int_0^z x^{n-1} (z-x)^{m-1} dx \end{aligned}$$

We now make a change of variable $x = zu$ so that $dx = zdu$ to obtain

$$f_Z(z) = \frac{z^{n+m-1} e^{-\lambda z} \lambda^{n+m}}{(n-1)!(m-1)!} \int_0^1 u^{n-1} (1-u)^{m-1} du$$

Define

$$c(n, m) = \frac{\int_0^1 u^{n-1} (1-u)^{m-1} du}{(n-1)!(m-1)!}.$$

Thus we have the probability density of Z is given by,

$$f_Z(z) = \begin{cases} c(n,m) \cdot \lambda^{n+m} z^{n+m-1} e^{-\lambda z} & \text{if } z \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

To evaluate $c(n,m)$ we use the following fact. From Proposition 5.5.1 $f_Z(\cdot)$ (given by (5.5.1)) is a Probability density function. Therefore,

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f_Z(z)dz \\ &= c(n,m)\lambda^{n+m} \int_0^{\infty} z^{n+m-1} e^{-\lambda z} dz \\ &= c(n,m)[(n+m-1)!], \end{aligned}$$

where in the last line we have used (5.5.2) with n replaced by $n+m$. So $c(n,m) = \frac{1}{(n+m-1)!}$. Hence Z has Gamma $(n+m, \lambda)$ distribution. From the definition of $c(n,m)$ we also have

$$\int_0^1 u^{n-1}(1-u)^{m-1} du = \frac{(n+m-1)!}{(n-1)!(m-1)!}.$$

The above calculation is easily extended by an induction argument to obtain the fact that if $\lambda > 0$, X_i, $1 \leq i \leq n$ are independent Gamma(n_i, λ) distributed random variables (respectively). Then $Z = \sum_{i=1}^{n} X_i$ has Gamma $(\sum_{i=1}^{n} n_i, \lambda)$ distribution.

As Exponential (λ) is the same as Gamma$(1, \lambda)$ random variable, the above implies that the sum of n independent Exponential (λ) random variables is a Gamma(n, λ) random variable. ■

It is possible to define the Gamma distribution when the shape parameter is not necessarily an integer.

Definition 5.5.7. *$X \sim$ **Gamma**(α, λ): Let $\lambda > 0$ and $\alpha > 0$. Then X is said to be Gamma distributed with shape parameter α and rate parameter λ if it has the density*

$$f(x) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \tag{5.5.4}$$

where $x \geq 0$ and for $\alpha > 0$

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx \tag{5.5.5}$$

One can imitate the calculation done in Example 5.5.6 as well for such a Gamma distribution.

The distribution function of a gamma random variable involves an indefinite form of the integral in (5.5.5). Such integrals are known as incomplete gamma functions, and have no closed-form solution in terms of simple functions. In R, $F(x)$ for the gamma distribution

$$F(x) = P(X \leq x) = \int_0^x \frac{\lambda^\alpha}{\Gamma(\alpha)} z^{\alpha-1} e^{-\lambda z} dz \, , x > 0$$

can be evaluated numerically with a function call of the form `pgamma(x, alpha, lambda)`. For example,

```
pgamma(1, 2, 1)
```

```
[1] 0.2642411
```

```
pgamma(3, 4.5, 1.5)
```

```
[1] 0.5627258
```

Similarly, the density function $f(x)$ in (5.5.4) involves the normalising constant $\Gamma(\alpha)$ (also known as the gamma function) which usually cannot be computed explicitly when α is not an integer. Using R, one can evaluate $f(x)$ numerically using the `dgamma()` function as

```
dgamma(1, 2, 1)
```

```
[1] 0.3678794
```

```
dgamma(3, 4.5, 1.5)
```

```
[1] 0.2769272
```

5.5.2 *Distributions of Quotients of Independent Random Variables*

Let X and Y be two independent continous random variables with densities f_X and f_Y. In this section we shall find out the probability density function of $Z = \frac{X}{Y}$. As $P(Y = 0) = 0$, Z is well defined random variable.

Proposition 5.5.8. *(Quotient of two independent random variables) Let X and Y be two independent random variables with marginal densities given by $f_X : \mathbb{R} \to \mathbb{R}$ and $f_Y : \mathbb{R} \to \mathbb{R}$. Then $Z = \frac{X}{Y}$ has a probability density function $f_Z : \mathbb{R} \to \mathbb{R}$ given by*

$$f_Z(z) = \int_{-\infty}^{\infty} \mid y \mid f_X(zy) f_Y(y) dy. \tag{5.5.6}$$

Proof- Let us find an expression for the distribution function of Z.

$$\begin{aligned}
F(z) &= P(Z \leq z) \\
&= P(\frac{X}{Y} \leq z) \\
&= \int \int_{\{(x,y): y \neq 0, \frac{x}{y} \leq z\}} f_X(x) f_Y(y) dy dx \\
&= \int \int_{\{(x,y): y<0, \frac{x}{y} \leq z\}} f_X(x) f_Y(y) dy dx + \int \int_{\{(x,y): y>0, \frac{x}{y} \leq z\}} f_X(x) f_Y(y) dy dx \\
&= \int \int_{\{(x,y): y<0, x \geq yz\}} f_X(x) f_Y(y) dy dx + \int \int_{\{(x,y): y>0, x \leq yz\}} f_X(x) f_Y(y) dy dx \\
&= \int_{-\infty}^{0} \int_{yz}^{\infty} f_X(x) f_Y(y) dx dy + \int_{0}^{\infty} \int_{-\infty}^{yz} f_X(x) f_Y(y) dx dy \\
&= I + II
\end{aligned}$$

Let us make a u-substituion $x = yu$ in both I and II. For I, $y < 0$, so we will obtain,

$$\begin{aligned}
I &= \int_{-\infty}^{0} \int_{z}^{-\infty} y f_X(yu) f_Y(y) du dy \\
&= \int_{-\infty}^{0} \int_{-\infty}^{z} (-y) f_X(yu) f_Y(y) du dy \\
&= \int_{-\infty}^{z} \int_{-\infty}^{0} (-y) f_X(yu) f_Y(y) dy du,
\end{aligned}$$

where in the last line we have changed the order of integration[1]. For II, $y > 0$ so we will obtain (similarly as in I),

$$\begin{aligned}
II &= \int_{0}^{\infty} \int_{-\infty}^{z} y f_X(yu) f_Y(y) du dy \\
&= \int_{-\infty}^{z} \int_{0}^{\infty} y f_X(yu) f_Y(y) dy du,
\end{aligned}$$

[1] The change of order of integration is justifiable under certain hypothesis for the integrand. We shall assume these are satisfied, as it is not possible to state and verify them within the scope of this book

Therefore

$$\begin{aligned} F(z) &= I + II \\ &= \int_{-\infty}^{z}\int_{-\infty}^{0}(-y)f_X(yu)f_Y(y)dydu + \int_{-\infty}^{z}\int_{0}^{\infty} yf_X(yu)f_Y(y)dydu \\ &= \int_{-\infty}^{z}\int_{-\infty}^{\infty} \mid y \mid f_X(yu)f_Y(y)dydu \end{aligned}$$

As $f_X(\cdot)$ and $f_Y(\cdot)$ are densities, it can be shown that the integrand is a piecewise continuous function. Hence the F is of the form (5.2.4) and Theorem 5.2.5 implies that the probability density function of Z is given by (5.5.6). ■

Using the above method for finding the distribution of quotient of two random variables, we shall present three examples that will lead us to standard continuous distributions that are useful in applications. We begin with an example that constructs the Cauchy distribution.

Example 5.5.9. Let X and Y be two independent Normal random variables with mean 0 and variance $\sigma^2 \neq 0$. Let $Z = \frac{X}{Y}$. We know that the probability density function of Z is given by (5.5.6). Further, for any $y, z \in \mathbb{R}$

$$f_X(zy)f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2y^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{y^2}{2\sigma^2}} = \frac{1}{2\pi\sigma^2}\exp\left(-\left(\frac{1+z^2}{2\sigma^2}\right)y^2\right)$$

Fix $z \in \mathbb{R}$.

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} \mid y \mid \frac{1}{2\pi\sigma^2}\exp\left(-\left(\frac{1+z^2}{2\sigma^2}\right)y^2\right)dy \\ &= \frac{1}{2\pi\sigma^2}\left[\int_{-\infty}^{0} \mid y \mid \exp\left(-\left(\frac{1+z^2}{2\sigma^2}\right)y^2\right)dy + \int_{0}^{\infty} \mid y \mid \exp\left(-\left(\frac{1+z^2}{2\sigma^2}\right)y^2\right)dy\right] \\ &= \frac{1}{2\pi\sigma^2}\left[\int_{-\infty}^{0}(-y)\exp\left(-\left(\frac{1+z^2}{2\sigma^2}\right)y^2\right)dy + \int_{0}^{\infty} y\exp\left(-\left(\frac{1+z^2}{2\sigma^2}\right)y^2\right)dy\right] \end{aligned}$$

It is easy to see that two integrals are the same (perform a substitution of $u = -y$ in the first integral). So the above is

$$= \frac{1}{\pi\sigma^2}\int_{0}^{\infty} y\exp\left(-\left(\frac{1+z^2}{2\sigma^2}\right)y^2\right)dy.$$

Now perform a substitution $\left(\frac{1+z^2}{2\sigma^2}\right) y^2 = t$, so $\frac{1+z^2}{\sigma^2} y dy = dt$.

$$\begin{aligned} f_Z(z) &= \frac{1}{\pi\sigma^2} \frac{\sigma^2}{1+z^2} \int_0^\infty \exp(-t) dt. \\ &= \frac{1}{\pi(1+z^2)} (-e^{-t} \mid_0^\infty) = \frac{1}{\pi(1+z^2)}. \end{aligned}$$

Therefore Z has the Cauchy distribution, which we first saw in the context of Example 5.3.4. ■

The next example considers the ratio of two gamma random variables. This motivates a standard distribution called the F-distribution, which we will encounter in Chapter 8.

Example 5.5.10. Let $m \in \mathbb{N}, n \in \mathbb{N}, \lambda > 0$, X and Y be two independent Gamma (m, λ) and Gamma (n, λ) random variables respectively. Let $Z = \frac{X}{Y}$. We know that the probability density function of Z is given by (5.5.6). Further,

$$\begin{aligned} f_X(zy) f_Y(y) &= \begin{cases} \frac{\lambda^m}{(m-1)!} (zy)^{m-1} e^{-\lambda(zy)} \frac{\lambda^n}{(n-1)!} y^{n-1} e^{-\lambda y} & \text{if } y \geq 0, zy \geq 0 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{\lambda^{n+m}}{(n-1)!(m-1)!} y^{n+m-2} z^{m-1} e^{-\lambda(1+z)y} & \text{if } y \geq 0, z \geq 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Fix $z > 0$,

$$\begin{aligned} f_Z(z) &= \int_0^\infty y \frac{\lambda^{n+m}}{(n-1)!(m-1)!} y^{n+m-2} z^{m-1} e^{-\lambda(1+z)y} dy \\ &= \frac{z^{m-1}\lambda^{n+m}}{(n-1)!(m-1)!} \int_0^\infty y^{n+m-1} e^{-\lambda(1+z)y} dy \end{aligned}$$

Now perform a substition $(1+z)y = t$, so $(1+z)dy = dt$ and the above is

$$= \frac{z^{m-1}}{(1+z)^{m+n}} \frac{\lambda^{m+n}}{(m-1)!(n-1)!} \int_0^\infty t^{m+n-1} e^{-\lambda t} dt$$

Using (5.5.2) we have that

$$f_Z(z) = \begin{cases} \frac{(m+n-1)!}{(m-1)!(n-1)!} z^{m-1} (1+z)^{-(m+n)} & \text{if } z \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.5.7}$$

■

Our next example is a construction of the Beta-distribution.

Example 5.5.11. Let $m \in \mathbb{N}, n \in \mathbb{N}, \lambda > 0$. Let X and Y be two independent Gamma (m, λ) and Gamma (n, λ) random variables respectively. Let $Z = \frac{X}{X+Y}$.

Let $W = \frac{Y}{X}$. Note that $Z = \frac{1}{1+W}$. In Example 5.5.10 we found the probability density function of W. We shall use this to find the distribution funciton of Z. As $P(W \geq 0) = 1$,

$$P(Z \leq z) = \begin{cases} 0 & \text{if } z < 0 \\ 1 & \text{if } z > 1. \end{cases}$$

For $0 < z < 1$,

$$\begin{aligned} P(Z \leq z) &= P(\frac{1}{1+W} \leq z) = P(W \geq \frac{1-z}{z}) \\ &= 1 - P(W \leq \frac{1-z}{z}) \end{aligned}$$

Using (5.5.7) we obtain that the above is

$$\begin{aligned} &= 1 - \int_0^{\frac{1-z}{z}} \frac{(m+n-1)!}{(m-1)!(n-1)!} u^{m-1}(1+u)^{-(m+n)} du \\ &= 1 - \frac{(m+n-1)!}{(m-1)!(n-1)!} \int_0^{\frac{1-z}{z}} u^{m-1}(1+u)^{-(m+n)} du \end{aligned}$$

For $0 < z < 1$, by the fundamental theorem of calculus, differentiating in z

$$\begin{aligned} f_Z(z) &= \frac{1}{z^2} \cdot \frac{(m+n-1)!}{(m-1)!(n-1)!} \left(\frac{1-z}{z}\right)^{m-1} \left(1 + \frac{1-z}{z}\right)^{-(m+n)} \\ &= \frac{(m+n-1)!}{(m-1)!(n-1)!} z^{n-1}(1-z)^{m-1} \end{aligned}$$

Z is said to have the Beta(m, n) distribution. ■

We define the distribution in general next.

Definition 5.5.12. $X \sim$ **Beta**(α, β)*: Let $\alpha > 0$ and $\beta > 0$. Then X is said to be Beta distributed with parameters α and β if it has the density*

$$f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & 0 < x < 1 \\ 0 & \textit{otherwise.} \end{cases} \tag{5.5.8}$$

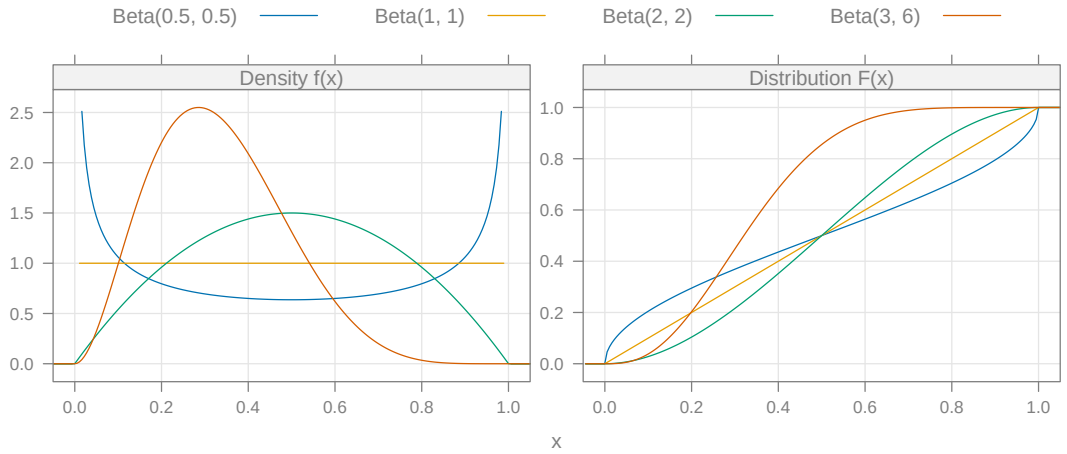


Figure 5.16: The Beta density and cumulative distribution functions for selected shape parameters.

The distribution function of a beta random variable is given by an indefinite integral which in general has no closed-form solution in terms of simple functions. In R, $F(x)$ for the beta distribution

$$F(x) = P(X \leq x) = \int_0^x \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-1}(1-u)^{\beta-1} du\,, 0 < x < 1$$

can be evaluated numerically with a function call of the form `pbeta(x, alpha, beta)`. For example,

```
pbeta(0.5, 0.5, 0.5)
```

```
[1] 0.5
```

```
pbeta(0.5, 3, 6)
```

```
[1] 0.8554688
```

```
pbeta(0.2, 6, 1)
```

```
[1] 6.4e-05
```

```
pbeta(0.2, 1, 6)
```

```
[1] 0.737856
```

In the special case where either α or β equals 1, the distribution function of X can be computed explicitly. Another special case is the standard arcsine law we previously encountered in Exercise 5.2.7 in terms of its explicit distribution function; it is easy to see that this is the same as the Beta$(\frac{1}{2}, \frac{1}{2})$ distribution. The semicircular distribution encountered in Exercise 5.2.6 is also related, in the sense that it can be viewed as a location and scale transformed beta random variable.

EXERCISES

Ex. 5.5.1. Suppose that X and Y are random variables with joint probability density

$$f(x,y) = \begin{cases} \frac{4}{5}(xy+1) & \text{if } 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Compute the marginal densities of X and Y ?

(b) Compute the conditional density $(X|Y = y)$ (for appropriate y).

(c) Are X and Y independent?

Ex. 5.5.2. Let X and Y be two random variables with the joint p.d.f given by

$$f(x,y) = \begin{cases} \lambda^2 e^{-\lambda y} & 0 \leq x \leq y \\ \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the marginal distribution of X and Y.

(b) Find the conditional distribution of $(Y \mid X = x)$ for some $x > 0$

Ex. 5.5.3. Let $a, b > 0$. Let $X \sim$ Gamma (a, b) and $Y \sim$ Exponential (X).

(a) Find the joint density of X and Y.

(b) Find the marginal density of Y.

(c) Find the conditional density of $(X \mid Y = y)$.

Ex. 5.5.4. Let X_1, X_2, X_3 be independent and identically distributed Uniform $(0,1)$ random variables. Let $A = X_1 X_3$ and $B = X_2^2$. Find the $P(A < B)$.

Ex. 5.5.5. Let X and Y be two independent exponential random variables each with mean 1.

(a) Find the density of $U_1 = X^{\frac{1}{2}}$.

(b) Find the density of $U_2 = X + Y + 1$.

(c) Find $P(\max\{X, Y\} > 1)$.

Ex. 5.5.6. Suppose X is a uniform random variable in the interval $(0,1)$ and Y is an independent exponential(2) random variable. Find the distribution of $Z = X + Y$.

Ex. 5.5.7. Let $\alpha > 0, \beta > 0, \lambda > 0$, X and Y be two independent Gamma(α, λ) and Gamma(β, λ) random variables respectively. Then $Z = X + Y$ is distributed as a Gamma $(\alpha + \beta, \lambda)$.

Ex. 5.5.8. Let X and Y be two independent random variables with probability density function $f_X(\cdot)$ and $f_Y(\cdot)$. Show that $X + Y$ and $Y + X$ have the same distribution by showing that the integral expression defining $f_X \star f_Y(\cdot)$ is equal to the integral expression defining $f_Y \star f_X(\cdot)$).

Ex. 5.5.9. Let $\alpha > 0$ and $\Gamma(\alpha)$ as in (5.5.5).

(a) Using the same technique as in Lemma 5.5.4, show that $0 < \Gamma(\alpha) < \infty$.

(b) Show that $\Gamma(\frac{1}{2}) = \int_0^\infty x^{-0.5} e^{-x} dx = \sqrt{\pi}$.

Ex. 5.5.10. Let $\alpha > 0, \delta > 0, \lambda > 0$. Let X and Y be two independent Gamma (α, λ) and Gamma (δ, λ) random variables respectively.

(a) Let $W = \frac{Y}{X}$. Find the probability density function of W.

(b) Let $Z = \frac{X}{X+Y}$. Find the probability density function of Z.

(c) Are X and Z independent ?
Hint: Compute the joint density and see if it is a product of the marginals.

Ex. 5.5.11. Suppose X, Y are independent random variables each normally distributed with mean 0 and variance 1.

(a) Find the probability density function of $R = \sqrt{X^2 + Y^2}$

(b) Find the probability density function of $Z = \frac{X}{Y}$

(c) Find the probability density function of $\theta = \arctan\left(\frac{X}{Y}\right)$

(d) Are R, θ independent random variables ?
Hint: Compute the joint density using the change of variable indicated in Exercise 5.1.10. Decide if it is a product of the marginals

6

SUMMARISING CONTINUOUS RANDOM VARIABLES

In this chapter we shall revisit concepts that have been discussed for discrete random variables and see their analogues in the continuous setting. We then introduce generating functions and conclude this chapter with a discussion on bivariate normal random variables.

6.1 EXPECTATION AND VARIANCE

The notion of expected value carries over from discrete to continuous random variables, but instead of being described in terms of sums, it is defined in terms of integrals.

Definition 6.1.1. *Let X be a continuous random variable with piecewise continuous density $f(x)$. Then the expected value of X is given by*

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx.$$

provided that the integral converges absolutely.[a] *In this case we say that X has "finite expectation". If the integral diverges to $\pm\infty$ we say the random variable has infinite expectation. If the integral diverges, but not to $\pm\infty$ we say the expected value is undefined.*

[a] That is, $\lim\limits_{\substack{M\to-\infty\\N\to\infty}} \int_M^N |x|\, f(x)\, \mathrm{d}x < \infty$.

The next three examples illustrate the three posibilities: the first is an example where expectation exists as a real number; the next is an example of an infinite expected value; and the final example shows that the expected value may not be defined at all.

Example 6.1.2. Let $X \sim \text{Uniform}(a, b)$. Then the expected value of X is given by

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx = \int_a^b x \cdot \frac{1}{b-a}\, dx = \frac{1}{2(b-a)}(b^2 - a^2) = \frac{b+a}{2}.$$

This result is intuitive since it says that the average value of a Uniform(a, b) random variable is the midpoint of its interval. ■

Example 6.1.3. Let $0 < \alpha < 1$ and $X \sim \text{Pareto}(\alpha)$ which is defined to have the probability density function

$$f(x) = \begin{cases} \frac{\alpha}{x^{\alpha+1}} & 1 \leq x < \infty \\ \\ 0 & \text{otherwise} \end{cases}$$

$$E[X] = \int_1^\infty x \cdot \frac{\alpha}{x^{\alpha+1}}\, dx = \alpha \lim_{M\to\infty} \int_1^M x^{-\alpha} dx = \frac{\alpha}{-\alpha+1}(-1 + \lim_{M\to\infty} M^{-\alpha+1}) = \infty$$

as $0 < \alpha < 1$.

Thus this Pareto random variable has an infinite expected value. ■

Example 6.1.4. Let $X \sim \text{Cauchy}(0, 1)$. Then the probability density function of X is given by

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2} \text{ for all } x \in \mathbb{R}.$$

Now,

$$E[X] = \int_{-\infty}^\infty x \cdot \frac{1}{\pi(1+x^2)}\, dx$$

Now by Exercise 6.1.10, we know that as $M \to -\infty, N \to \infty$ the $\int_M^N \frac{x}{1+x^2} dx$ does not converge or diverge to $\pm\infty$. So $E[X]$ is not defined for this Cauchy random variable. ■

Expected values of functions of continuous random variables may be computed using their respective probability density function by the following theorem.

Theorem 6.1.5. *Let X be continuous random variables with probability density function $f_X : \mathbb{R} \to \mathbb{R}$.*

(a) Let $g : \mathbb{R} \to \mathbb{R}$ be piecewise continuous and $Z = g(X)$ Then the expected value of Z given by

$$E[g(X)] \quad = \quad \int_{-\infty}^\infty g(x) f_X(x)\, dx$$

(b) Let Y be a continuous random variable such that (X, Y) have a joint probability density function $f : \mathbb{R}^2 \to \mathbb{R}$. Suppose $h : \mathbb{R}^2 \to \mathbb{R}$ be piecewise continuous. Then,

$$E[h(X, Y)] = \int_{-\infty}^\infty \int_{-\infty}^\infty h(x, y) f(x, y)\, dx\, dy.$$

Proof- The proof is beyond the scope of this book. For (a) when g is as in Exercise 5.3.10 then one can provide the proof using only the tools of basic calculus (we will leave this case as an exercise to the reader) ■

We illustrate the use of the above theorem with a couple of examples.

Example 6.1.6. A piece of equipment breaks down after a functional lifetime that is a random variable $T \sim Exp(\frac{1}{5})$. An insurance policy purchased on the equipment pays a dollar amount equal to $1000 - 200t$ if the equipment breaks down at a time $0 \leq t \leq 5$ and pays nothing if the equipment breaks down after time $t = 5$. What is the expected payment of the insurance policy?

For $t \geq 0$ the policy pays $g(t) = \max\{1000 - 200t, 0\}$ so,

$$\begin{aligned} E[g(T)] &= \int_0^\infty \frac{1}{5} e^{(1/5)t} \max\{1000 - 200t, 0\}\, dt \\ &= \int_0^5 \frac{1}{5} e^{(1/5)t} (1000 - 200t)\, dt \\ &= 1000e^{-1} \approx \$367.88 \end{aligned}$$

■

Example 6.1.7. Let $X, Y \sim \text{Uniform}(0, 1)$. What is the expected value of the larger of the two variables?

We offer two methods of solving this problem. The first is to define $Z = \max\{X, Y\}$ and then determine the density of Z. To do so, we first find its distribution. $F_Z(z) = P(Z \leq z)$, but $\max\{X, Y\}$ is less than or equal to z exactly when both X and Y are less than or equal to z. So for $0 \leq z \leq 1$,

$$\begin{aligned} F_Z(z) &= P((X \leq z) \cap (Y \leq z)) \\ &= P(X \leq z) \cdot P(Y \leq z) \\ &= z^2 \end{aligned}$$

Therefore $f_Z(z) = F_Z'(z) = 2z$ after which the expected value can be obtained through integration

$$E[Z] = \int_0^1 z \cdot 2z\, dz = \frac{2}{3} z^3 \,|_0^1 = \frac{2}{3}.$$

An alternative method is to use Theorem 6.1.5 (b) to calculate the expectation directly without finding a new density. Since X and Y are independent, their joint distribution is the product of their marginal distributions. That is,

$$f(x, y) = f_X(x) f_Y(y) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Therefore,

$$\begin{aligned} E[\max\{X,Y\}] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \max\{x,y\}\cdot f(x,y)\,dx\,dy \\ &= \int_0^1\int_0^1 \max\{x,y\}\cdot 1\,dx\,dy \end{aligned}$$

The value of $\max\{x,y\}$ is x if $0 < y \leq x < 1$ and it is y if $0 < x \leq y < 1$. So,

$$\begin{aligned} E[\max\{X,Y\}] &= \int_0^1\int_0^y y\,dx\,dy + \int_0^1\int_y^1 x\,dx\,dy \\ &= \int_0^1 xy\,|_{x=0}^{x=y}\,dy + \int_0^1 \frac{1}{2}x^2\,|_{x=y}^{x=1}\,dy \\ &= \int_0^1 y^2\,dy + \int_0^1 \frac{1}{2} - \frac{1}{2}y^2\,dy \\ &= \frac{1}{3} + \frac{1}{3} = \frac{2}{3}. \end{aligned}$$

■

Results from calculus may be used to show that the linearity properties from Theorem 4.1.7 such as apply to continuous random variables as well as to discrete ones. We restate it here for completeness.

Theorem 6.1.8. *Suppose that X and Y are continuous random variables with piecewise continuous joint density function function $f : \mathbb{R}^2 \to \mathbb{R}$. Assume that both have finite expected value. If a and b are real numbers then*

(a) $E[aX] = aE[X]$;

(b) $E[aX + b] = aE[X] + b$

(c) $E[X + Y] = E[X] + E[Y]$; *and*

(d) $E[aX + bY] = aE[X] + bE[Y]$.

(e) If $X \geq 0$ *then* $E[X] \geq 0$.

Proof- See Exercise 6.1.11. ■

We will use these now-familiar properties in the continuous setting. As in the discrete setting we can define the variance and standard deviation of a continuous random variable.

Definition 6.1.9. *Let X be a random variable with probability density function $f : \mathbb{R} \to \mathbb{R}$. Suppose X has finite expectation. Then*

(a) the variance of the random variable is written as $Var[X]$ and is defined as

$$Var[X] = E[(X - E[X])^2] = \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x) dx,$$

(b) the standard deviation of X is written as $SD[X]$ and is defined as

$$SD[X] = \sqrt{Var[X]}$$

Since the above terms are expected values, there is the possibility that they may be infinite because the integral describing the expectation diverges to infinity. As the integrand is strictly positive, it isn't possible for the integral to diverge unless it diverges to infinity.

The properties of variance and standard deviation of continuous random variables match those of their discrete counterparts. A list of these properties follows below.

Theorem 6.1.10. *Let $a \in \mathbb{R}$ and let X be a continuous random variable with finite variance (and thus, with finite expected value as well). Then,*

(a) $Var[X] = E[X^2] - (E[X])^2$.

(b) $Var[aX] = a^2 \cdot Var[X]$;

(c) $SD[aX] = |a| \cdot SD[X]$;

(d) $Var[X + a] = Var[X]$; *and*

(e) $SD[X + a] = SD[X]$.

If Y is another independent continuous random variable with finite variance (and thus, with finite expected value as well) then

(f) $E[XY] = E[X]E[Y]$;

(g) $Var[X + Y] = Var[X] + Var[Y]$; *and*

(h) $SD[X + Y] = \sqrt{(SD[X])^2 + (SD[Y])^2}$.

Proof- The proof is essentially an imitation of the proofs presented in Theorem 4.1.10, Theorem 4.2.5, Theorem 4.2.4, and Theorem 4.2.6. One needs to use the respective densities, integrals in lieu of sums, and use Theorem 6.1.11 and Theorem 6.1.5 when needed. We will leave this as an exercise to the reader. ■

Example 6.1.11. Let $X \sim$ Normal $(0, 1)$. In this example we shall show that $E[X] = 0$ and $Var[X] = 1$. Before that we collect some facts about the probability density function of X, given by (5.2.7). Using (5.2.9) with $z = 0$, we can conclude that

$$\int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \frac{1}{2} \tag{6.1.1}$$

Observe that there exists $c_1 > 0$ such that

$$\max\{|\ x\ |, x^2\} e^{-\frac{x^2}{2}} \leq c_1 e^{-c_1|x|}$$

for all $x \in \mathbb{R}$. Hence

$$\begin{aligned} \int_{-\infty}^\infty |\ x\ | \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx &\leq c_1 \int_{-\infty}^\infty e^{-c_1|x|} < \infty \\ \int_{-\infty}^\infty x^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx &\leq c_1 \int_{-\infty}^\infty e^{-c_1|x|} < \infty \end{aligned} \tag{6.1.2}$$

Using the above we see that

$$E[X] = \int_{-\infty}^\infty x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx < \infty$$

So we can split integral expression in definition of $E[X]$ as

$$E[X] = \int_{-\infty}^0 x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx + \int_0^\infty x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

Further the change of variable $y = -x$ will imply that

$$\int_{-\infty}^0 x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = -\int_0^\infty y \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy.$$

So $E[X] = 0$. Again by (6.1.2),

$$Var[X] = \int_{-\infty}^\infty (x - E[X])^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \int_{-\infty}^\infty x^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx < \infty$$

To evaluate the integral we make a change of variable to obtain

$$\int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = \int_{-\infty}^{0} x^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx + \int_{0}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 2\int_{0}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

Then we use integration by parts like Lemma 5.5.4. Set $u(x) = x$ and $v(x) = e^{-\frac{x^2}{2}}$, which imply $u'(x) = 1$ and $v'(x) = -xe^{-\frac{x^2}{2}}$. Therefore for $a > 0$,

$$\begin{aligned} \int_0^a x^2 e^{-\frac{x^2}{2}} dx &= \int_0^a u(x)(-v'(x))dx = u(x)(-v(x)) \,|_0^a - \int_0^a u'(x)(-v(x))dx \\ &= a^2 e^{-\frac{a^2}{2}} + \int_0^a e^{-\frac{x^2}{2}} \end{aligned}$$

Using the fact that $\lim_{a\to\infty} a^2 e^{-\frac{a^2}{2}} = 0$ and (6.1.1) we have

$$\begin{aligned} Var[X] &= 2\frac{1}{\sqrt{2\pi}} \int_0^\infty x^2 e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{\pi}} \lim_{a\to\infty} \int_0^a x^2 e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{\pi}} \lim_{a\to\infty} \left[a^2 e^{-\frac{a^2}{2}} + \int_0^a e^{-\frac{x^2}{2}} dx \right] \\ &= \frac{1}{\sqrt{\pi}} \left[0 + \int_0^\infty e^{-\frac{x^2}{2}} dx \right] = \frac{1}{\sqrt{\pi}} [0 + \sqrt{\pi}] = 1 \end{aligned}$$

Suppose $Y \sim$ Normal (μ, σ^2) then we know by Corollary 5.3.3 that $W = \frac{Y-\mu}{\sigma} \sim$ Normal $(0, 1)$. By Example 6.1.11, $E[W] = 0$ and $Var[W] = 1$. Also $Y = \sigma W + \mu$, so by Theorem 6.1.8(b) $E[Y] = \sigma E[W] + \mu = \mu$ and by Theorem 6.1.10 (d) and (b) $Var[Y] = \sigma^2 Var[W] = \sigma^2$. ■

Example 6.1.12. Let $X \sim$ Uniform(a, b). To calculate the variance of X first note that Theorem 6.1.5(a) gives

$$E[X^2] = \int_{-\infty}^{\infty} x^2 \cdot f(x)\, dx = \int_a^b x^2 \cdot \frac{1}{b-a}\, dx = \frac{1}{3(b-a)}(b^3 - a^3) = \frac{b^2 + ab + a^2}{3}.$$

Now, since $E[X] = \frac{b+a}{2}$ (see Example 6.1.2), the variance may be found as

$$Var[X] = E[X^2] - (E[X])^2 = \frac{b^2 + ab + a^2}{3} - (\frac{b+a}{2})^2 = \frac{(b-a)^2}{12}.$$

Taking square roots, we obtain $SD[X] = \frac{b-a}{\sqrt{12}}$. So the standard deviation of a continuous, uniform random variable is $\frac{1}{\sqrt{12}}$ times of the length of its interval. ■

The Markov and Chebychev inequalities also apply to continuous random variables. As with discrete variables, these help to estimate the probabilities that a random variable will fall within a certain number of standard deviations from its expected value.

Theorem 6.1.13. *Let X be a continuous random variable with probability density function f and finite non-zero variance.*

(a) **(Markov's Inequality)** *Suppose X is supported on non-negative values, i.e. $f(x) = 0$ for all $x < 0$. Then for any $c > 0$,*

$$P(X \geq c) \leq \frac{\mu}{c}.$$

(b) **(Chebychev's Inequality)** *For any $k > 0$,*

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Proof - (a) By definition of μ and assumptions on f, we have

$$\mu = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{\infty} xf(x)dx.$$

Using an elementary fact from integrals we know that

$$\int_{0}^{\infty} xf(x)dx = \int_{0}^{c} xf(x)dx + \int_{c}^{\infty} xf(x)dx$$

We note that the first integral is non-negative so we have

$$\mu \quad \geq \quad \int_{c}^{\infty} xf(x)dx.$$

As $f(\cdot) \geq 0$, we have $xf(x) \geq cf(x)$ whenever $x > c$. So again using facts about integrals

$$\mu \quad \geq \quad \int_{c}^{\infty} cf(x)dx = c\int_{c}^{\infty} f(x)dx = cP(X > c).$$

The last equality follows from definition. Hence we have the result.

(b) The event $(|X - \mu| \geq k\sigma)$ is the same as the event $((X - \mu)^2 \geq k^2\sigma^2)$. The random variable $(X - \mu)^2$ is certainly non-negative, is continuous by Exercise 5.3.9, and its expected value is the variance of X which we have assumed to be finite. Therefore we may apply Markov's inequality to $(X - \mu)^2$ to get

$$P(|X - \mu| \geq k\sigma) \quad = \quad P((X - \mu)^2 \geq k^2\sigma^2) \leq \frac{E[(X - \mu)^2]}{k^2\sigma^2} = \frac{Var[X]}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}.$$

Though the theorem is true for all $k > 0$, it doesn't give any useful information unless $k > 1$.

EXERCISES

Ex. 6.1.1. Suppose X has probability density function given by

$$f_X(x) = \begin{cases} 1 - |\, x \,| & -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Compute the distribution function of X.

(b) Compute $E[X]$ and $Var[X]$.

Ex. 6.1.2. Suppose X has probability density function given by

$$f_X(x) = \begin{cases} \frac{\cos(x)}{2} & -\frac{\pi}{2} \leq x \leq \frac{\pi}{2} \\ 0 & \text{otherwise} \end{cases}$$

(a) Compute the distribution function of X.

(b) Compute $E[X]$ and $Var[X]$.

Ex. 6.1.3. Find $E[X]$ and $Var[X]$ in the following situations:

(a) $X \sim \text{Normal}(\mu, \sigma^2)$, with $\mu \in \mathbb{R}$ and $\sigma > 0$.

(b) X has probability density function given by

$$f_X(x) = \begin{cases} x & 0 \leq x \leq 1 \\ 2 - x & 1 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Ex. 6.1.4. Let $1 < \alpha$ and $X \sim \text{Pareto}(\alpha)$. Calculate $E[X]$ to show that it is finite.

Ex. 6.1.5. Let X be a random variable with density $f(x) = 2x$ for $0 < x < 1$ (and $f(x) = 0$ otherwise).

(a) Calculate $E[X]$. You should get a result larger than $\frac{1}{2}$. Explain why this should be expected even without computations.

(b) Calculate $SD[X]$.

Ex. 6.1.6. Let $X \sim \text{Uniform}(a, b)$ and let $k > 0$. Let μ and σ be the expected value and standard deviation calculated in Example 6.1.12.

(a) Calculate $P(|X - \mu| \leq k\sigma)$. Your final answer should depend on k, but not on the values of a or b.

(b) What is the value of k such that results of more than k standard deviations from expected value are unachievable for X?

Ex. 6.1.7. Let $X \sim \text{Exponential}(\lambda)$.

(a) Prove that $E[X] = \frac{1}{\lambda}$ and $SD[X] = \frac{1}{\lambda}$.

(b) Let μ and σ denote the mean and standard deviation of X respectively. Use your computations from (a) to calculate $P(|X - \mu| \leq k\sigma)$. Your final answer should depend on k, but not on the value of λ.

(c) Is there a value of k such that results of more than k standard deviations from expected value are unachievable for X?

Ex. 6.1.8. Let $X \sim \text{Gamma}(n, \lambda)$ with $n \in \mathbb{N}$ and $\lambda > 0$. Using Example 5.5.3, Exercise 6.1.7(a) and Theorem 6.1.8(c) calculate $E[X]$. Using Theorem 6.1.10 calculate $Var[X]$.

Ex. 6.1.9. Let $X \sim \text{Uniform}(0, 10)$ and let $g(x) = \max\{x, 4\}$. Calculate $E[g(X)]$.

Ex. 6.1.10. Show that as $M \to -\infty, N \to \infty$ $\int_M^N \frac{x}{1+x^2} dx$ does not have a limit.

Ex. 6.1.11. Using the hints provided below prove the respective parts of Theorem 6.1.8.

(a) For $a = 0$ the result is clear. Let $a \neq 0$ and $f_X : \mathbb{R} \to \mathbb{R}$ be the probability density function of X. Use Lemma 5.3.2 to find the probability density function of aX. Compute the expectation of aX to obtain the result. Alternatively use Theorem 6.1.5(a).

(b) Use Theorem 6.1.5(b).

(c) Use the joint density of (X, Y) to write $E[X + Y]$. Then use (5.4.2) an (5.4.3) to prove the result.

(d) Use the same technique as in (b).

(e) If $X \geq 0$ then its marginal density $f_X : \mathbb{R} \to \mathbb{R}$ is positive only when the $x \geq 0$. The result immediately follows from definition of expectation.

Ex. 6.1.12. Prove Theorem 6.1.10.

6.2 COVARIANCE, CORRELATION, CONDITIONAL EXPECTATION AND CONDITIONAL VARIANCE

Covariance of continuous random variables (X, Y) is used to describe how the two random variables relate to each other. The properties proved about covariances for discrete random variables in Section 4.5 apply to continuous random variables as well via essentially the same arguments. We define covariance and state the properties next.

Definition 6.2.1. *Let X and Y be random variables with joint probability density function $f : \mathbb{R}^2 \to \mathbb{R}$. Suppose X and Y have finite expectation. Then the covariance of X and Y is defined as*

$$Cov[X,Y] = E[(X - E[X])(Y - E[Y])] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - E[X])(y - E[Y])f(x,y)dxdy, \tag{6.2.1}$$

Since it is defined in terms of an expected value, there is the possibility that the covariance may be infinite or not defined at all. We now state the properties of Covariance.

Theorem 6.2.2. *Let X, Y be continuous random variables such that they have joint probability density function. Assume that $0 \neq \sigma_x^2 = Var(X) < \infty, 0 \neq \sigma_y^2 = Var(Y) < \infty$. Then*

(a) $Cov[X,Y] = E[XY] - E[X]E[Y]$.

(b) $Cov[X,Y] = Cov[Y,X]$;

(c) $Cov[X,X] = Var[X]$.

(d) $-\sigma_X\sigma_Y \leq Cov[X,Y] \leq \sigma_X\sigma_Y$

(e) If X and Y are independent then $Cov[X,Y] = 0$.

Let a, b be real numbers. Suppose Z is another continuous random variable, and $\sigma_z = Var(Z) < \infty$. Further (X,Z), (Y,Z), $(X, aY + bZ)$, and $(aX + bY, Z)$ all have (their respective) joint probability functions. Then

(f) $Cov[X, aY + bZ] = a \cdot Cov[X,Y] + b \cdot Cov[X,Z]$;

(g) $Cov[aX + bY, Z] = a \cdot Cov[X,Z] + b \cdot Cov[Y,Z]$;

Proof. See Exercise 6.2.13. ■

Definition 6.2.3. *Let (X,Y) be continuous random variables both with finite variance and covariance. From Theorem 6.2.2(d) the quantity $\rho[X,Y] = \frac{Cov[X,Y]}{\sigma_X \sigma_Y}$ is in the interval $[-1,1]$. It is known as the "correlation" of X and Y. As discussed earlier, both the numerator and denominator include the units of X and the units of Y. The correlation, therefore, has no units associated with it. It is thus a dimensionless rescaling of the covariance and is frequently used as an absolute measure of trends between the two continuous random variables as well.*

EXAMPLE 6.2.4. Let $X \sim$ Uniform $(0,1)$ and be independent of $Y \sim$ Uniform $(0,1)$. Let $U = \min(X,Y)$ and $V = \max(X,Y)$. We wish to find $\rho[U,V]$. First, $0 < u < 1$

$$P(U \leq u) = 1 - P(U > u) = 1 - P(X > u, Y > u) = 1 - P(X > u)P(Y > u) = 1 - (1-u)^2,$$

as X, Y are independent uniform random variables. Second, for $0 < v < 1$,

$$P(V \leq v) = P(X \leq v, Y \leq v) = P(X \leq v)P(Y \leq v) = v^2,$$

as X, Y are independent uniform random variables. Therefore the distribution function of U and V are given by

$$F_U(u) = \begin{cases} 0 & \text{if } u < 0 \\ 1-(1-u)^2 & \text{if } 0 < u < 1 \\ 1 & \text{if } u \geq 1. \end{cases} \quad \text{and } F_V(v) = \begin{cases} 0 & \text{if } v < 0 \\ v^2 & \text{if } 0 < v < 1 \\ 1 & \text{if } v \geq 1. \end{cases}$$

As F_U, F_V are piecewise differentiable, the probability density function of U and V are obtained by differentiating F_U and F_V respectively.

$$f_U(u) = \begin{cases} 2(1-u) & \text{if } 0 < u < 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{and } f_V(v) = \begin{cases} 2v & \text{if } 0 < v < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Thirdly, $0 < u < v < 1$

$$\begin{aligned} P(U \leq u, V \leq v) &= P(V \leq v) - P(U > u, V \leq v) \\ &= v^2 - P(u < X \leq v, u < Y \leq v) \\ &= v^2 - P(u < X \leq v)P(u < Y \leq v) \\ &= v^2 - (v-u)^2, \end{aligned}$$

where we have used the formula for distribution function of V and the fact that X, Y are independent uniform random variables. It is easily seen that $P(U \leq u, V \leq v) = 0$ for all other possibilities of (u, v). As the joint distribution function is piecewise differentiable in each variable, the joint probability density function of U and V, $f : \mathbb{R}^2 \to \mathbb{R}$, exists and is obtained by differentiating it partially in u and v.

$$f(u,v) = \begin{cases} 2 & \text{if } 0 < u < v < 1 \\ 0 & \text{otherwise} \end{cases}$$

Now,

$$\begin{aligned}
E[U] &= \int_0^1 u2(1-u)du = u^2 - 2\frac{u^3}{3} \,|_0^1 = \frac{1}{3} \\
E[V] &= \int_0^1 v2vdv = 2\frac{v^3}{3} \,|_0^1 = \frac{2}{3} \\
E[U^2] &= \int_0^1 u^2 2(1-u)du = 2\frac{u^3}{3} - 2\frac{u^4}{4} \,|_0^1 = \frac{1}{6} \\
E[V^2] &= \int_0^1 v^2 2vdv = 2\frac{v^4}{4} \,|_0^1 = \frac{1}{2} \\
E[UV] &= \int_0^1 \left[\int_0^v uv2du\right] dv = \int_0^1 2v\left[\frac{u^2}{2} \,|_0^1\right] dv = \int_0^1 2v\frac{v^2}{2}dv = \frac{v^4}{4} \,|_0^1 = \frac{1}{4}
\end{aligned}$$

Therefore

$$\begin{aligned}
Var[U] &= E[U^2] - (E[U])^2 = \frac{2}{3} - \frac{1}{9} = \frac{5}{9} \\
Var[V] &= E[V^2] - (E[V])^2 = \frac{1}{2} - \frac{4}{9} = \frac{1}{18} \\
Cov[U,V] &= E[UV] - E[U]E[V] = \frac{1}{4} - \frac{1}{3}\frac{2}{3} = \frac{5}{36} \\
\rho[U,V] &= \frac{Cov[U,V]}{\sqrt{Var[V]}\sqrt{Var[U]}} = \frac{\frac{5}{36}}{\sqrt{\frac{5}{9}}\sqrt{\frac{1}{18}}} = \frac{1}{2\sqrt{2}}
\end{aligned}$$

■

As seen in Theorem 6.2.2 (e), independence of X and Y guarantees that they are uncorrelated (i.e $\rho[X, Y] = 0$). The converse is not true (See Example 4.5.6 for discrete case). It is possible that $Cov[X, Y] = 0$ and yet that X and Y are dependent, as the next example shows.

Example 6.2.5. Let $X \sim$ Uniform $(-1,1)$. Let $Y = X^2$. Note from Example 6.1.2 and Example 6.1.12 we have $E[X] = 0, E[Y] = E[X^2] = \frac{1}{3}$. Further using the probability density function of X,

$$E[XY] = E[X^3] = \int_{-1}^{1} x^3 \frac{1}{2} = \frac{x^4}{8} \,|_{-1}^{1} = 0.$$

So $\rho[X,Y] = 0$. Clearly X and Y are not independent. We verify this precisely as well. Consider the

$$P(X \leq -\frac{1}{4}, Y \leq \frac{1}{4}) = P(X \leq -\frac{1}{4}, X^2 \leq \frac{1}{4}) = P(-\frac{1}{2} \leq X \leq -\frac{1}{4}) = \frac{1}{8},$$

as $X \sim$ Uniform $(-1,1)$. Whereas,

$$P(X \leq -\frac{1}{4})P(Y \leq \frac{1}{4}) = P(X \leq -\frac{1}{4})P(X^2 \leq \frac{1}{4}) = P(X \leq -\frac{1}{4})P(-\frac{1}{2} \leq X \leq \frac{1}{2}) = \frac{3}{8}\frac{1}{2} = \frac{3}{16}.$$

Clearly

$$P(X \leq -\frac{1}{4}, Y \leq \frac{1}{4}) \neq P(X \leq -\frac{1}{4})P(Y \leq \frac{1}{4})$$

implying they are not independent. ■

We are now ready to define conditional expectation and variance.

Definition 6.2.6. *Let (X,Y) be continuous random variables with a piecewise continuous joint probability density function f. Let f_X be the marginal density of X. Assume x is a real number for which $f_x(x) \neq 0$. The conditional expectation of Y given $X = x$ is defined by*

$$E[Y \mid X = x] = \int_{-\infty}^{\infty} y f_{Y|X=x}(y) dy = \int_{-\infty}^{\infty} y \frac{f(x,y)}{f_X(x)} dy$$

whenever it exists. The conditional variance of Y given $X = x$ is defined by

$$\begin{aligned} Var[Y|X=x] &= E[(Y - E[Y|X=x])^2 | X = x] \\ &= \int_{-\infty}^{\infty} \left(y - \int_{-\infty}^{\infty} y \frac{f(x,y)}{f_X(x)} dy \right)^2 \frac{f(x,y)}{f_X(x)} dy. \end{aligned}$$

The results proved in Theorem 4.4.4, Theorem 4.4.6, Theorem 4.4.8, and Theorem 4.4.9 are all applicable when X and Y are continuous random variables having joint probability density function f. The proofs of these results in the continuous setting follow very similarly (though using facts about integrals from analysis).

Theorem 6.2.7. *Let (X,Y) be continuous random variables with joint probability density function $f:\mathbb{R}\to\mathbb{R}$. Assume that $h,g:\mathbb{R}\to\mathbb{R}$ be defined as*

$$g(y)=\begin{cases}E[X|Y=y] & \text{if } f_Y(y)>0\\ 0 & \text{otherwise}\end{cases} \quad \text{and } h(y)=\begin{cases}Var[X|Y=y] & \text{if } f_Y(y)>0\\ 0 & \text{otherwise}\end{cases}$$

are well-defined piecewise continuous functions. Let $k:\mathbb{R}\to\mathbb{R}$ be a piecewise continous function. Then

$$E[k(X)\mid Y=y]=\int_{-\infty}^{\infty}k(x)f_{X|Y=y}(x)dx, \tag{6.2.2}$$

$$E[g(Y)]=E[X], \tag{6.2.3}$$

and

$$Var[X]=E[h(Y)]+Var[g(Y)]. \tag{6.2.4}$$

Proof- The proof of (6.2.2) is beyond the scope of this book. We shall omit it. To prove (6.2.3) we use the definition of g and Theorem 6.1.8 (a) to write

$$E[g(Y)] \;=\; \int_{-\infty}^{\infty}g(y)f_Y(y)dy=\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}xf_{X|Y=y}(x)dx\right]f_Y(y)dy$$

Using the definition of conditional density and rearranging the order of integration we obtain that the above is

$$=\;\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}x\frac{f(x,y)}{f_Y(y)}dx\right]f_Y(y)dy=\int_{-\infty}^{\infty}x\left[\int_{-\infty}^{\infty}f(x,y)dy\right]dx=\int_{-\infty}^{\infty}xf_X(x)dx=E[X].$$

So we are done. To prove (6.2.4), using Exercise 6.2.8

$$h(y)=E[X^2\mid Y=y]-(E[X\mid Y=y])^2=E[X^2\mid Y=y]-(g(y))^2$$

From the above we have,

$$\begin{aligned}E[h(Y)] &= E[X^2]+E[g(Y)^2]\\ Var[g(Y)] &= E[g(Y)^2]-(E[g(Y)])^2=E[g(Y)^2]-(E[X])^2\end{aligned}$$

Therefore summing the two equations we have (6.2.4). ■

As before it is common to use $E[X|Y]$ to denote $g(Y)$ after which the result may be expressed as $E[E[X|Y]]=E[X]$. This can be slightly confusing notation, but one must

keep in mind that the exterior expected value in the expression $E[E[X|Y]]$ refers to the averge of $E[X|Y]$ viewed as a function of Y.

Similarly one denotes $h(Y)$ by $Var[X|Y]$. Then we can rewrite (6.2.4) as

$$Var[X] = E[Var[X|Y]] + Var[E[X|Y]].$$

Example 6.2.8. Let $X \sim$ Uniform $(0,1)$ and be independent of $Y \sim$ Uniform $(0,1)$. Let $U = \min(X,Y)$ and $V = \max(X,Y)$. In Example 6.2.4 we found $\rho[U,V]$. During that computation we showed that the marginal densities of U and V were given by

$$f_U(u) = \begin{cases} 2(1-u) & \text{if } 0 < u < 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{and } f_V(v) = \begin{cases} 2v & \text{if } 0 < v < 1 \\ 0 & \text{otherwise.} \end{cases}$$

and the joint density of (U,V) was given by

$$f(u,v) = \begin{cases} 2 & \text{if } 0 < u < v < 1 \\ 0 & \text{otherwise} \end{cases}$$

Let $0 < u < 1$. The conditional density of $V \mid U = u$, is given by

$$f_{V|U=u}(v) = \frac{f(u,v)}{f_U(u)}, \text{ for } v \in \mathbb{R}.$$

So,

$$f_{V|U=u}(v) = \begin{cases} \frac{1}{1-u} & \text{if } u < v < 1 \\ 0 & \text{otherwise} \end{cases}$$

Therefore $(V \mid U = u) \sim$ Uniform $(u,1)$. So the conditional expectation is given by

$$E[V \mid U = u] = \int_u^1 \frac{v}{1-u} dv = \frac{1-u^2}{2(1-u)} = \frac{1+u}{2}.$$

The conditional variance is given by

$$\begin{aligned} Var[V \mid U = u] &= E[V^2 \mid U = u] - (E[V \mid U = u])^2 \\ &= \int_u^1 \frac{v^2}{1-u} dv - \left(\frac{1+u}{2}\right)^2 \\ &= \frac{1-u^3}{3(1-u)} dv - \frac{(1+u)^2}{4} = \frac{(1-u)^2}{12}. \end{aligned}$$

We could have also concluded these from properties of Uniform distribution computed in Example 6.1.2 and Example 6.1.12. We will use this approach in the next example. ■

Example 6.2.9. Let (X, Y) have joint probability density function f given by

$$f(x,y) = \frac{\sqrt{3}}{4\pi} e^{-\frac{1}{2}(x^2 - xy + y^2)} \quad -\infty < x, y < \infty.$$

These random variables were considered in Example 5.4.12. We showed there that X is a Normal random variable with mean 0 and variance $\frac{4}{3}$ and Y is also a Normal random variable with mean 0 and variance $\frac{4}{3}$. We observed that they are not independent as well and the conditional distribution of Y given $X = x$ was Normal with mean $\frac{x}{2}$ and variance 1. Either by direct computation or by definition we observe that

$$E[Y \mid X = x] = \frac{x}{2} \quad Var[Y \mid X = x] = 1.$$

We could compute the $Var[Y]$ using (6.2.4), i.e

$$\begin{aligned} Var[Y] &= Var[E[Y \mid X]] + E[Var[Y \mid X = x]] \\ &= Var[\frac{X}{2}] + E[1] \\ &= \frac{1}{4} Var[X] + 1 = \frac{1}{4}\frac{4}{3} + 1 = \frac{4}{3}. \end{aligned}$$

■

EXERCISES

Ex. 6.2.1. Let (X, Y) be uniformly distributed on the triangle $0 < x < y < 1$.

(a) Compute $E[X|Y = \frac{1}{6}]$.

(b) Compute $E[(X - Y)^2]$.

Ex. 6.2.2. X is a random variable with mean 3 and variance 2. Y is a random variable with mean -1 and variance 6. The covariance of X and Y is -2. Let $U = X + Y$ and $V = X - Y$. Find the correlation coefficient of U and V.

Ex. 6.2.3. Suppose X and Y are both uniformly distributed on $[0, 1]$. Suppose $Cov[X, Y] = \frac{-1}{24}$. Compute the variance of $X + Y$.

Ex. 6.2.4. A dice game between two people is played by a pair of dice being thrown. One of the dice is green and the other is white. If the green die is larger than the white die, player number one earns a number of points equal to the value on the green die. If the

green die is less than or equal to the white die, then player number two earns a number of points equal to the value of the green die. Let X be the random variable representing the number of points earned by player one after one throw. Let Y be the random variable representing the number of points earned by player two after one throw.

(a) Compute the expected value of X and of Y.

(b) Without explicitly computing it, would you expect $Cov[X, Y]$ to be positive or negative? Explain.

(c) Calculate $Cov[X, Y]$ to confirm your intuition.

Ex. 6.2.5. Suppose X has variance σ_X^2, Y has variance σ_Y^2, and the pair (X, Y) has correlation coefficient $\rho[X, Y]$.

(a) In terms of σ_X, σ_Y, and $\rho[X, Y]$, find $Cov[X, Y]$ and $Cov[X + Y, X - Y]$.

(b) What must be true of σ_X^2 and σ_Y^2 if $X + Y$ and $X - Y$ are uncorrelated?

Ex. 6.2.6. Let (X, Y) have the joint probability density function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f_{X,Y}(x, y) = \begin{cases} 3(x + y) & \text{if } x > 0,\ y > 0, \text{ and } x + y < 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find $E[X|Y = \frac{1}{2}]$ and $Var[X|Y = \frac{1}{2}]$

(b) Are X and Y independent ?

Ex. 6.2.7. Suppose Y is uniformly distributed on $(0, 1)$, and suppose for $0 < y < 1$ the conditional density of $X \mid Y = y$ is given by

$$f_{X|Y=y}(x) = \begin{cases} \frac{2x}{y^2} & \text{if } 0 < x < y \\ 0 & \text{otherwise.} \end{cases}$$

(a) Show that, as a function of x, $f_{X|Y=y}(x)$ is a density.

(b) Compute the joint p.d.f. of (X, Y) and the marginal density of X.

(c) Compute the expected value and variance of X given that $Y = y$, with $0 < y < 1$.

Ex. 6.2.8. Let (X, Y) have joint probability density function $f : \mathbb{R}^2 \to \mathbb{R}$. Show that $Var[X \mid Y = y] = E[X^2 \mid Y = y] - (E[X \mid Y = y])^2$.

Ex. 6.2.9. For random variables (X, Y) as in Exercise 5.4.1, find

(a) $E[X]$ and $E[Y]$

(b) $Var[X]$ and $Var[Y]$

(c) $Cov[X, Y]$ and $\rho[X, Y]$

Ex. 6.2.10. From Example 5.4.12, consider(X, Y) have joint probability density function f given by

$$f(x, y) = \frac{\sqrt{3}}{4\pi} e^{-\frac{1}{2}(x^2 - xy + y^2)} \quad -\infty < x, y < \infty.$$

Find

(a) $E[X]$ and $E[Y]$

(b) $Var[X]$ and $Var[Y]$

(c) $Cov[X, Y]$ and $\rho[X, Y]$

Ex. 6.2.11. From Example 5.4.13, suppose $T = \{(x, y) \mid 0 < x < y < 4\}$ and let $(X, Y) \sim$ Uniform (T). Find

(a) $E[X]$ and $E[Y]$

(b) $Var[X]$ and $Var[Y]$

(c) $Cov[X, Y]$ and $\rho[X, Y]$

Ex. 6.2.12. From Example 5.4.9, consider the open disk in $\mathbb{R}^2$ given by $C = \{(x, y) : x^2 + y^2 < 25\}$ and $| C | = 25\pi$ denote its area. Let (X, Y) have a joint density $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x, y) = \begin{cases} \frac{1}{|C|} & \text{if } (x, y) \in C \\ 0 & \text{otherwise.} \end{cases}$$

Find

(a) $E[X]$ and $E[Y]$

(b) $Var[X]$ and $Var[Y]$

(c) $Cov[X, Y]$ and $\rho[X, Y]$

Ex. 6.2.13. Using the hints provided below prove the respective parts of Theorem 6.2.2

(a) Use the linearity properties of the expected value from Theorem 6.1.8.

(b) Use definition of covariance.

(c) Use the definitions of variance and covariance.

(d) Imitate the proof of Theorem 4.5.7.

(e) Use part (a) of this problem and part (f) of Theorem **??**.

(f) Use the linearity properties of the expected value from Theorem 6.1.8.

(g) Use the linearity properties of the expected value from Theorem 6.1.8.

Ex. 6.2.14. Let X, Y be continuous random variable with piecewise continuous densities $f(x)$ and $g(y)$ and well-defined expected values. Suppose $X \leq Y$ then show that $E[X] \leq E[Y]$.

Ex. 6.2.15. Let T be the triangle bounded by the lines $y = 0$, $y = 1 - x$, and $y = 1 + x$. Suppose a random vector (X, Y) has a joint p.d.f.

$$f_{(X,Y)}(x, y) = \begin{cases} 3y & \text{if} \\ 0 & \text{otherwise.} \end{cases}$$

Compute $E[Y|X = \frac{1}{2}]$.

Ex. 6.2.16. Let (X, Y) be random variables with joint probability density function $f : \mathbb{R}^2 \to \mathbb{R}$. Assume that both random variables have finite variances and that their covariance is also finite.

(a) Show that $Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$.

(b) Show that when X and Y are positively correlated (i.e. $\rho[X, Y] > 0$) then $Var[X + Y] > Var[X] + Var[Y]$, while when X and Y are negatively correlated (i.e. $\rho[X, Y] < 0$), then $Var[X + Y] < Var[X] + Var[Y]$.

6.3 MOMENT GENERATING FUNCTIONS

We have already seen for the distribution of a discrete random variable or a continuous random variable is determined by its distribution function. In this section we shall discuss the concept of moment generating functions. Under suitable assumptions, these functions will determine the distribution of random variables. They are also serve as tools in computations and come in handy for convergence concepts that we will discuss.

The moment generating function generates or determine the moments which in turn, under suitable hypothesis determine the distribution of the corresponding random variable. We begin with a definition of a moment.

Definition 6.3.1. *Suppose X is a random variable. For a positive integer k, the quantity*

$$m_k = E[X^k]$$

is known as the "k-th moment of X". As before the existence of a given moment is determined by whether the above expectation exists or not.

We have previously seen many computations of the first moment $E[X]$ and also seen that the second moment $E[X^2]$ is related to the variance of the random variable. The next theorem states that if a moment exists then it guarantees the existence of all lesser moments.

Theorem 6.3.2. *Let X be a random variable and let k be a positive integer. If $E[X^k] < \infty$ then $E[X^j] < \infty$ for all positive integers $j < k$.*

Proof - Suppose X is a continuous random variable. Suppose $E[X^k]$ exists and is finite, so that $E[|X^k|] < \infty$. Divide $\mathbb{R}$ in two pieces by letting $R_1 = \{x \in T : |x| < 1\}$ and letting $R_2 = \{x \in T : |x| \geq 1\}$. If $j < k$ then $|x|^j \leq |x|^k$ for $x \in R_2$ so,

$$\begin{aligned}
E[|X^j|] &= \int_{\mathbb{R}} |x|^j f_X(x)\,dx = \int_{R_1} |x|^j f_X(x)\,dx + \int_{R_2} |x|^j f_X(x)\,dx \\
&\leq \int_{R_1} 1 \cdot f_X(x)\,dx + \int_{R_2} |x|^k f_X(x)\,dx \\
&\leq \int_{R_1} f_X(x)\,dx + \int_{R_2} |x|^k f_X(x)\,dx \\
&= 1 + E[|X^k|] < \infty
\end{aligned}$$

Therefore $E[X^j]$ exists and is finite. See Exericse 6.3.7 when X is a discrete random variable. ■

When a random variable has finite moments for all positive integers, then these moments provide a great deal of information about the random variable itself. In fact, in some cases, these moments serve to completely describe the distribution of the random variable. One way to simultaneously describe all moments of such a variable in terms of a single expression is through the use of a "moment generating function".

Definition 6.3.3. *Suppose X is a random variable and $D = \{t \in \mathbb{R} : E[e^{tX}] \text{ exists}\}$. The function $M : D \to \mathbb{R}$ given by*

$$M(t) = E[e^{tX}],$$

is called the moment generating function for X.

The notation $M_X(t)$ will also be used when clarification is needed as to which variable a particular moment generating function belongs. Note that $M(0) = 1$ will always be true, but for other values of t, there is no guarantee that the function is even defined as the expected value might be infinite. However, when $M(t)$ has derivatives defined at zero, these values incorporate information about the moments of X. For a discrete random variable $X : S \to T$ with $T = \{x_i : i \in \mathbb{N}\}$, then for $t \in D$ (as in Definition 6.3.3)

$$M_X(t) = \sum_{i \geq 1} e^{tx_i} P(X = x_i).$$

For a continuous random variable X with probability density function $f_X : \mathbb{R} \to \mathbb{R}$ then for $t \in D$ (as in Definition 6.3.3)

$$M_X(t) = \int_{\mathbb{R}} e^{tx} f_X(x) dx.$$

We compute moment generating function for a Poisson (λ) and a Gamma (n, λ), with $n \in \mathbb{N}, \lambda > 0$.

Example 6.3.4. Suppose $X \sim$ Poisson (λ) then for all $t \in \mathbb{R}$,

$$M_X(t) = \sum_{k=0}^{\infty} e^{tk} P(X = k) = \sum_{k=0}^{\infty} e^{tk} \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^t \lambda)^k}{k!} = e^{-\lambda} e^{e^t \lambda} = e^{-\lambda(1+e^t)}.$$

So the moment generating function of X exists for all $t \in \mathbb{R}$. Suppose $Y \sim$ Gamma (n, λ) then $t < \lambda$,

$$M_Y(t) = \int_R e^{ty} \frac{\lambda^n}{\Gamma(n)} y^{n-1} e^{-\lambda y} dy = \frac{\lambda^n}{\Gamma(n)} \int_R y^{n-1} e^{-(\lambda - t)y} dy = \frac{\lambda^n}{\Gamma(n)} \frac{\Gamma(n)}{(\lambda - t)^n} = \left(\frac{\lambda}{\lambda - t}\right)^n,$$

where we have used (5.5.3). The moment generating function of Y will not be finite if $t \geq \lambda$. ■

We summarily compile some facts about moment generating functions. The proof of some of the results are beyond the scope of this text.

Theorem 6.3.5. *Suppose for a random variable X, there exists $\delta > 0$ such that $M_X(t)$ exists $(-\delta, \delta)$.*

(a) The k-th moment of X exists and is given by

$$E[X^k] = M_X^{(k)}(0),$$

where $M_X^{(k)}$ denotes the k-th derivative of M_X.

(b) For $0 \neq a \in \mathbb{R}$ such that $at, t \in (-\delta, \delta)$ we have

$$M_{aX}(t) = M_X(at).$$

(c) Suppose Y is another independent random variable such that $M_Y(t)$ exists for $t \in (-\delta, \delta)$. Then

$$M_{X+Y}(t) = M_X(t)M_Y(t).$$

for $t \in (-\delta, \delta)$.

Proof - (a) A precise proof is beyond the scope of this book. We provide a sketch. Express e^{tX} as a power series in t.

$$e^{tX} = 1 + tX + \frac{t^2X^2}{2} + \cdots + \frac{t^nX^n}{n!} + \ldots$$

The expected value of the left hand side is the moment generating function for X while linearity may be used on the right hand side. So the power series of $M(t)$ is given by

$$M(t) = 1 + t \cdot E[X] + \frac{t^2}{2} \cdot E[X^2] + \cdots + \frac{t^n}{n!} \cdot E[X^n] + \ldots$$

Taking k derivatives of both sides of the equation (which is valid in the interval of convergence) yields

$$M^{(k)}(t) = E[X^k] + t \cdot E[X^{k+1}] + \frac{t^2}{2} \cdot E[X^{k+2}] + \ldots$$

Finally, when evaluating both sides at $t = 0$ all but one term on the right hand side vanishes and the equation becomes simply $M^{(k)}(0) = E[X^k]$.

(b) $M_{aX}(t) = E[e^{(aX)t}] = E[e^{X(at)}] = M_X(at)$.

(c) Using Theorem 4.1.10 or Theorem 6.1.10 (f) we have

$$M_{X+Y}(t) = E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E[e^{tX}]E[e^{tY}] = M_X(t)M_Y(t).$$

■

Theorem 6.3.5 applies equally well for both discrete and continuous variables. A discrete example is presented next.

Example 6.3.6. Let $X \sim \text{Geometric}(p)$. We shall find $M_X(t)$ and use this function to calculate the expected value and variance X. For any $t \in \mathbb{R}$,

$$\begin{aligned} M_X(t) &= E[e^{tX}] = \sum_{n=1}^{\infty} e^{tn} P(X=n) = \sum_{n=1}^{\infty} (e^t)^n \cdot p(1-p)^{n-1} = pe^t \cdot \sum_{n=1}^{\infty} (e^t \cdot (1-p))^{n-1} \\ &= \frac{pe^t}{1-e^t(1-p)} \end{aligned}$$

Having completed that computation, the expected value and variance can be computed simply by calculating derivatives.

$$M_X'(t) = \frac{pe^t}{[1-(1-p)e^t]^2}$$

and so $E[X] = M_X'(0) = \frac{p}{p^2} = \frac{1}{p}$. Similarly,

$$M_X''(t) = \frac{pe^t + p(1-p)e^{2t}}{[1-(1-p)e^t]^3}$$

and so $E[X^2] = M_X''(0) = \frac{2p-p^2}{p^3} = \frac{2}{p^2} - \frac{1}{p}$. Therefore, $Var[X] = E[X^2] - (E[X])^2 = \frac{1-p}{p^2}$. Both the expected value and variance are in agreement with the previous computations for the goemetric random variable.

Let $Y \sim \text{Normal}(\mu, \sigma^2)$. The density of Y is $f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/2\sigma^2}$. For any $t \in \mathbb{R}$,

$$\begin{aligned} M_Y(t) &= E[e^{tY}] = \int_{-\infty}^{\infty} e^{ty} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/2\sigma^2}\, dy = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(y^2-(2\mu y+2\sigma^2 ty)+\mu^2)/2\sigma^2}\, dy \\ &= e^{\mu t+(1/2)\sigma^2 t^2} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-(\mu+\sigma^2 t))^2/2\sigma^2}\, dy \\ &= e^{\mu t+(1/2)\sigma^2 t^2} \end{aligned} \tag{6.3.1}$$

where the integral in the final step is equal to one since it integrates the density of a $\text{Normal}(\mu+\sigma^2 t, \sigma^2)$ random variable. One can easily verify that the $M_Y'(0) = \mu$ and $M_Y''(0) = \mu^2 + \sigma^2$. ■

As with the expected value and variance, moment generating functions behave well when applied to linear combinations of independent variables (courtesy Theorem 6.3.5 (b) and (c)).

Example 6.3.7. Supppose we wish to find the moment generating function of $X \sim \text{Binomial}(n, p)$. We have seen that such a random variable may arise as the sum of indpendent Bernoulli variables. That is, $X = Y_1 + \cdots + Y_n$ where $Y_j \sim \text{Bernoulli}(p)$. But it is routine to compute

$$M_{Y_j}(t) = E[e^{tY_j}] = e^{t\cdot 1}P(Y_j = 1) + e^{t\cdot 0}P(Y_j = 0) = pe^t + (1-p).$$

Therefore by linearity (inductively applying Theorem 6.3.5 (c)),

$$M_X(t) = M_{Y_1+\cdots+Y_n}(t) = M_{Y_1}(t) \cdot \ldots \cdot M_{Y_n}(t) = (pe^t + (1-p))^n.$$

■

Moment generating functions are an extraordinarily useful tool in analyzing the distributions of random variables. Two particularly useful tools involve the uniqueness and limit properties of such generating functions. Unfortunately these theorems require analysis beyond the scope of this text to prove. We will state the uniqueness fact (unproven) below and the limit property in Chapter 8. First we generalize the definition of moment generating functions to pairs of random variables.

Definition 6.3.8. *Suppose X and Y are random variables. Then the function*

$$M(s,t) = E[e^{sX+tY}]$$

is called the (joint) moment generating function for X and Y. The notation $M_{X,Y}(s,t)$ will be used when confusion may arise as to which random variables are being represented.

Moment generating functions completely describe the distributions of random variables. We state the result precisely.

Theorem 6.3.9. *(M.G.F. Uniqueness Theorem)*

(a) (One variable) Suppose X and Y are random variables and $M_X(t) = M_Y(t)$ in some open interval containing the origin. Then X and Y are equal in distribution.

(b) (Two variable) Suppose (X, W) and (Y, Z) are pairs of random variables and suppose $M_{X,W}(s,t) = M_{Y,Z}(s,t)$ in some rectangle containing the origin. Then (X, W) and (Y, Z) have the same joint distribution.

An immediate application of the theorem is an alternate proof of Corollary 5.3.3 based on moment generating functions.

Example 6.3.10. Let $X \sim \text{Normal}(\mu, \sigma^2)$ and let $Y = \frac{X-\mu}{\sigma}$. Show that $Y \sim \text{Normal}(0, 1)$.

We know X is normal, (6.3.1) shows that the moment generating function of X is $M_X(t) = e^{\mu t + (1/2)\sigma^2 t^2}$, for all $t \in \mathbb{R}$. So consider the moment generating function of Y. For all $t \in \mathbb{R}$

$$\begin{aligned} M_Y(t) &= E[e^{tY}] = E[e^{t(X-\mu)/\sigma}] = E[e^{tX/\sigma} e^{-t\mu/\sigma}] = e^{-t\mu/\sigma} \cdot M_X(\frac{t}{\sigma}) \\ &= e^{-t\mu/\sigma} \cdot e^{\mu(t/\sigma) + (1/2)\sigma^2 (t/\sigma)^2} = e^{\frac{t^2}{2}}. \end{aligned}$$

But this expression is the moment generating function of a Normal$(0, 1)$ random variable. So by the uniqueness of moment generating functions, Theorem 6.3.9 (a), the distribution of Y is Normal$(0, 1)$. ■

Just as the joint density of a pair of random variables factors as a product of marginal densities exactly when the variables are independent (Theorem 5.4.7), a similar result holds for moment generating functions.

Theorem 6.3.11. *Suppose (X, Y) are a pair of continuous random variables with moment generating function $M(s,t)$. Then X and Y are indpendent if and only if*

$$M(s,t) = M_X(s) \cdot M_Y(t).$$

Proof - One direction of the proof follows from basic facts about independence. If X and Y are independent, then by Exercise 6.3.4 , we have

$$M(s,t) = E[e^{sX+tY}] = E[e^{sX} e^{tY}] = E[e^{sX}] E[e^{tY}] = M_X(s) \cdot M_Y(t).$$

To prove the opposite direction, we shall use Theorem 6.3.9(b). Let $\hat{X}$ and $\hat{Y}$ be independent, but have the same distributions as X and Y respectively. Since $M_{X,Y}(s,t) = M_X(s)M_Y(t)$ we have the following series of equalities:

$$M_{X,Y}(s,t) = M_X(s)M_Y(t) = M_{\hat{X}}(s)M_{\hat{Y}}(t) = M_{\hat{X},\hat{Y}}(s,t).$$

By Theorem 6.3.9(b), this means that (X,Y) and $(\hat{X},\hat{Y})$ have the same distribution. This would imply that

$$P(X \in A, Y \in B) = P(\hat{X} \in A, \hat{Y} \in B) = P(\hat{X} \in A)P(\hat{Y} \in B) = P(X \in A)P(Y \in B),$$

for any events A and B. Hence X and Y are independent. ■

Notice that the method employed in Example 6.3.10 did not require considering integrals directly. Since the manipulation of integrals can be complicated (particularly when dealing with multiple integrals), the moment generating function method will often be simpler as the next example illustrates.

Example 6.3.12. Let a, b be two real numbers. Let $X \sim \text{Normal}(\mu_1, \sigma_1^2)$ and $Y \sim \text{Normal}(\mu_2, \sigma_2^2)$ be independent. Observe that

$$M_{aX+bY}(t) \quad = \quad M_{X,Y}(at, bt)$$

Using Theorem 6.3.11, we have that the above is

$$M_X(at)M_Y(bt) = e^{a\mu_1 t + (1/2)a^2\sigma_1^2 t^2} e^{b\mu_2 t + (1/2)b^2\sigma_2^2 t^2} = e^{(a\mu_1 + b\mu_2)t + (1/2)(a^2\sigma_1^2 + b^2\sigma_2^2)t^2}$$

which is the moment generating function of a Normal random variable with mean $a\mu_1 + b\mu_2$ and variance $a^2\sigma_1^2 + b^2\sigma_2^2$). So $aX + bY \sim \text{Normal}(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$. ■

We conclude this section with a result on finite linear combinations of independent normal random variables.

Theorem 6.3.13. *Let $X_1, X_2, \ldots, X_n$ be independent, normally distributed random variables with mean μ_i and variance σ_i^2 respectively for $i = 1, 2, \ldots n$. Let $a_1, a_2, \ldots, a_n$ be real-valued numbers, not all of which are zero. Then then the linear combination $Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$ is also normally distributed with mean $\sum_{i=1}^n a_i\mu_i$ and variance $\sum_{i=1}^n a_i^2\sigma_i^2$.*

Proof- This follows from the preceeding example by induction and is left as an exercise. ■

EXERCISES

Ex. 6.3.1. Let $X \sim \text{Normal}(0, 1)$. Use the moment generating function of X to calcluate $E[X^4]$.

Ex. 6.3.2. Let $Y \sim \text{Exponential}(\lambda)$.

(a) Calculate the moment generating function $M_Y(t)$.

(b) Use (a) to calculate $E[Y^3]$ and $E[Y^4]$, the third and fourth moments of an exponential distriubtion.

Ex. 6.3.3. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables.

(a) Let $Y = X_1 + \cdots + X_n$. Prove that $M_Y(t) = [M_{X_1}(t)]^n$.

(b) Let $Z = (X_1 + \cdots + X_n)/n$. Prove that $M_Z(t) = [M_{X_1}(\frac{t}{n})]^n$.

Ex. 6.3.4. Let X and Y be two independent discrete random variables. Let $h : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$. Show that

$$E[h(X)g(Y)] = E[h(X)]E[g(Y)].$$

Show that the above holds if X and Y are independent continous random variables.

Ex. 6.3.5. Suppose X is a discrete random variable and $D = \{t \in \mathbb{R} : E[t^X] \text{ exists}\}$. The function $\psi : D \to \mathbb{R}$ given by

$$\psi(t) = E[t^X],$$

is called the probability generating function for X. Calculate the probability generating function of X when X is

(a) $X \sim \text{Bernoulli}(p)$, with $0 < p < 1$.

(b) $X \sim \text{Binomial}(n, p)$, with $0 < p < 1$, $n \geq 1$.

(c) $X \sim \text{Geometric}(p)$, with $0 < p < 1$.

(d) $X \sim \text{Poisson}\ (\lambda)$, with $0 < \lambda$.

Ex. 6.3.6. Let $X, Y : S \to T$ be dicrete random variables with the number of elements in T is finite. Prove part (a) of Theorem 6.3.9 in this case.

Ex. 6.3.7. Prove Theorem 6.3.2 when X is a discrete random variable.

6.4 BIVARIATE NORMALS

In Example 6.3.12, we saw that if X and Y are independent, normally distributed random variables, any linear combination $aX + bY$ is also normally distributed. In such a case the joint density of (X, Y) is determined easily (courtesy Theorem 5.4.7). We would like to understand random variables that are not independent but have normally distributed marginals. Motivated by the observations in Example 6.3.12 we provide the following definition.

Definition 6.4.1. *A pair of random variables (X, Y) is called "bivariate normal" if $aX + bY$ is a normally distributed random variable for all real numbers a and b.*

We need to be somewhat cautious in the above definition. Since the variables are dependent it may turn out that $aX + bY = 0$ or some constant. (E.g: $Y = -X$,or $Y = -X + 2$ with $a = 1, b = 1$). We shall follow the convention that a constant c random variable in such cases is a normal random variable with mean c and variance 0.

If (X, Y) are bivariate normal then as $X = X + 0Y$ and $Y = 0X + Y$ both X and Y individually are normal random variables. The converse if not true (See Exercise 6.4.3). However the joint distribution of bivariate normal random variables are determined by their means, variances and covariances. This fact is proved next.

Theorem 6.4.2. *Suppose (X, Y) and (Z, W) are two bivariate normal random variables. If*

$$\begin{aligned} E[X] = E[Z] = \mu_1, \qquad & E[Y] = E[W] = \mu_2 \\ Var[X] = Var[Z] = \sigma_1^2, \qquad & Var[Y] = Var[W] = \sigma_2^2 \\ & \textit{and} \\ Cov[X, Y] = Cov[Z, W] = \sigma_{12} & \end{aligned} \tag{6.4.1}$$

then (X, Y) and (Z, W) have the same joint distribution.

Proof- As (X, Y) and (Z, W) are bivariate normal random variables, given real numbers s, t $sX + tY$ and $sZ + tW$ are normal random variables. Using (6.4.1) and the properties of mean and covariance (see Theorem 6.2.2) we have

$$\begin{aligned}
E[sX + tY] &= sE[X] + tE[Y] = s\mu_1 + t\mu_2, \\
E[sZ + tW] &= sE[Z] + tE[W] = s\mu_1 + t\mu_2, \\
Var[sX + tY] &= s^2 Var[X] + t^2 Var[Y] + 2stCov[X, Y] \\
&= s^2\sigma_1^2 + t^2\sigma_2^2 + 2st\sigma_{12}, \\
&\quad \text{and} \\
Var[sZ + tW] &= s^2 Var[Z] + t^2 Var[W] + 2stCov[Z, W] \\
&= s^2\sigma_1^2 + t^2\sigma_2^2 + 2st\sigma_{12}.
\end{aligned}$$

From the above, $sX + tY$ and $sZ + tW$ have the same mean and variance. So they have the same distribution (as normal random variables are determined by their mean and variances). By Theorem 6.3.9 (a) they have the same moment generating function. So, the (joint) moment generating function of (X, Y) at (s,t) is

$$M_{X,Y}(s, t) = E[e^{sX+tY}] = M_{sX+tY}(1) = M_{sZ+tW}(1) = E[e^{sZ+tW}] = M_{Z,W}(s, t)$$

Therefore (Z, W) has the same joint m.g.f. as (X,Y) and Theorem 6.3.9 (b) implies that they have the same joint distribution. ■

Though, in general, two variables which are uncorrelated may not be independent, it is a remarkable fact that the two concepts are equivalent for bivariate normal random variables.

Theorem 6.4.3. *Let (X, Y) be a bivariate normal random variable. Then $Cov[X, Y] = 0$ if and only if X and Y are independent.*

Proof - That independence implies a zero covariance is true for any pair of random variables (use Theorem 6.1.10 (e)), so we need to only consider the reverse implication.

Suppose $Cov[X, Y] = 0$. Let μ_X and σ_X^2 denote the expected value and variance of X and μ_Y and σ_Y^2 the corresponding values for Y. Let s and t be real numbers. Then, by the bivariate normality of (X, Y), we know $sX + tY$ is normally distributed. Moreover by properties of expected value and variance we have

$$E[sX + tY] = sE[X] + tE[Y] = s\mu_X + t\mu_Y$$

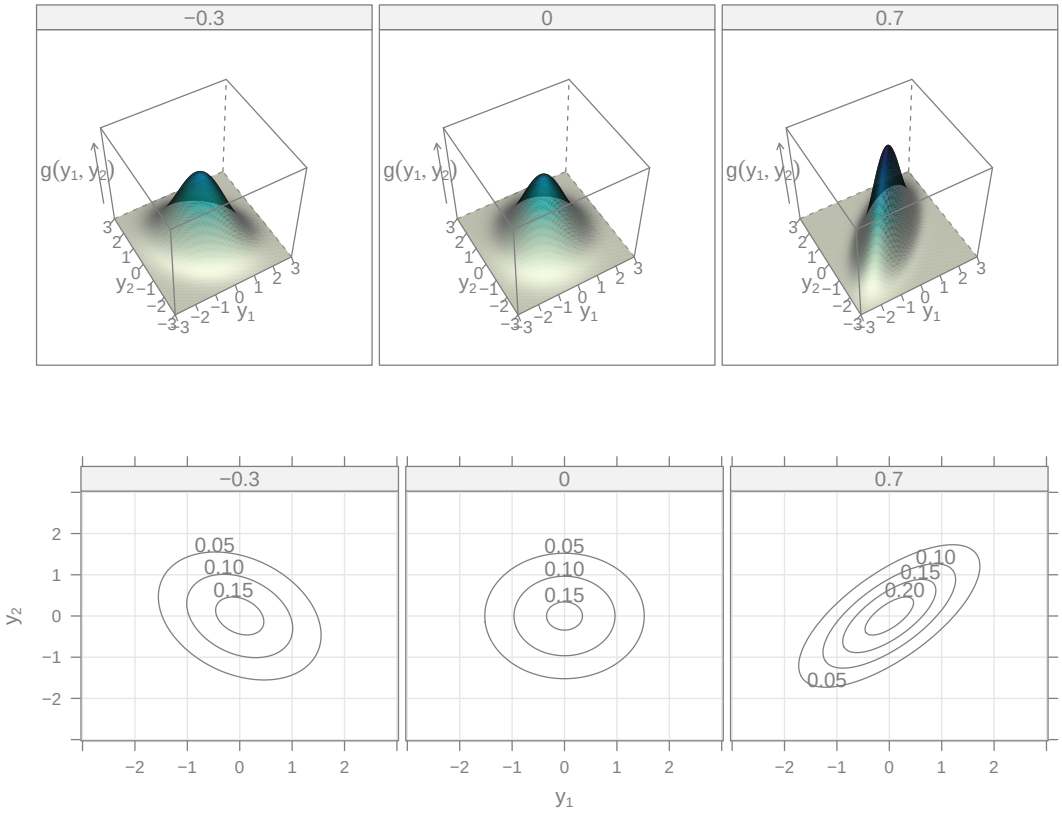


Figure 6.1: The density function of Bivariate Normal distributions. The set of panels on top show a three-dimensional view of the density function for various values of the correlation ρ. The bottom set of panels show contour plots, where each ellipse corresponds to the (y_1, y_2) pairs corresponding to a constant value of $g(y_1, y_2)$.

and

$$Var[sX + tY] = s^2 Var[X] + 2stCov[X, Y] + t^2 Var[Y] = s^2\sigma_X^2 + t^2\sigma_Y^2.$$

That is, $sX + tY \sim \text{Normal}(s\mu_X + t\mu_Y, s^2\sigma_X^2 + t^2\sigma_Y^2)$. So for all $s, t \in \mathbb{R}$

$$\begin{aligned} M_{X,Y}(s,t) &= E[e^{sX+tY}] = M_{sX+tY}(1) = e^{(s\mu_X + t\mu_Y) + (1/2)(s^2\sigma_X^2 + t^2\sigma_Y^2)} \\ &= e^{s\mu_x + (1/2)s^2\sigma_X^2} \cdot e^{t\mu_Y + (1/2)t^2\sigma_Y^2} \\ &= M_X(s) \cdot M_Y(t). \end{aligned}$$

Hence by Theorem 6.3.11 X and Y are independent. ■

We conclude this section by finding the joint density of a Bivariate normal random variable. See Figure 6.1 for a graphical display of this density.

Theorem 6.4.4. *Let (Y_1, Y_2) be a bivariate Normal random variable, with $\mu_1 = E[Y_1], \mu_2 = E[Y_2], 0 \neq \sigma_1^2 = Var[Y_1], 0 \neq \sigma_2^2 = Var[Y_2]$, and $\sigma_{12} = Cov[Y_1, Y_2]$. Assume that the correlation coefficient $|\rho[Y_1, Y_2]| \neq 1$. Then the joint probability density function of (Y_1, Y_2), $g : \mathbb{R}^2 \to [0, \infty)$ is given by*

$$g(y_1, y_2) = \frac{\exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{y_1-\mu_1}{\sigma_1}\right)\left(\frac{y_2-\mu_2}{\sigma_2}\right)\right]\right)}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \tag{6.4.2}$$

Proof- Let a, b be two real numbers. We will show that

$$P(Y_1 \leq a, Y_2 \leq b) = \int_{-\infty}^{a}\int_{-\infty}^{b} g(y_1, y_2)dy_2dy_1. \tag{6.4.3}$$

From the discussion that follows (5.4.1), we can then conclude that the joint density of (Y_1, Y_2) is indeed given by g. To show (6.4.3) we find an alternate description of (Y_1, Y_2) which is the same in distribution. Let Z_1, Z_2 be two independent standard normal random variables. Define

$$\begin{aligned} U &= \sigma_1 Z_1 + \mu_1 \\ V &= \sigma_2(\rho Z_1 + \sqrt{1-\rho^2}Z_2) + \mu_2 \end{aligned} \tag{6.4.4}$$

Let $\alpha, \beta \in \mathbb{R}$. Then

$$\alpha U + \beta V = (\alpha\sigma_1 + \beta\sigma_2\rho)Z_1 + (\beta\sigma_2\sqrt{1-\rho^2})Z_2 + \alpha_1\mu_1 + \beta\mu_2.$$

As Z_1 and Z_2 are independent standard normal random variables by Theorem 6.3.13, $(\alpha\sigma_1 + \beta\sigma_2\rho)Z_1 + (\beta\sigma_2\sqrt{1-\rho^2})Z_2 \sim$ Normal $(0, (\alpha\sigma_1 + \beta\sigma_2\rho)^2 + (\beta\sigma_2\sqrt{1-\rho^2})^2)$. Further using Corollary 5.3.3 (a) we have that $\alpha U + \beta V \sim$ Normal $(\alpha_1\mu_1 + \beta\mu_2, (\alpha\sigma_1 + \beta\sigma_2\rho)^2 + (\beta\sigma_2\sqrt{1-\rho^2})^2)$. As α, β were arbitrary real numbers by Definition 6.4.1, (U, V) is a bivariate normal random variable.

Using Theorem 6.1.8 and Theorem 6.1.10 (d) that,

$$\mu_1 = E[U], \mu_2 = E[V], \text{Var}[U] = \sigma_1^2.$$

Also in addition, using Exercise 6.2.16 and Theorem 6.2.2 (f), we have

$$\begin{aligned} Var[V] &= \sigma_2^2\rho^2 Var[Z_1] + \sigma_2^2(1-\rho^2)Var[Z_2] + 2(\sigma_2(\rho+\sqrt{1-\rho^2})Cov[Z_1,Z_2] \\ &= \sigma_2^2\rho^2 + \sigma_2^2(1-\rho^2) + 0 = \sigma_2^2 \\ &\quad \text{and} \\ Cov[U,V] &= Cov[\sigma_1 Z_1 + \mu_1, \sigma_2(\rho Z_1 + \sqrt{1-\rho^2}Z_2)] \\ &= \sigma_1\sigma_2\rho Cov[Z_1,Z_1] + \sigma_1\sigma_2\sqrt{1-\rho^2}Cov[Z_1,Z_2] \\ &= \sigma_1\sigma_2\rho + 0 = \sigma_{12}. \end{aligned}$$

As bivariate normal random variables are by their means and covariances (by Theorem 6.4.2), (Y_1, Y_2) and (U, V) have the same joint distribution. By the above, we have

$$P(Y_1 \leq a, Y_2 \leq b) = P(U \leq a, V \leq b). \tag{6.4.5}$$

By elementary algebra we can also infer from (6.4.4)

$$Z_1 = \frac{U-\mu_1}{\sigma_1},\ Z_2 = \frac{V-\mu_2}{\sigma_2\sqrt{1-\rho^2}} - \frac{\rho Z_1}{\sqrt{1-\rho^2}}.$$

So

$$\{U \leq a,\ V \leq b\} = \left\{ Z_1 \leq \frac{a-\mu_1}{\sigma_1},\ Z_2 \leq \frac{b-\mu_2}{\sigma_2\sqrt{1-\rho^2}} - \frac{\rho Z_1}{\sqrt{1-\rho^2}} \right\}$$

So, using this fact in (6.4.5) we get

$$\begin{aligned} P(Y_1 \leq a, Y_2 \leq b) &= P\left(Z_1 \leq \frac{a-\mu_1}{\sigma_1},\ Z_2 \leq \frac{b-\mu_2}{\sigma_2\sqrt{1-\rho^2}} - \frac{\rho Z_1}{\sqrt{1-\rho^2}} \right) \\ &= \int_{-\infty}^{\frac{a-\mu_1}{\sigma_1}} \int_{-\infty}^{\frac{b-\mu_2}{\sigma_2\sqrt{1-\rho^2}} - \frac{\rho z_1}{\sqrt{1-\rho^2}}} \frac{\exp(-\frac{z_1^2+z_2^2}{2})}{2\pi} dz_2 dz_1 \end{aligned} \tag{6.4.6}$$

First performing a u-substitution in the inner integral for each fixed z_1,

$$z_2 = \frac{y_2-\mu_2}{\sigma_2\sqrt{1-\rho^2}} - \frac{\rho z_1}{\sqrt{1-\rho^2}}$$

yields that the inner integral in (6.4.6) for each $z_1 \in \mathbb{R}$

$$\begin{aligned}
&\int_{-\infty}^{\frac{b-\mu_2}{\sigma_2\sqrt{1-\rho^2}}-\frac{\rho z_1}{\sqrt{1-\rho^2}}} \frac{\exp(-\frac{z_1^2+z_2^2}{2})}{2\pi} dz_2 = \int_{-\infty}^{b} \frac{\exp(-\frac{z_1^2+\left(\frac{y_2-\mu_2}{\sigma_2\sqrt{1-\rho^2}}-\frac{\rho z_1}{\sqrt{1-\rho^2}}\right)^2}{2})}{2\pi\sigma_2\sqrt{1-\rho^2}} dy_2 \\
&= \int_{-\infty}^{b} \frac{\exp(-\frac{1}{2(1-\rho^2)}[(1-\rho^2)z_1^2 + (\frac{y_2-\mu_2}{\sigma_2} - \rho z_1)^2])}{2\pi\sigma_2\sqrt{1-\rho^2}} dy_2 \\
&= \int_{-\infty}^{b} \frac{\exp(-\frac{1}{2(1-\rho^2)}[z_1^2 + (\frac{y_2-\mu_2}{\sigma_2})^2 - 2\rho(\frac{y_2-\mu_2}{\sigma_2})z_1])}{2\pi\sigma_2\sqrt{1-\rho^2}} dy_2.
\end{aligned}$$

Substituting the above into (6.4.6), we have

$$P(Y_1 \le a, Y_2 \le b) = \int_{-\infty}^{\frac{a-\mu_1}{\sigma_1}} \int_{-\infty}^{b} \frac{\exp(-\frac{1}{2(1-\rho^2)}[z_1^2 + (\frac{y_2-\mu_2}{\sigma_2})^2 - 2\rho(\frac{y_2-\mu_2}{\sigma_2})z_1])}{2\pi\sigma_2\sqrt{1-\rho^2}} dy_2.dz_1 \tag{6.4.7}$$

Performing a u-subsitution

$$z_1 = \frac{y_1-\mu_1}{\sigma_1}$$

on the outer integral above we obtain

$$\begin{aligned}
&P(Y_1 \le a, Y_2 \le b) \\
&= \int_{-\infty}^{a} \int_{-\infty}^{b} \frac{\exp(-\frac{1}{2(1-\rho^2)}[\left(\frac{y_1-\mu_1}{\sigma_1}\right)^2 + (\frac{y_2-\mu_2}{\sigma_2})^2 - 2\rho(\frac{y_2-\mu_2}{\sigma_2})\left(\frac{y_1-\mu_1}{\sigma_1}\right)])}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} dy_2 dy_1
\end{aligned}$$

Thus we have established (6.4.3). ■

EXERCISES

Ex. 6.4.1. Let X_1, X_2 be two independent Normal random variables with mean 0 and variance 1. Show that (X_1, X_2) is a bivariate normal random variable.

Ex. 6.4.2. Let (X_1, X_2) be a bivariate normal random variable. Assume that the correlation coefficient $|\rho[X_1, X_2]| \neq 1$. Show that X_1 and X_2 are Normal random variables by calculating their marginal densities.

Ex. 6.4.3. Let X_1, X_2 be two independent normal random variables with mean 0 and variance 1. Let (Y_1, Y_2) be a bivariate normal random variable with zero means, variances equal to 1 and correlation $\rho = \rho[Y_1, Y_2]$, with $\rho^2 \neq 1$. Let f be the joint probability density function of (X_1, X_2)

and g be the joint probability density function of (Y_1, Y_2). For $0 < \alpha < 1$, let (Z_1, Z_2) be a bivariate random variable with joint density given by

$$h(z_1, z_2) = \alpha g(z_1, z_2) + (1 - \alpha) f(z_1, z_2),$$

for any real numbers z_1, z_2.

(a) Write down the exact expressions for f and g.

(b) Verify that h is indeed a probability density function.

(c) Show that Z_1 and Z_2 are Normal random variables by calculating their marginal densities.

(d) Show that (Z_1, Z_2) is not a bivariate normal random variable.

Ex. 6.4.4. Suppose $X_1, X_2, \ldots, X_n$ are independent and normaly distributed. Let $Y = c_1 X_1 + \cdots + c_n X_n$ and let $Z = d_1 X_1 + \cdots + d_n X_n$ be linear combinations of these variables (for real numbers c_j and d_j). Then (Y, Z) is bivariate normal.

Ex. 6.4.5. Prove Theorem 6.3.13. Specifically, suppose for $i = 1, 2, \ldots, n$ that $X_i \sim \text{Normal}(\mu_i, \sigma_i^2)$ with $X_1, X_2, \ldots, X_n$ independent. Let $a_1, a_2, \ldots, a_n$ be real numbers, not all zero, and let $Y = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$. Prove that Y is normally distributed and find its mean and variance in terms of the a's, μ's, and σ's.

Ex. 6.4.6. Let (X_1, X_2) be a bivariate Normal random variable. Define

$$\Sigma = \begin{bmatrix} Cov[X_1, X_1] & Cov[X_1, X_2] \\ \\ Cov[X_1, X_2] & Cov[X_2, X_2] \end{bmatrix}$$

$$\text{and } \mu_1 = E[X_1], \mu_2 = E[X_2], \mu_{2\times 1} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}.$$

Σ is referred to as the covariance matrix of (X_1, X_2) and μ is the mean matrix of (X_1, X_2).

(a) Compute $\det(\Sigma)$.

(b) Show that the joint density of (X_1, X_2) can be rewritten in matrix notation as

$$g(x_1, x_2) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 & x_2 - \mu_2 \end{bmatrix} \Sigma^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right)$$

(c)

$$A_{2\times 2} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \eta_{2\times 1} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}$$

such that a_{ij} are real numbers. Suppose we define

$$Y = AX = \begin{bmatrix} a_{11} X_1 + a_{12} X_2 + \eta_1 \\ a_{21} X_1 + a_{22} X_2 + \eta_2 \end{bmatrix}.$$

Then (Y_1, Y_2) is also a bivariate Normal random variable, with covariance matrix $A\Sigma A^T$ and mean matrix $A\mu + \eta$.
Hint: Compute means, variances and covariances of Y_1, Y_2 and use Theorem 6.4.2

7

SAMPLING AND DESCRIPTIVE STATISTICS

The distinction between Probability and Statistics is somewhat blurred, but largely has to do with the perspective of what is known versus what is to be determined. One may think of Probability as the study of models for (random) experiments when the model is fully known. When the model is not fully known and one tries to infer about the unknown aspects of the model based on observed outcomes of the experiment, this is where Statistics enters the picture. In this chapter we will be interested in problems where we assume we know the outputs of random variables, and wish to use that information to say what we can about their (unknown) distributions.

Suppose, for instance, we sample from a large population and record a numerical fact associated with each selection. This may be recording the heights of people, recording the arsenic content of water samples, recording the diameters of randomly selected trees, or anything else that may be thought of as repeated, random measurements. Sampling an individual from a population in this case may be viewed as a random experiment. If the sampling were done at random with replacement with each selection independent of any other, we could view the resulting numerical measurements as i.i.d. random variables $X_1, X_2, \ldots, X_n$. A more common situation is sampling without replacement, but we have previously seen (see Section 2.3) that when the sample size is small relative to the size of the population, the two sampling methods are not dramatically different. In this case we have the results of n samples from a distribution, but we do not actually know the distribution itself. How might we use the samples to attempt to predict or "infer" such things as expected value and variance?

7.1 DESCRIPTIVE STATISTICS

A natural quantity we can create from the observed data, regardless of the underlying distribution that generated it, is a discrete distribution that puts equal probability on each observed point. This distribution is known as the empirical distribution. Inferences based on the empirical distribution are traditionally referred to as "descriptive statistics". In later chapters, we will see that making additional assumptions lets us make "better" inferences, provided the additional assumptions are valid.

We will assume that the random variables $X_1, X_2, \ldots, X_n$ are i.i.d. from some common distribution, usually unknown. Some values of X_i can of course be repeated, so the empirical distribution (and the empirical cummulative distribution function) is formally defined as follows.

Definition 7.1.1. *Let $X_1, X_2, \ldots, X_n$ be random variables be i.i.d with distribution X. The "empirical distribution" based on these is the discrete distribution with probability mass function given by*

$$f_n(t) = \frac{1}{n} \left| \{i : X_i = t\} \right|,$$

for $t \in \mathbb{R}$. Further, for $x \in \mathbb{R}$

$$F_n(x) = \frac{|\{i : X_i \leq x\}|}{n},$$

is known as the "empirical cumulative distribution function" or ECDF of $X_1, X_2, \ldots, X_n$.

Given a realisation of $X_1, X_2, \ldots, X_n$ ECDF are easy to compute and provide information about the underlying distribution. One can also show that as $n \to \infty$ the ECDF will converge to the underlying distribution function.

Example 7.1.2. Suppose we surveyed 10 random people and asked them how many litres of water they consume in a day. Suppose the data collected was the following:

$$3 \quad 4 \quad 2 \quad 5 \quad 2 \quad 4 \quad 4 \quad 6 \quad 3 \quad 4$$

We can compute the empirical probability mass function and the empirical cummulative distribution function. That is,

$$f_{10}(t) = \begin{cases} \frac{2}{10} & \text{if } t = 2, \\ \frac{2}{10} & \text{if } t = 3, \\ \frac{4}{10} & \text{if } t = 4, \\ \frac{1}{10} & \text{if } t = 5, \\ \frac{1}{10} & \text{if } t = 6 \\ \text{and} & \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad F_{10}(x) = \begin{cases} 0 & \text{if } x < 2, \\ \frac{2}{10} & \text{if } 2 \leq x < 3, \\ \frac{4}{10} & \text{if } 3 \leq x < 4, \\ \frac{8}{10} & \text{if } 4 \leq x < 5, \\ \frac{9}{10} & \text{if } 5 \leq x < 6, \\ \text{and} & \\ 1 & \text{if } 6 \leq x \end{cases}$$

R has an built function called `ecdf` which will compute the empirical cumulative distribution function given the data with options for plotting as indicated below.

```
x = c(3, 4,  2 , 5 , 2 , 4 , 4 , 6 , 3 , 4 )
F= ecdf(x)
plot(F)
```

■

Note that, the empirical distribution is a random object, as it is defined in terms of random variables. However, for any *fixed* realisation of these random variables $X_1, X_2, \ldots, X_n$, the corresponding empirical distribution is a fixed probability distribution, so we can now study it using

the tools of probability. Doing so does not make any additional assumptions about the underlying distribution.

It is important to realize that the empirical distribution is itself a random quantity, as each sample realisation will produce a different discrete distribution. We intuitively expect it to carry information about the underlying distribution, especially as the sample size n grows. For example, the expectation computed from the empirical distribution should be closely related to the true underlying expectation, probabilities of events computed from the empirical distribution should be related to the true probabilities of those events, and so on. In the remainder of this chapter, we will make this intuition more precise and describe some tools to investigate the properties of the empirical distribution.

7.1.1 *Sample Mean and Sample Variance*

Given a sample of observations $X_1, X_2, \ldots, X_n$ from a distribution X, we define the sample mean to be the familiar definition of average. We shall present the precise definition first and then a result that describes how well does the sample mean work as an estimate of the true mean of the distribution X.

Definition 7.1.3. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with distribution X. The "sample mean" of these is*

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

It is easy to see that $\overline{X}$ is the expected value of a *random variable* whose distribution is the empirical distribution based on $X_1, X_2, \ldots, X_n$ (see Exercise 7.1.5). Suppose the X_j random variables have a finite expected value μ. The sample mean $\overline{X}$ is not the same as this expected value. In particular μ is a fixed constant while $\overline{X}$ is a random variable. From the statistical perspective, μ is usually assumed to be an unknown quantity while $\overline{X}$ is something that may be computed from the results of the sample $X_1, X_2, \ldots, X_n$. The next theorem is a first step in answering how well does $\overline{X}$ work as an estimate of μ.

Theorem 7.1.4. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample of random variables whose distribution has finite expected value μ and finite variance σ^2. Let $\overline{X}$ represent the sample mean. Then*

$$E[\overline{X}] = \mu \quad \text{and} \quad SD[\overline{X}] = \frac{\sigma}{\sqrt{n}}.$$

Proof. We can write

$$\begin{aligned} E\left[\overline{X}\right] &= E\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] \\ &= \frac{E[X_1] + E[X_2] + \cdots + E[X_n]}{n} \\ &= \frac{n\mu}{n} = \mu \end{aligned}$$

To calculate the standard deviation, we consider the variance and use Theorem 4.2.6 and Exercise 6.1.12 to obtain

$$\begin{aligned} Var\left[\overline{X}\right] &= Var\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] \\ &= \frac{Var[X_1] + Var[X_2] + \cdots + Var[X_n]}{n^2} \\ &= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \end{aligned}$$

Taking square roots then shows $SD\left[\overline{X}\right] = \frac{\sigma}{\sqrt{n}}$. ■

The fact that $E[\overline{X}] = \mu$ implies that, on average, the quantity $\overline{X}$ is accurately describing the unknown mean μ. In the language of statistics $\overline{X}$ is said to be an "unbiased estimator" of the quantity μ. Note also that $SD[\overline{X}] \to 0$ as $n \to \infty$ meaning that the larger the sample size, the more accurately $\overline{X}$ reflects its average of μ. In other words, if there is an unknown distribution from which it is possible to sample, averaging a large sample should produce a value close to the expected value of the distribution. In technical terms, this is considered as a notion of consistency and we say that the sample mean is a "consistent estimator" of the population mean μ.

Given a sample of observations from a given distribution one may try to estimate the variance of the distribution via the sample variance which we define below.

Definition 7.1.5. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables. The "sample variance" of these is*

$$S^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}.$$

Note that this definition is not universal; it is common to define sample variance with n (instead of $n-1$) in the denominator, in which case the definition matches the variance of the empirical distribution of $X_1, X_2, \ldots, X_n$ (Exercise 7.1.5). The definition given here produces a quantity that is unbiased for the underlying population variance, a fact that follows from the next theorem.

Theorem 7.1.6. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample of random vairables whose distribution has finite expected value μ and finite variance σ^2. Then S^2 is an unbiased estimator of σ^2, i.e.,*

$$E[S^2] = \sigma^2.$$

Proof. First note that

$$E[\overline{X}^2] = Var[\overline{X}] + (E[\overline{X}])^2 = \frac{\sigma^2}{n} + \mu^2$$

whereas

$$E[X_j^2] = Var[X_j] + E[X_j]^2 = \sigma^2 + \mu^2.$$

Now consider the quantity $(n-1)S^2$.

$$\begin{aligned} E[(n-1)S^2] &= E[(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2] \\ &= E[X_1^2 + X_2^2 + \cdots + X_n^2] - 2E[(X_1 + X_2 + \cdots + X_n)\overline{X}] \\ &\qquad + E[\overline{X}^2 + \overline{X}^2 + \cdots + \overline{X}^2] \end{aligned}$$

But $X_1 + X_2 + \cdots + X_n = n\overline{X}$, so

$$\begin{aligned} E[(n-1)S^2] &= E[X_1^2 + X_2^2 + \cdots + X_n^2] - 2nE[\overline{X}^2] + nE[\overline{X}^2] \\ &= E[X_1^2 + X_2^2 + \cdots + X_n^2] - nE[\overline{X}^2] \\ &= n(\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2) = (n-1)\sigma^2 \end{aligned}$$

Dividing by $n-1$ gives the desired result, $E[S^2] = \sigma^2$. ■

A more important property (than unbiasedness) is that S^2 and its variant with n in the denominator are both "consistent" for σ^2, just as $\overline{X}$ was for μ, in the sense that $Var[S^2] \to 0$ as $n \to \infty$ under some mild conditions (See Exercise 7.1.7). One may also try to estimate σ from S but due to vagaries of averaging (in turn expectation) one will typically loose the unbiasedness property (See Exercise 7.1.8).

Expectation and variance are commonly used summaries of a random variable, but they do not characterize its distribution completely. In the next subsection we shall see how to use the idea of sample proportion to understand the underlying distribution better.

7.1.2 *Sample proportion*

In general, the distribution of a random variable X is fully known if we can compute $P(X \in A)$ for any event A. On the other hand if the distribution is not known and we have an event A of interest then we can use the empirical distribution to estimate the probability $P(X \in A)$.

Given a sample of i.i.d. observations $X_1, X_2, \ldots, X_n$ from a common distribution defined by a random variable X, let Y be the random variable that has the same distribution as the empirical distribution based on sample. More precisely,

$$\text{Range}(Y) = \{X_1, X_2, \ldots, X_n\} \qquad \text{and} \qquad P(Y = t) = \frac{|\{i : X_i = t|\}}{n}, \text{ for } t \in \text{Range}(Y).$$

Let A be the event of interest, then

$$P(Y \in A) = \frac{|\{i : X_i \in A\}|}{n}.$$

In other words, $P(Y \in A)$ is simply the proportion of sample observations for which the event A happened. Not surprisingly, $P(Y \in A)$ is a good estimator of $P(X \in A)$.

Theorem 7.1.7. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample of random variables with the same distribution as a random variable X. Suppose that we are interested in the value $p = P(X \in A)$ for an event A. Let*

$$\hat{p}_n = \frac{|\{i : X_i \in A\}|}{n}.$$

Then, $E[\hat{p}_n] = P(X \in A)$ and $Var(\hat{p}_n) \to 0$ as $n \to \infty$.

Proof. For $1 \leq i \leq n$, let

$$Z_i = \begin{cases} 1 & \text{if } X_i \in A \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that

$$|\{i : X_i \in A\}| = \sum_{i=1}^{n} Z_i$$

and Z_i's are independent because X_i's are independent (See Theorem 3.3.6 and Exercise 7.1.2) and identically distributed with

$$P(Z_i = 1) = P(X_i \in A) = p.$$

Thus, $\sum_{i=1}^{n} Z_i$ has the Binomial distribution with parameters n and p, with expectation np and variance $np(1-p)$. It is immediate that

$$E[\hat{p}_n] = E[\frac{\sum_{i=1}^{n} Z_i}{n}] = \frac{1}{n} E[\sum_{i=1}^{n} Z_i] = p$$

and

$$Var[\hat{p}_n] = Var[\frac{\sum_{i=1}^{n} Z_i}{n}] = \frac{1}{n^2} Var[\sum_{i=1}^{n} Z_i] = p(1-p)/n. \tag{7.1.1}$$

The result follows. ∎

This result is a special case of the more general "law of large numbers" we will encounter in Section 8.2. It is important because it gives formal credence to our intuition that the probability of an event measures the limiting relative frequency of that event over repeated trials of an experiment.

Example 7.1.8. Suppose that U and V are independent Uniform$(0, 1)$, and we interpret (U, V)as coordinates of a point in $\mathbb{R}^2$. Let A be the event that the point (U, V) is inside the unit circle

and we wish to estimate the probability of $p = P((U,V) \in A) = P(U^2 + V^2 < 1)$. We can find the answer by a direct computation (See Exercise 7.1.6) that $p = \frac{\pi}{4}$. We can estimate this probability using the sampling proportion. We simulate the experiment a large number of times, and computing the proportion of cases where the simulated Z is less than 1. That is generate samples $\{(U_i, V_i) : 1 \leq i \leq n\}$

```
replicate(10, {
    u <- runif(10000)
    v <- runif(10000)
    z <- sqrt(u^2 + v^2)
    sum(z < 1) / 10000
})
```

```
 [1] 0.7820 0.7791 0.7882 0.7834 0.7888 0.7802 0.7861 0.7816 0.7813
[10] 0.7872
```

We can see that our estimates are quite good with $n = 10000$ that $\hat{p}_n$ is very close to p. A little thought tells us that the true probability $P(Z < 1) = \pi/4 \approx 0.7854$. The simulation experiment we have performed above is in fact one way of estimating π, although it is not a particularly efficient one. We illustrate this below by repeating the experiment with 1000000 trials and multiplying the observed sample proportion by 4. Note that in this experiment $z < 1 \iff z^2 < 1$, so calculating the square root is unnecessary.

```
u <- runif(1000000)
v <- runif(1000000)
zsq <- u^2 + v^2
4 * mean(zsq < 1)
```

```
[1] 3.13966
```

As the variance of the sample proportion $\hat{p}$ is given by $\frac{1}{n}\, p(1-p)$ (see (7.1.1)), increasing the number of replications by a factor of 100 (from 10^4 to 10^6) leads to an improvement in the accuracy (in terms of standard deviation) of the estimate of π by a factor of $\sqrt{100} = 10$. ■

Example 7.1.9. Suppose we are given A, B, C are independent Poisson random variables with parameters α, β and γ respectively. What is the probablity that the equation $Ax^2 - Bx + C = 0$ has a real solution? To answer this question one would have to calculate the probability that $B^2 - 4AC > 0$. That would imply evaluating

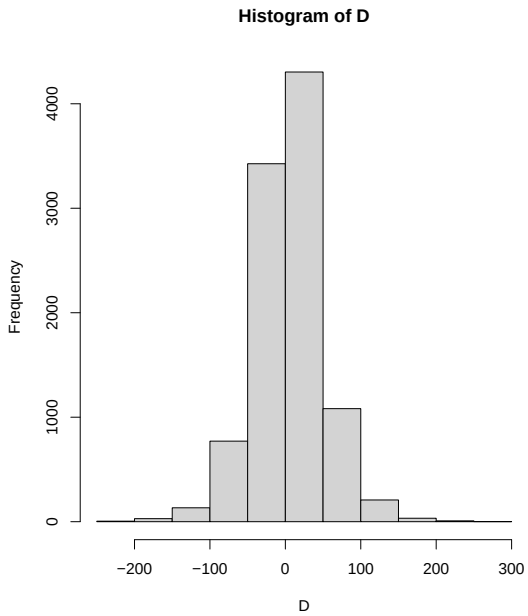
$$\sum_{a=0}^{\infty} \sum_{b=0}^{\infty} \sum_{c=0}^{\infty} \left(1(b^2 - 4ac \geq 0) \exp(-\alpha - \beta - \gamma) \frac{(\alpha)^a (\beta)^b (\gamma)^c}{a!\, b!\, c!} \right),$$

which would require some combinatorial effort. However we can use the strong law of large numbers and try to estimate the number via simulations.

```
B <- rpois(10000,6)
A <- rpois(10000,3)
C <- rpois(10000,3)
D <- B^2-4*A*C
mean(D >=0)
```

```
[1] 0.5906
```

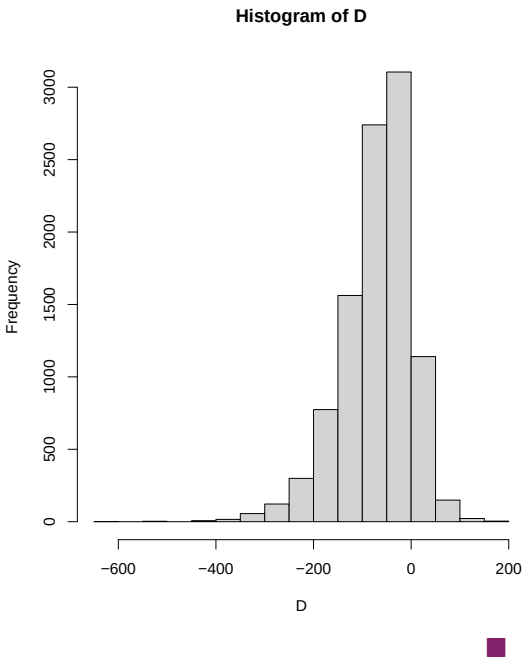
```
hist(D)
```




```
B <- rpois(10000,5)
A <- rpois(10000,5)
C <- rpois(10000,5)
D <- B^2-4*A*C
mean(D >=0)
```

```
[1] 0.1397
```

```
hist(D)
```



EXERCISES

Ex. 7.1.1. Consider the following data

13 40 23 15 21 4 44 16 32 14

(a) Compute the probability mass function of the empirical distribution from the data and also the corresponding ECDF.

(b) Using `ecdf` and `plot` command in R do part (a).

Ex. 7.1.2. Verify that the proofs of Theorem 3.3.5 and Theorem 3.3.6 hold for continuous random variables.

Ex. 7.1.3. Let X and Y be two continuous random variables having the same distribution. Let $f : \mathbb{R} \to \mathbb{R}$ be a piecewise continuous function. Then show that $f(X)$ and $f(Y)$ have the same distribution.

Ex. 7.1.4. Let X and Y be two discrete variables having the same distribution. Let $f : \mathbb{R} \to \mathbb{R}$ be a piecewise continuous function. Then show that $f(X)$ and $f(Y)$ have the same distribution.

Ex. 7.1.5. Let P be the empirical distribution defined by sample observations $X_1, X_2, \ldots, X_n$. In other words, P is the discrete distribution with probability mass function given in Definition 7.1.1. Let Y be a random variable with distribution P.

(a) Show that $E[Y] = \overline{X}$.

(b) Show that $Var[Y] = \frac{n-1}{n} S^2$.

Ex. 7.1.6. Suppose that U and V are independent Uniform$(0, 1)$, and we interpret (U, V) as coordinates of a point in $\mathbb{R}^2$.

1. Let $Z = \sqrt{U^2 + V^2}$. Find $p := P(Z < 1)$.
2. Can you modify the above `R-code` given in Example 7.1.8 to provide an estimate for π ?
3. Can you modify the above `R-code` to observe that the variance of the estimator of p goes to 0?

Ex. 7.1.7. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with finite expectation μ, finite variance σ^2, and finite $\gamma = E[X_1 - \mu]^4$. Compute $Var(S^2)$ in terms of μ, σ^2, and γ and show that $Var(S^2) \to 0$ as $n \to \infty$.

Ex. 7.1.8. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with finite expectation μ and finite variance σ^2. let $S = \sqrt{S^2}$, the non-negative root of the sample variance. The quantity S is called the "sample standard deviation". Although $E[S^2] = \sigma^2$, it is not true that $E[S] = \sigma$. In other words, S is not an unbiased estimator for σ. Follow the steps below to see why.

(a) Let Z be a random variable with finite mean and finite variance. Prove that $E[Z^2] \geq E[Z]^2$ and give an example to show that equality may not hold. (Hint: Consider how these quantities relate to the variance of Z).

(b) Use (a) to explain why $E[S] \leq \sigma$ and give an example to show that equality may not hold.

7.2 SIMULATION

The preceding discussion gives several mathematical statements about random samples, but it is difficult to develop any intuition about what these statements mean unless we look at actual data. Data is of course abundant in our world; however, the problem with real data is that we do not usually know for certain the random variable that generated it. To hone our intuition, it is therefore useful to be able to generate random samples from a distribution we specify. The process of doing so using a computer program is known as "simulation".

Simulation is not an easy task, because computers are by nature not random. Simulation is in fact not a random process at all; it is a completely deterministic process that tries to mimic

randomness. We will not go into how simulation is done, but simply use R to obtain simulated random samples.

R supports simulation from many distributions, including all the ones we have encountered. The general pattern of usage is that each distribution has a corresponding function that is called with the sample size an argument, and further arguments specifying parameters. The function returns the simulated observations as a vector. For example, 30 Binomial$(100, 0.75)$ samples can be generated by

```
rbinom(30, size = 100, prob = 0.75)
```

```
 [1] 63 72 77 75 82 73 69 78 68 76 87 67 75 73 68 64 71 74 65 79 72 79
[23] 76 72 70 74 72 69 74 72
```

We usually want to do more than just print simulated data, so we typically store the result in a variable and make further calculations with it; for example, compute the sample mean, or the sample proportion of cases where a particular event happens.

```
x <- rbinom(30, size = 100, prob = 0.75)
mean(x)
```

```
[1] 75.66667
```

```
sum(x >= 75) / length(x)
```

```
[1] 0.5666667
```

R has a useful function called `replicate` that allows us to repeat such an experiment several times.

```
replicate(15, {
    x <- rbinom(30, size = 100, prob = 0.75)
    mean(x)
})
```

```
 [1] 77.06667 76.60000 74.90000 73.66667 74.53333 76.40000 74.00000
 [8] 76.00000 74.33333 75.03333 74.76667 74.36667 75.16667 75.46667
[15] 75.26667
```

```
replicate(15, {
    x <- rbinom(30, size = 100, prob = 0.75)
    sum(x >= 75) / length(x)
})
```

```
 [1] 0.5666667 0.5333333 0.6000000 0.5333333 0.4333333 0.6000000
 [7] 0.5666667 0.5000000 0.7000000 0.5333333 0.6000000 0.4333333
[13] 0.7000000 0.6000000 0.3666667
```

This gives us an idea of the variability of the sample mean and sample proportion computed from a sample of size 30. We know of course that the sample mean has expectation $100 \times 0.75 = 75$, and we can use R to compute the expected value of the proportion as follows.

```
1 - pbinom(74, size = 100, prob = 0.75)
```

```
[1] 0.5534708
```

So the correponding estimates are close to the expected values, but with some variability. We expect the variability to go down if the sample size increases, say, from 30 to 3000.

```
replicate(15, {
    x <- rbinom(3000, size = 100, prob = 0.75)
    mean(x)
})
```

```
 [1] 75.01667 75.00100 74.90767 75.08067 75.01767 75.02867 74.97900
 [8] 74.98433 74.94667 75.08833 74.93900 74.94833 74.91500 74.96400
[15] 74.92700
```

```
replicate(15, {
    x <- rbinom(3000, size = 100, prob = 0.75)
    sum(x >= 75) / length(x)
})
```

```
 [1] 0.5516667 0.5403333 0.5583333 0.5520000 0.5520000 0.5423333
 [7] 0.5590000 0.5493333 0.5403333 0.5556667 0.5543333 0.5563333
[13] 0.5563333 0.5496667 0.5446667
```

Indeed we see that the estimates are much closer to their expected values now.

We can of course repeat this process for other events of interest, and indeed for many other distributions. We will see in the next section how we can simulate observations following the normal distribution using the funtion `rnorm`, and the exponential distribution using the funtion `rexp`. It is

also interesting to think about how one can simulate observations from a given distribution when a function to do so is not already available.

Recall from Lemma 5.3.7, that suppose $U \sim$ Uniform $(0,1)$ random variable and X is a continuous random variable such that its distribution function, F_X, is a strictly increasing continous function then $Y = F_X^{-1}(U)$ has the same distribution as X. This approach can be used to be simulate distributions (both discrete and continuous) from Uniform samples. The following examples explore this approach. We begin with an example on how to simulate Poisson samples from Uniform.

Example 7.2.1. When trying to formulate a method to simulate random variables from a new distribution, it is customary to assume that we already have a method to generate random variables from Uniform$(0,1)$. Let us see this can be used to generate random observations from a Poisson(λ) distribution using its probability mass function.

Let X denote an observation from the Poisson(λ) distribution, and $U \sim$ Uniform$(0,1)$. Denote $p_i = P(X = i)$. An algorithm to generate a random variable with the same distribution as X is suggested by the following observation.

$$\begin{aligned} & & p_0 &= P(U \le p_0), \\ P(U \le p_0 + p_1) = p_0 + p_1 & \Rightarrow & p_1 &= P(p_0 < U < p_0 + p_1), \\ P(U \le p_0 + p_1 + p_2) = p_0 + p_1 + p_2 & \Rightarrow & p_2 &= P(p_0 + p_1 < U < p_0 + p_1 + p_2), \end{aligned}$$

and so on. Thus, if we set Y to be 0 if $U \le p_0$, and k if U satisfies $\sum_{i=0}^{k-1} p_i < U < \sum_{i=0}^{k} p_i$, then Y has the same distribution as X. To use this idea to generate 50 observations from Poisson(5), we can use the following code in R, noting that $\sum_{i=0}^{k} p_i = P(X \le k)$.

```
replicate(50,
          {
              U <- runif(1)
              Y <- 0
              while (U > ppois(Y, lambda = 5)) Y <- Y + 1
              Y
          })
```

```
 [1]  5  3  3  5  5  3  4  4  4  6  6  8 10  2 10  9  4  5  5  9  2  9
[23]  2  6  4  5  7  3  4  6  8  3  4  2  4  6  4  2  3  0  4  5  5  5
[45]  4  9  7  6  3  4
```

Of course, there is nothing in this procedure that is specific to the Poisson distribution. By replacing the call to `ppois()` suitably, the same process can be used to simulate random observations from any discrete distribution supported on the non-negative integers. ■

The process described in the previous example cannot be used for continuous random variables. In such cases, Lemma 5.3.7 often proves useful. We illustrate how to generate samples from Exponential distribution using Uniform.

Example 7.2.2. Consider the case where we want X to have the Exp(1) distribution. Then, $F_X(x) = 1 - e^{-x}$ for $x > 0$. Solving for $F_X(x) = u$, we have

$$\begin{aligned} 1 - e^{-x} &= u \\ \Rightarrow e^{-x} &= 1 - u \\ \Rightarrow x &= -\log(1-u), \end{aligned}$$

that is, $F_X^{-1}(u) = -log(1-u)$. Thus, we can simulate 50 observations from the Exp(1) distribution using the following R code.

```
-log(1 - runif(50))
```

```
 [1] 0.233065598 0.280462639 0.154108577 1.372431939 1.821648686
 [6] 0.138321757 0.436222604 0.542218451 1.380248674 0.766713622
[11] 2.144642933 0.301378922 1.269994489 0.333248095 1.149509378
[16] 0.117884391 2.562483778 0.153560104 1.522516377 0.596372217
[21] 1.185683422 0.601846501 0.937989125 1.488177881 0.414706054
[26] 0.199426507 0.820649856 1.291911171 0.678224144 0.161773024
[31] 0.356192943 0.491640264 1.168934205 0.695930930 0.340648718
[36] 0.232449735 0.698323026 0.061983953 1.152038946 0.608277765
[41] 0.252472685 0.662322314 0.152055387 1.440552940 1.382534336
[46] 0.004607301 0.554657905 0.989217878 1.185984115 0.254887938
```

This takes advantage of the ability of `runif()` to generate multiple values at once, and the fact that the expression for $F_X^{-1}(u)$ can be easily vectorized. We can multiply the resulting observations by $1/\lambda$ to simulate observations from the Exp(λ) distribution. ■

The approach illustrated in the last two examples has a disadvantage when the distribution function F or its generalised inverse cannot be computed explicitly. This is the case when one wishes to simulate samples from standard Normal distribution. We will discuss a few instances in the Exercises next.

EXERCISES

Ex. 7.2.1. Let U_1 and U_2 be i.i.d Normal random variables.

(a) Let $\Theta = 2\pi U_1$. Find the distribution of Θ

(b) $R = \sqrt{-2\log(U_2)}$. Find the distribution of R.

(c) Let $X_1 = R\cos(\Theta)$ and $X_2 = R\sin(\Theta)$. Show that X_1, X_2 are i.i.d standard Normal random variables.

(d) Write a R code to simulate 100 samples from standard Normal distribution from Uniform$(0, 1)$.

Ex. 7.2.2. Let Z_1, Z_2 be i.i.d. standard normal random variables. $\mu_1, \mu_2 \in R, \sigma_1, \sigma_2 > 0$ and $-1 < \rho < 1$. Suppose $X_1 = \sigma_1 Z_1 + \mu_1$ and $X_2 = \sigma_2(\rho Z_1 + \sqrt{1-\rho^2} Z_2) + \mu_2$.

(a) Find the joint distribution X_1, X_2 *Hint: Proof of Theorem 6.4.4*

(b) Use the Exercise 7.2.1 (d) and write an R code to simulate 100 samples from a bivariate Normal disstribution where the correlation $\rho = \frac{1}{2}$ and marginals are standard Normal random variables.

Ex. 7.2.3. Use the approach in Example 7.2.1 write an R code to simulate from Uniform(0,1), 100 samples of

(a) Binomial $(10, \frac{1}{3})$

(b) Geometric$(\frac{1}{4})$

Ex. 7.2.4. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the Poisson(λ) distribution, and suppose we are interested in estimating λ.

(a) Show that both the sample mean and the sample variance of $X_1, X_2, \ldots, X_n$ are unbiased estimators of λ.

(b) Which of these estimators is better? To answer this question, simulate random observations from the Poisson(λ) distribution for various values of λ using the R function `rpois`. Explore the behaviour of the two estimates by varying λ as well as the sample size.

Ex. 7.2.5. Exercise 2.3.7 described the technique called "capture-recapture" which biologists use to estimate the size of the population of a species when it cannot be directly counted. Suppose the unknown population size is N, and fifty members of the species are selected and given an identifying mark. Sometime later a sample of size twenty is taken from the population, and it is found to contain X of the twenty previously marked. Equating the proportion of marked members in the second sample and the population, we have $\frac{X}{20} = \frac{50}{N}$, giving an estimate of $\hat{N} = \frac{1000}{X}$.

Recall that X has a hypergeometric distribution that involves N as a parameter. It is not easy to compute $E[\hat{N}]$ and $Var[\hat{N}]$. However, Hypergeometric random variables can be simulated in R using the function `rhyper`. For each $N = 50$, 100, 200, 300, 400, and 500, use this function to simulate 1000 values of $\hat{N}$ and use them to estimate $E[\hat{N}]$ and $Var[\hat{N}]$. Plot these estimates as a function of N.

Ex. 7.2.6. Suppose p is the unknown probability of an event A, and we estimate p by the sample proportion $\hat{p}$ based on an i.i.d. sample of size n.

(a) Write $Var[\hat{p}]$ and $SD[\hat{p}]$ as functions of n and p.

(b) Using the relations derived above, determine the sample size n, as a function of p, that is required to acheive $SD(\hat{p}) = 0.01$. How does this required value of n vary with p?

(c) Design and implement the following simulation study to verify this behaviour. For $p = 0.01$, 0.1, 0.25, 0.5, 0.75, 0.9, and 0.99,

(i) Simulate 1000 values of $\hat{p}$ with $n = 500$.

(ii) Simulate 1000 values of $\hat{p}$ with n chosen according to the formula derived above.

In each case, you can think of the 1000 values as i.i.d. samples from the distribution of $\hat{p}$, and use the sample standard deviation as an estimate of $SD[\hat{p}]$. Plot the estimated values of $SD(\hat{p})$ against p for both choices of n. Your plot should look similar to Figure 7.1.

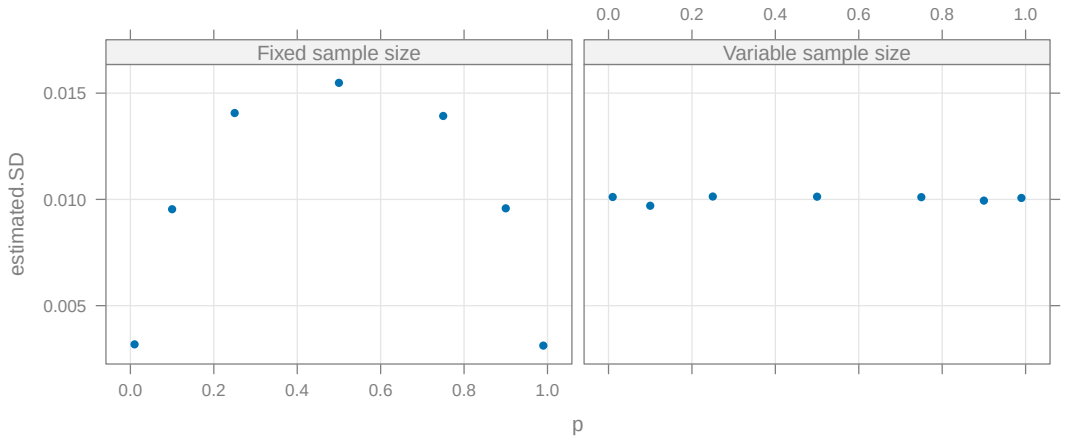


Figure 7.1: Estimated standard deviation in estimating a probability using sample proportion as a function of the probability being estimated. See exercise 7.2.6.

7.3 PLOTS

As we will see in later chapters, making more assumptions about the underlying distribution of X allows us to give concrete answers to many important questions. This is indeed a standard and effective approach to doing statistics, but in following that approach there is a danger of forgetting that assumptions have been made, which we should guard against by doing our best to convince ourselves beforehand that the assumptions we are making are reasonable.

Doing this is more of an art than a science, and usually takes the form of staring at plots obtained from the sample observations, with the hope of answering the question: "does this plot look like what I would have expected it to look like had my assumptions been valid?" Remember that the sample $X_1, X_2, \ldots, X_n$ is a random sample, so any plot derived from it is also a "random plot". Unlike simple quantities such as sample mean and sample variance, it is not clear what to "expect" such plots to look like, and the only way to really hone our instincts to spot anomalies is through experience. In this section, we introduce some commonly used plots and use simulated data to give examples of how such plots might look like when the usual assumptions we make are valid or invalid.

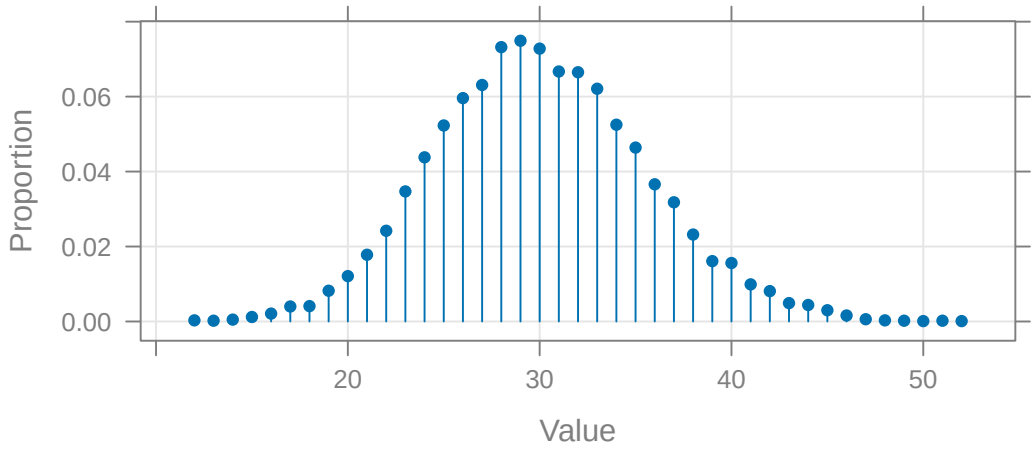


Figure 7.2: Empirical frequency distribution of 10000 random samples from the Poisson(30) distribution.

7.3.1 *Empirical Distribution Plot for Discrete Distributions*

The typical assumption made about a random sample is that the underlying random variable belongs to a *family* of distributions rather than a very specific one. For example, we may assume that the random variable has a Poisson(λ) distribution for some $\lambda > 0$, without placing any further restriction on λ, or a Binomial(n, p) distribution for some $0 < p < 1$. Such families are known as parametric families.

When the data $X_1, X_2, \ldots, X_n$ are from a discrete distribution, the simplest representation of the data is its empirical distribution, which is essentially a table of the frequencies of each value that appeared. For example, if we simulate 1000 samples from a Poisson distribution with mean 3, its frequency table may look like

```
x <- rpois(1000, lambda = 3)
table(x)
```

```
x
  0   1   2   3   4   5   6   7   8
 49 168 238 215 155  99  48  18  10
```

```
prop.table(table(x))
```

```
x
    0     1     2     3     4     5     6     7     8
0.049 0.168 0.238 0.215 0.155 0.099 0.048 0.018 0.010
```

The simplest graphical representation of such a table is through a plot similar to Figure 7.2, which represents a larger Poisson sample with mean 30, resulting in many more distinct values. Although in theory all non-negative integers have positive probability of occurring, the probabilites are too small to be relevant beyond a certain range. This plot does not have a standard name, although it may be considered a variant of the Cleveland Dot Plot. We will refer to it as the *Empirical Distribution Plot* from now on.

We can make similar plots for samples from Binomial or any other distribution. Unfortunately, looking at this plot does not necessarily tell us whether the underlying distribution is Poisson, in part because the shape of the Poisson distribution varies with the λ parameter. A little later, We will discuss a modification of the empirical distribution plot, known as a rootogram, that helps make this kind of comparison a little easier.

7.3.2 *Histograms for Continuous Distributions*

In the case of continuous distributions, we similarly want to make assumptions about a random sample being from a parametric family of distributions. For example, we may assume that the random variable has a Normal(μ, σ^2) distribution without placing any further restriction on the parameters μ or σ^2 (except of course that $\sigma^2 > 0$), or that it has an Exponential(λ) distribution with any value of the parameter $\lambda > 0$. Such families, as noted earlier, are known as parametric families. For both these examples, the *shape* of the distribution does not depend on the parameters, and this makes various diagnostic plots more useful.

The empirical distribution plot above is not useful for data from a continuous distribution, because by the very nature of continuous distributions, all the data points will be distinct with probability 1, and the value of the empirical distribution function will be exactly $1/n$ at these points.

The plot that is most commonly used instead to study distributions is the histogram. It is similar to the empirical distribution plot, except that it does not retain all the information contained in the empirical distribution. Instead, it divides the range of the data into arbitrary bins and counts the frequencies of data points falling into each bin, effectively discretizing the data. More precisely, the histogram estimates the probability density function of the underlying random variable by estimating the density in each bin as a quantity such that the probability of each bin is proportional to the number of observations in that bin. By choosing the bins judiciously, for example by having more of them as sample size increases, the histogram strikes a balance that ensures that the histogram "converges" to the true underlying density as $n \to \infty$.

Figure 7.3 gives examples of histograms where data are simulated from the normal and exponential distributions for varying sample sizes. Five replications are shown for each sample size. We can see that for large sample sizes, the shape of the histograms are recognizably similar to the shapes of the corresponding theoretical distributions seen in Figure 5.1 and Figure 5.2 in Chapter 5.

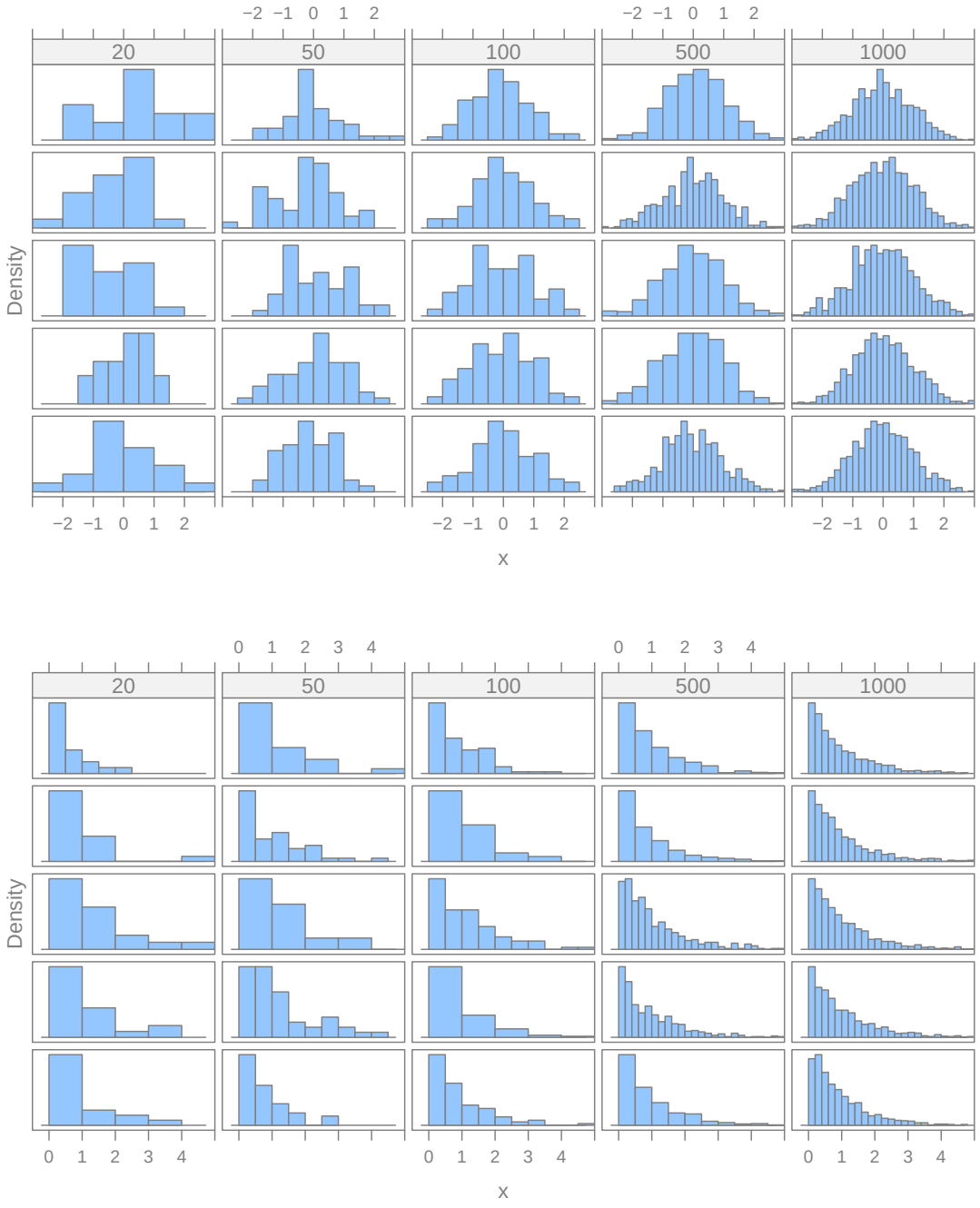


Figure 7.3: Histograms of random samples from the Normal$(0, 1)$ (top) and Exponential(1) (bottom) distributions. Columns represent increasing sample sizes, and rows are independent repetitions of the experiment.

Moreover, the shape is consistent over the five replications. This is not true, however, for small sample sizes. Remember that the histograms are based on the observed data, and are therefore random objects themselves. As we saw with numerical properties like the mean, estimates have higher variability when the sample size is small, and get less variable as sample size increases. The same holds for graphical estimates, although making this statement precise is more difficult.

There are several ways to create histograms in R, which we will not go into here, but one approach is explored in the exercises.

7.3.3 *Hanging Rootograms for Comparing with Theoretical Distributions*

Graphical displays of data are almost always used for some kind of comparison. Sometimes these are implicit comparisons, asking, say, "how many peaks does a density have", or "is it symmetric?" More often, they are used to compare samples from two subpopulations, say, the distribution of height in males and females. Sometimes, as discussed above, they are used to compare an observed sample to a hypothesized distribution.

In the case of the empirical distribution plot, a simple modification is to add the probability mass function of the theoretical distribution. This, although a reasonable modification, is not optimal. Research into human perception of graphical displays indicates that the human eye is more adept at detecting departures from straight lines than from curves. Taking this insight into account, John Tukey suggested "hanging" the vertical lines in an empirical distribution plot (which are after all nothing but sample proportions) from their expected values under the hypothesized distribution. He further suggested a transformation of what is plotted: instead of the sample proportions and the correponding expected probabilities, he suggested plotting their square roots, thus leading to the name *hanging rootogram* for the resulting plot. The reason for making this transformation is as follows. Recall that for a proportion $\hat{p}$ obtained from a sample of size n,

$$Var[\hat{p}] = \frac{p(1-p)}{n} \approx \frac{p}{n}$$

provided p is close to 0. In Chapter 8, we will encounter the Central Limit Theorem and the Delta Method, which can be used to show (see Example 8.5.4) that as the sample size n grows large, $Var[\sqrt{\hat{p}}] \approx c/n$ for a constant c. This means that unlike $\hat{p} - p$, the variance of $\sqrt{\hat{p}} - \sqrt{p}$ will be approximately independent of p.

Figure 7.4 gives examples of hanging rootograms. These examples have been created using the `rootogram()` function in the `latticeExtra` package. The following R code is an example of its use.

```
library(package = "latticeExtra")
xbin30 <- rbinom(10000, 100, 0.3)
rootogram(~ xbin30, dfun = function(x) dpois(x, lambda = 30), grid = TRUE)
```

This requires the `latticeExtra` package to be already installed on your system, which it most likely will not be. To install it, type

```
install.packages("latticeExtra")
```

and follow instructions.

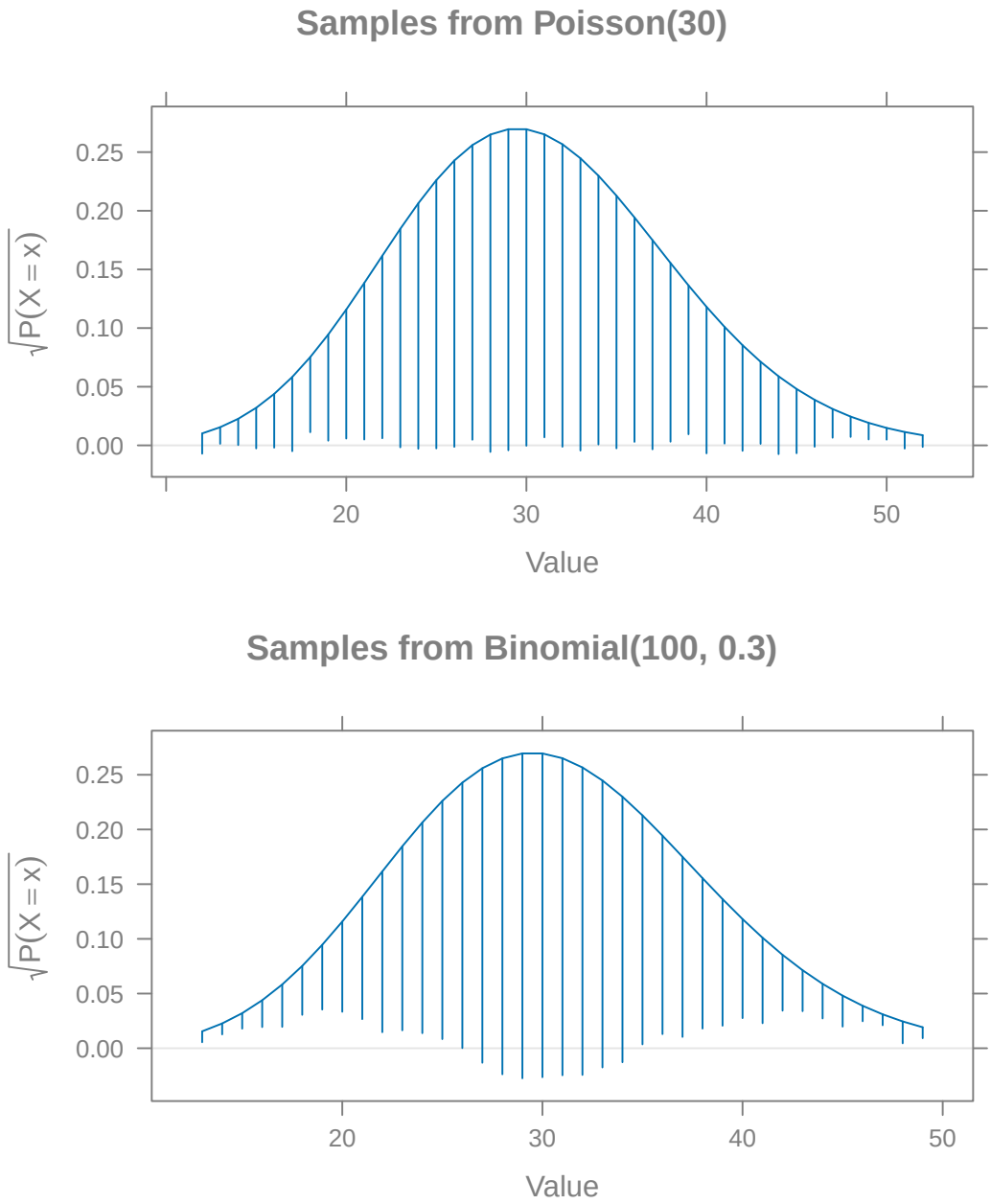


Figure 7.4: Hanging rootogram of 10000 random samples compared with the Poisson(30) distribution. In the top plot, the samples are also from Poisson(30), whereas in the bottom plot the samples are from the Binomial(100, 0.3) distribution, which has the same mean but different variance. Note the similarities with Figure 2.2

7.3.4 *Q-Q Plots for Continuous Distributions*

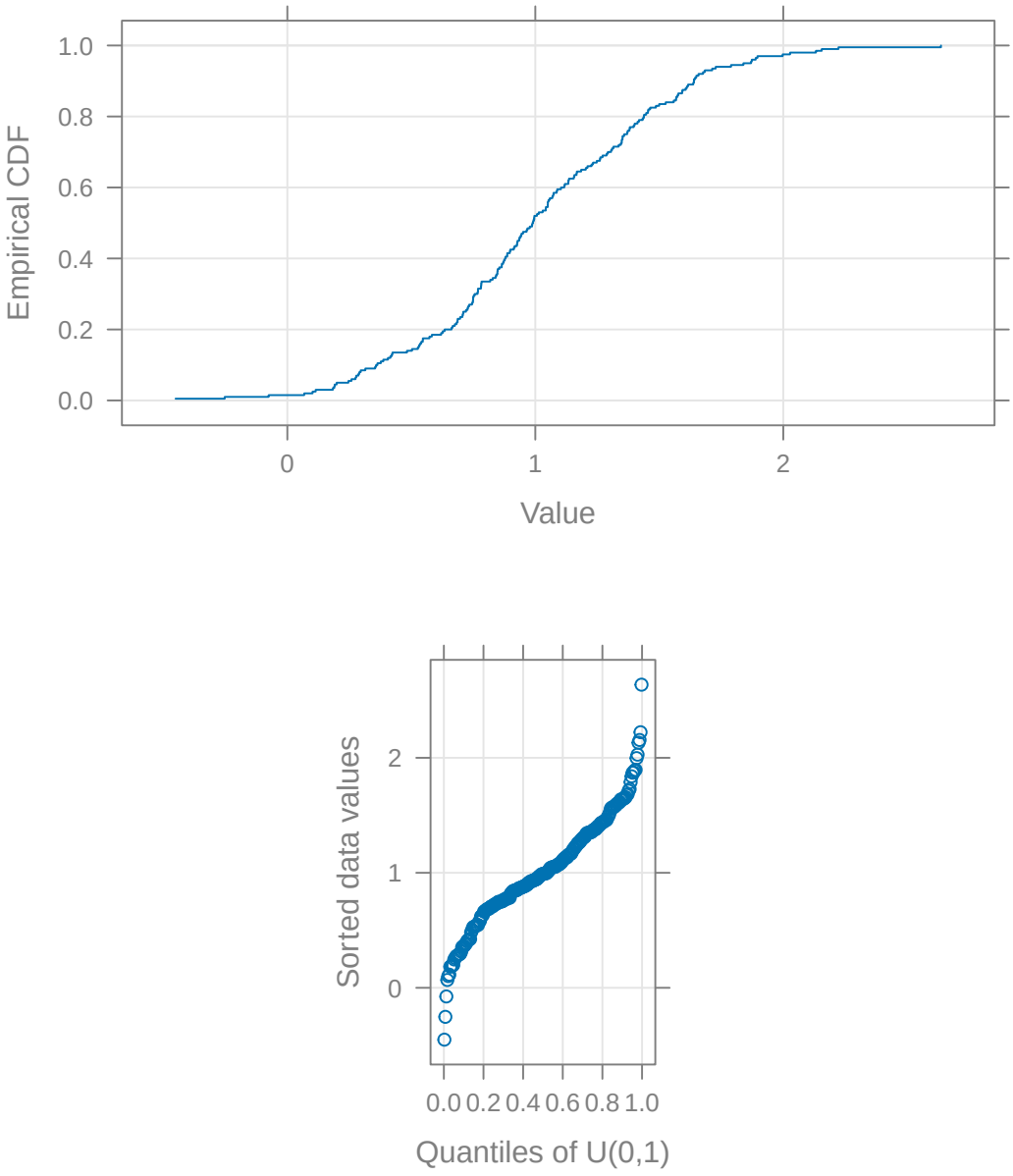


Figure 7.5: Conventional ECDF plot (top) and its "inverted" version (bottom), with x- and y-axes switched, and points instead of lines.

Just as histograms were binned versions of the empirical distribution plot, we can plot binned versions of hanging rootograms for data from a continuous distribution as well. It is more common

however, to look at quantile-quantile plots (QQ plots), which do not bin the data, but instead plot what is essentially a transformation of the empirical CDF.

Recall from Definition 7.1.1 that the ECDF of observations $X_1, X_2, \ldots, X_n$ is given by

$$\hat{F}_n(t) = P(Y \leq t) = \frac{\#\{X_i \leq t\}}{n}$$

The top plot in Figure 7.5 is a conventional ECDF plot of 200 observations simulated from a Normal$(1, 0.5^2)$ distribution. The bottom plot has the sorted data values on the y-axis and 200 equally spaced numbers from 0 to 1. A little thought tells us that this plot is essentially the same as the ECDF plot, with the x- and y-axes switched, and using points instead of lines. Naturally, we expect that for reasonably large sample sizes, the ECDF plot obtained from a random sample will be close to the true cumulative distribution function of the underlying distribution. If we know the shape of the distribution we expect the data to be from, we can compare it with the shape seen in the plot.

Although this is a fine idea in principle, it is difficult in practice to detect small differences between the observed shape and the theorized or expected shape. Here, we are helped again by the insight that the human eye finds it easier to detect deviations from a straight line than from curves. By keeping the sorted data values unchanged, but transforming the equally spaced probability values to the corresponding quantile values of the theorized distribution, we obtain a plot that we expect to be linear. We define them formally below.

Definition 7.3.1. *Let F be a distribution function. For $0 < p < 1$, the p-th quantile of F is defined as*

$$q_p = \inf\{x \in \mathbb{R} : F(x) \geq p\}.$$

Note that $q_p = F^{-1}(p)$ when F^{-1} exists and $F(q_p-) \leq p \leq F(q_p)$. $q_{\frac{1}{2}}$ is referred to as the median of F. For a sample $X_1, X_2, \ldots, X_n$ from distribution F, the sample p-th quantile is defined as the p-th quantile of the empirical distribution function F_n.

Quantiles may be thought of informally as follows, generalizing the definition of median given in Exercise 5.2.10: For a given CDF F, the quantile corresponding to a probability value $p \in [0, 1]$ is a value x such that $F(x) = p$. Such an x may not exist for all p and F, or it may not be unique, and a formal definition of quantiles needs to be modified to take this into account.However, for most standard continuous distributions used in Q-Q plots, one may work with this informal notion. Such a plot with Normal $(0, 1)$ quantiles is shown in Figure 7.6 for simulated normal and exponential random samples. More examples are explored in the exercises.

EXERCISES

Ex. 7.3.1. The R functions `histogram()` and `qqmath()` in the `lattice` package can be used to generate histograms and Q-Q plots respectively (although there are other alternatives as well).

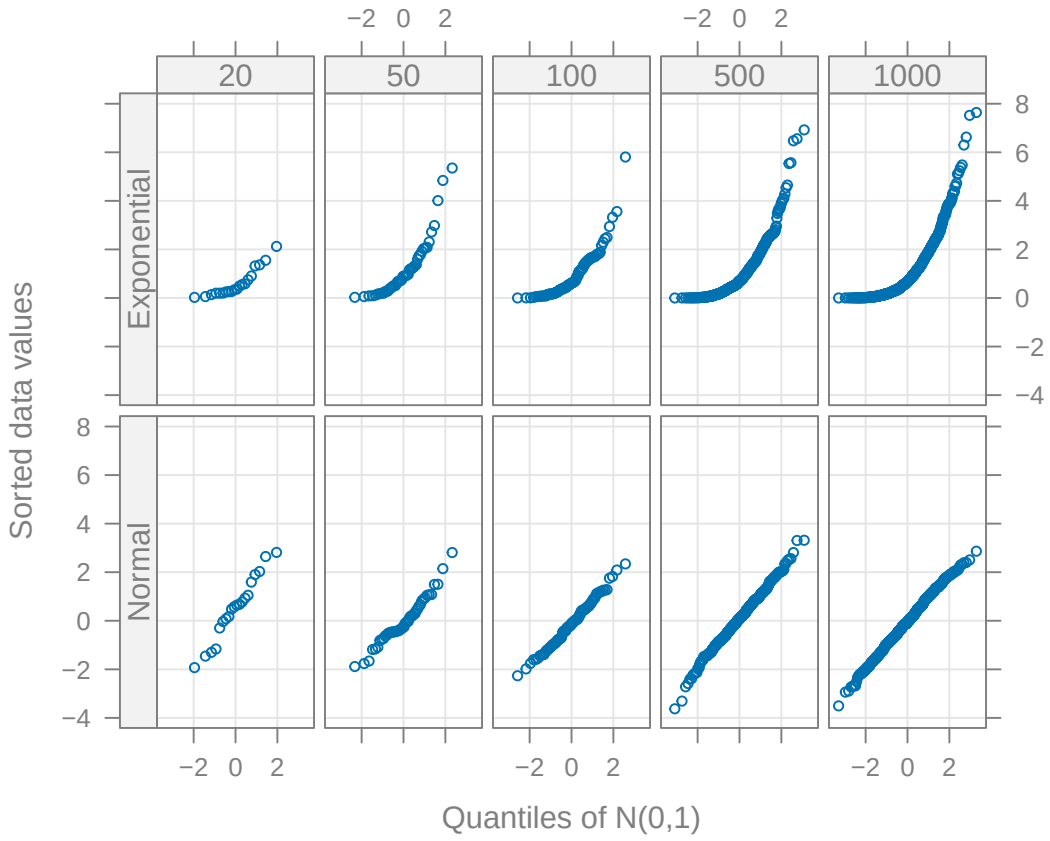


Figure 7.6: Normal Q-Q plots of data generated from Normal and Exponential distributions, with varying sample size. The Q-Q plots are more or less linear for Normal data, but exhibit curvature indicative of a relatively heavy right tail for exponential data. Not surprisingly, the difference becomes easier to see as the sample size increases.

This exercise guides you through the process of simulating data from a sampling distribution and creating corresponding histograms and Normal Q-Q plots.

(a) Suppose $Z_1, Z_2, \ldots, Z_n$ are independent Normal$(0, 1)$. Then the distribution of the mean of $Z_1, Z_2, \ldots, Z_n$ is Normal$(0, 1/n)$. To verify this, simulate such means for $n = 50$ using the following R code.

```
x.means <- replicate(1000, mean(rnorm(50)))
```

(b) Create a histogram of the simulated values using

```
library(package = "lattice") # needed only once, to load the package
histogram(~ x.means, nint = 15) # nint (optional) gives the number of bins
```

(c) Create a Normal Q-Q plot of the same values using

```
qqmath(~ x.means, grid = TRUE)
```

(d) Study the behaviour of these plots over multiple repetitions, as well as by varying n and the number of replications.

Ex. 7.3.2. If $Z_1, Z_2, \ldots, Z_n$ are independent Normal$(0, 1)$, what can you say about the distribution of the median of $Z_1, Z_2, \ldots, Z_n$? Use the `median()` function, using it to replace the call to `mean()` in the previous exercise, to simulate observations from this distribution. Use histograms and Normal Q-Q plots to study this distribution and compare it to the distribution of the mean. In particular, is the distribution of the median also Normal? Does it have lower or higher variance than the mean?

Ex. 7.3.3. Repeat the previous exercise, replacing the median by the minimum and maximum of n obsrvations $Z_1, Z_2, \ldots, Z_n$ that are independent Normal$(0, 1)$. What are the distingushing features of these histograms and Normal Q-Q plots?

8

SAMPLING DISTRIBUTIONS AND LIMIT THEOREMS

For $n \geq 1$, let $X_1, X_2, \ldots, X_n$ be an i.i.d. random sample from a population. Recall the sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

We have seen the use of these sample statistics in the previous chapter. In this chapter, we will discuss the distributional properties and limiting behaviour of such statistics. In Chapters 9 and 10, we will discuss how these properties can be effectively used to estimate parameters related to the underlying population and verify specific hypotheses about them. The corresponding fields of study are called Estimation and Hypothesis Testing.

We will spend most of our time in finding the distribution of the sample mean and the sample variance given the distribution of X_1. One immediately observes that these are somewhat complicated functions of independent random variables. However in Section 3.3 and Section 5.5 we have seen examples of functions for which we were able to explicitly compute the distribution. To understand sampling statistics we must also understand the notion of joint distribution of more than two continuous random variables (See Section 3.3 for discrete random variables).

8.1 MULTI-DIMENSIONAL CONTINUOUS RANDOM VARIABLES

In Chapter 3, while discussing discrete random variables, we had considered a finite collection of random variables $(X_1, X_2, \ldots, X_n)$. In Definition 3.2.7, we had described how to define their joint distribution and we used this to understand the multinomial distribution in Example 3.2.12. There are many instances in the continuous setting as well where it is relevant to study the joint distribution of a finite collection of random variables. Suppose X is a point chosen randomly inside the unit sphere in three dimensions. Then X has three coordinates, say $X = (X_1, X_2, X_3)$, where each X_i is a random variable in $(0, 1)$. These coordinates are dependent because we know that $\sqrt{X_1^2 + X_2^2 + X_3^2} \leq 1$. To reason about the properties of X, it is useful and necessary to understand the "joint distribution" of (X_1, X_2, X_3). Similarly, to understand the distribution of the sample mean and the sample variance, which are functions of $X_1, X_2, \ldots, X_n$, we first need to understand the joint distribution of $(X_1, X_2, \ldots, X_n)$. We begin by defining the joint distribution function.

Definition 8.1.1. *For $n \geq 1$, let $X_1, X_2, \ldots, X_n$ be random variables defined on the same probability space. The joint distribution function $F : \mathbb{R}^n \to [0,1]$ of $X_1, X_2, \ldots, X_n$ is given by*

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n), \tag{8.1.1}$$

for $x_1, x_2, \ldots, x_n \in \mathbb{R}$.

As in one-variable and two-variable situations, the joint distribution function determines the entire joint distribution of $(X_1, X_2, \ldots, X_n)$ for discrete random variables. More precisely, if all the random variables were discrete with $X_i : S \to T_i$, where T_i are countable subsets of $\mathbb{R}$ for $1 \leq i \leq n$, the from the joint distribution function one can determine

$$P(X_1 = t_1, X_2 = t_2, \ldots, X_n = t_n)$$

for all $t_i \in T_i, 1 \leq i \leq n$ (See Exercise 8.1.1). The joint distribution function determines the joint distribution in the continuous setting as well, but we need to introduce some notation before we can state this result rigorously.

For $n \geq 1$, let $f : \mathbb{R}^n \to \mathbb{R}$ be a non-negative function that is piecewise-continuous in each variable, and for which

$$\int_{\mathbb{R}^n} f(x_1, x_2, \ldots, x_n)\, \mathrm{d}x_1\, \mathrm{d}x_2 \ldots \mathrm{d}x_n = 1.$$

If for every Borel set $A \subset \mathbb{R}^n$ we have

$$P(A) = \int_A f(x_1, x_2, \ldots, x_n)\, \mathrm{d}x_1\, \mathrm{d}x_2 \ldots \mathrm{d}x_n,$$

then one can show as in Theorem 5.1.5 that P is a probability on $\mathbb{R}^n$. In this case, f is called the density function for P.

Density functions arise naturally from certain types of random variables. A collection of random variables $(X_1, X_2, \ldots, X_n)$ is said to have a joint density $f : \mathbb{R}^n \to \mathbb{R}$ if for every event $A \subset \mathbb{R}^n$,

$$P((X_1, X_2, \ldots, X_n) \in A) = \int_A f(x_1, x_2, \ldots, x_n)\, \mathrm{d}x_1\, \mathrm{d}x_2 \ldots \mathrm{d}x_n.$$

In this setting, the joint distribution of $(X_1, X_2, \ldots, X_n)$ is determined by their joint density f. Using multivariable calculus we can can state and prove a result similar to Theorem 5.2.5 for random variables $(X_1, X_2, \ldots, X_n)$ that have a joint density. In particular, we can conclude that as the joint densities are assumed to be piecewise continuous in each variable, the corresponding distribution functions are piecewise differentiable in each variable. Further, the joint distribution of the continuous random variables $(X_1, X_2, \ldots, X_n)$ are completely determined by their joint distribution function F. That is, if we know $F(x_1, x_2, \ldots, x_n)$ for all $x_1, x_2, \ldots, x_n \in \mathbb{R}$ we could use multivariable calculus to differentiate F to find f. Integrating this joint density over the event A, we can then calculate $P((X_1, X_2, \ldots, X_n) \in A)$.

As in the $n = 2$ case, one can recover the marginal density of each X_i for i between 1 and n by integrating over the other indices. So, the marginal density of X_i at a is given by

$$f_{X_i}(a) = \int_{\mathbb{R}^{n-1}} f(x_1, \ldots, x_{i-1}, a, x_{i+1}, \ldots, x_n)\, \mathrm{d}x_1 \ldots \mathrm{d}x_{i-1}\, \mathrm{d}x_{i+1} \ldots \mathrm{d}x_n.$$

Further, for $n \geq 3$, we can deduce the joint density for any sub-collection $m \leq n$ random variables by integrating over the other variables. For instance, if we were interested in the joint density of (X_1, X_3, X_7), we would obtain

$$f_{X_1,X_3,X_7}(a_1, a_3, a_7) = \int_{\mathbb{R}^{n-3}} f(a_1, x_2, a_3, x_4, x_5, x_6, a_7, x_8 \ldots, x_n)\, \mathrm{d}x_2\, \mathrm{d}x_4\, \mathrm{d}x_5\, \mathrm{d}x_6 \ldots \mathrm{d}x_n.$$

Suppose $X_1, X_2, \ldots, X_n$ are random variables defined on a single sample space S with joint density $f : \mathbb{R}^n \to \mathbb{R}$. Let $g : \mathbb{R}^n \to \mathbb{R}$ be a function of n variables for which $g(X_1, X_2, \ldots, X_n)$ is defined on the range of the X_j variables. Let B be an event in the range of g. Then, following the proof of Theorem 3.3.5, we can show that

$$P\left(g(X_1, X_2, \ldots, X_n) \in B\right) = P\left((X_1, X_2, \ldots, X_n) \in g^{-1}(B)\right).$$

Although the above provides an abstract method of finding the distribution of the random variable $Y = g(X_1, X_2, \ldots, X_n)$, it can be difficult to use for explicit calculations. For $n = 1$ we discussed this question in detail in Section 5.3, and for $n = 2$ we explored how to find the distributions of sums and ratios of independent random variables (see Section 5.5). This method could be extended by induction on n in a few cases, but in general this is not possible. In Appendix B, Section A.1.1, we discuss a more general Jacobian-based method of finding the joint density of functions of random variable.

The notion of independence, introduced in the discrete setting, also extends to multi-dimensional continuous random variables. As discussed in Definition 3.2.3, a finite collection of continuous random variables $X_1, X_2, \ldots, X_n$ is mutually independent if the sets $(X_j \in A_j)$ are mutually independent for all events A_j in the ranges of the corresponding X_j. As proved for the $n = 2$ case in Theorem 5.4.7, we can similarly deduce that if $(X_1, X_2, \ldots, X_n)$ are mutually independent continuous random variables with marginal densities f_{X_i} then their joint density is given by

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i), \tag{8.1.2}$$

for $x_i \in \mathbb{R}$ and $1 \leq i \leq n$. Further, for any finite sub-collection $(X_{i_1}, X_{i_2}, \ldots, X_{i_m})$ of the above independent random variables, the joint density is given by

$$f(a_1, a_2, \ldots, a_m) = \prod_{j=1}^{m} f_{X_{i_j}}(a_j). \tag{8.1.3}$$

We conclude this section with a result that we will repeatedly use.

Theorem 8.1.2. *Fix $n \geq 1$. For each $j \in \{1, 2, \ldots, n\}$, let $i \in \{1, 2, \ldots, m_j\}$ for some positive integer m_j. Suppose $X_{i,j}$ is an array of mutually independent continuous random variables. Define $Y_j = g_j(X_{1,j}, X_{2,j}, \ldots X_{m_j,j})$, where $g_j : \mathbb{R}^{m_j} \to \mathbb{R}$ are continuous functions. Then the resulting variables $Y_1, Y_2, \ldots, Y_n$ are mutually independent.*

Proof. Follows by the same proof presented in Theorem 3.3.6. ∎

8.1.1 *Order Statistics and their Distributions*

For $n \geq 1$, let $X_1, X_2, \ldots, X_n$ be an i.i.d. random sample from a population with common distribution function F. Arrange them in increasing order of magnitude, with the ordered observations denoted by

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}.$$

These ordered values are called the order statistics of the sample $X_1, X_2, \ldots, X_n$. For, $1 \leq r \leq n$, $X_{(r)}$ is called the r-th order statistic. The median of $X_1, X_2, \ldots, X_n$ is defined as $X_{\left(\frac{n+1}{2}\right)}$ when n is odd and $X_{\left(\frac{n}{2}\right)}$ when n is even.

One can compute $F_{(r)}$, the distribution function of $X_{(r)}$, for $1 \leq r \leq n$ in terms of n and F. We have,

$$\begin{aligned}
F_{(1)}(x) &= P(X_{(1)} \leq x) = 1 - P\left(X_{(1)} > x\right) = 1 - P\left(\bigcap_{i=1}^{n}(X_i > x)\right) \\
&= 1 - \prod_{i=1}^{n} P(X_i > x) = 1 - \prod_{i=1}^{n}(1 - P(X_i \leq x)) \\
&= 1 - (1 - F(x))^n, \\
F_{(n)}(x) &= P(X_{(n)} \leq x) = P\left(\bigcap_{i=1}^{n}(X_i \leq x)\right) = \prod_{i=1}^{n} P(X_i \leq x) = (F(x))^n,
\end{aligned}$$

and for $1 < r < n$,

$$\begin{aligned}
F_{(r)}(x) &= P(X_{(r)} \leq x) = P(\text{at least } r \text{ elements from the sample are} \leq x) \\
&= \sum_{j=r}^{n} P(\text{exactly } j \text{ elements from the sample are} \leq x) \\
&= \sum_{j=r}^{n} \binom{n}{j} P(\text{chosen } j \text{ elements from the sample are} \leq x) \times \\
&\qquad P((n-j) \text{ elements not chosen from the sample are } > x) \\
&= \sum_{j=r}^{n} \binom{n}{j} F(x)^j (1 - F(x))^{n-j}
\end{aligned}$$

If the distribution function F had a probability density function f then each $X_{(r)}$ has a probability density function $f_{(r)}$. This can be obtained by differentiating $F_{(r)}$ and is given by the following expression.

$$f_{(r)}(x) = \begin{cases} n(1-F(x))^{n-1}f(x) & r = 1 \\ nf(x)(F(x))^{n-1} & r = n \\ \frac{n!}{(r-1)!(n-r)!}f(x)(F(x))^{r-1}(1-F(x))^{n-r} & 1 < r < n \end{cases} \tag{8.1.4}$$

Example 8.1.3. Let $n \geq 1$ and let $X_1, X_2, \ldots, X_n$ be a i.i.d. random sample from a population whose common distribution F is an Exponential (λ) random variable. Then we know that

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0. \end{cases}$$

Therefore using (8.1.4) and substituting for F as above we have that the densities of the order statistics are given by

$$f_{(r)}(x) = \begin{cases} n(e^{-\lambda x}))^{n-1}\lambda e^{-\lambda x} & r = 1 \\ n\lambda e^{-\lambda x}(1-e^{-\lambda x})^{n-1} & r = n \\ \lambda e^{-\lambda x}\frac{n!}{(r-1)!(n-r)!}(1-e^{-\lambda x})^{r-1}(e^{-\lambda x})^{n-r} & 1 < r < n, \end{cases}$$

for $x > 0$. Simplifying the algebra we obtain,

$$f_{(r)}(x) = \begin{cases} n\lambda e^{-n\lambda x} & r = 1 \\ n\lambda e^{-\lambda x}(1-e^{-\lambda x})^{n-1} & r = n \\ \frac{\lambda n!}{(r-1)!(n-r)!}(1-e^{-\lambda x})^{r-1}(e^{-\lambda x})^{n-r+1} & 1 < r < n, \end{cases}$$

for $x > 0$. We note from the above that $X_{(1)}$, i.e minimum of exponentials, is Exponential $(n\lambda)$ random variable. However the other order statistics are not exponentially distributed. ■

In many applications, one is interested in the range of values a random variable X assumes. A method to understand this to sample $X_1, X_2, \ldots, X_n$ i.i.d. X and examine $R = X_{(n)} - X_{(1)}$. Suppose X has a probability density function $f : \mathbb{R} \to \mathbb{R}$ and distribution function $F : \mathbb{R} \to [0, 1]$. As before we can can calculate the joint density of $X_{(1)}, X_{(n)}$ by first computing the joint distribution

function. This is done by using the i.i.d. nature of the sample and the definition of the order statistics.

$$\begin{aligned} P(X_{(1)} \le x, X_{(n)} \le y) &= P(X_{(n)} \le y) - P(x < X_{(1)}, X_{(n)} \le y) \\ &= P\left(\bigcap_{i=1}^{n} \{X_i \le y\}\right) - P\left(\bigcap_{i=1}^{n} \{x < X_i \le y\}\right) \\ &= [P(X \le y)]^n - [P(x < X \le y)]^n \\ &= \begin{cases} [F(x)]^n - [F(y) - F(x)]^n & x < y \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

From the above, differentiating partially in x and y we see that the joint density of $(X_{(1)}, X_{(n)})$ is given by

$$f_{X_{(1)},X_{(n)}}(x,y) = \begin{cases} n(f(x) - f(y))[F(y) - F(x)]^{n-1} & x < y \\ 0 & \text{otherwise.} \end{cases} \tag{8.1.5}$$

To calculate the distribution of R, we compute its distribution function. For $r \le 0$, $P(R \le r) = 0$ and for $r > 0$, using the above joint density of $(X_{(1)}, X_{(n)})$ we have

$$\begin{aligned} P(R \le r) &= P(X_{(n)} \le X_{(1)} + r) \\ &= \int_{-\infty}^{\infty} \left[\int_0^r f_{X_{(1)},X_{(n)}}(x, z+x) dz \right] \mathrm{d}x \\ &= \int_0^r \left[\int_{-\infty}^{\infty} f_{X_{(1)},X_{(n)}}(x, z+x)\, \mathrm{d}x \right] dz, \end{aligned}$$

where we have done a change of variable $y = z + x$ in the second last line and a change in the order of integration in the last line. Differentiating the above we conclude that R has a joint density given by

$$f_R(r) = \begin{cases} \int_{-\infty}^{\infty} f_{X_{(1)},X_{(n)}}(x, r+x)\, \mathrm{d}x & \text{if } r > 0 \\ \\ 0 & \text{otherwise.} \end{cases} \tag{8.1.6}$$

Example 8.1.4. Let $X_1, X_2, \ldots, X_n$ be i.i.d. Uniform$(0,1)$. The probability density function and distribution function of a Uniform$(0,1)$ random variable are given by

$$f(x) = \begin{cases} 1 & \text{if } x \in (0,1) \\ 0 & \text{otherwise.} \end{cases} \qquad \text{and} \qquad F(x) = \begin{cases} 0 & \text{if } x \le 0 \\ x & \text{if } 0 < x < 1 \\ 1 & \text{if } x > 1. \end{cases}$$

Let $f_{X_{(r)}}$ be the probability density function of $X_{(r)}$ for $1 \leq r \leq n$. Then, using (8.1.4), we have

$$\begin{aligned} f_{X_{(1)}}(x) &= \begin{cases} n(1-x)^{n-1} & \text{if } x \in (0,1) \\ 0 & \text{otherwise,} \end{cases} \\ f_{X_{(n)}}(x) &= \begin{cases} nx^{n-1} & \text{if } x \in (0,1) \\ 0 & \text{otherwise, and} \end{cases} \\ \text{for } 1 < r < n, \quad f_{X_{(r)}}(x) &= \begin{cases} \frac{n!}{(r-1)!(n-r)!} x^{r-1}(1-x)^{n-r} & \text{if } x \in (0,1) \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Using (8.1.5), the joint density of $(X_{(1)}, X_{(n)})$ is given by

$$f_{X_{(1)},X_{(n)}}(x,y) = \begin{cases} n(n-1)(y-x)^{n-1} & \text{if } 0 \leq x \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Using (8.1.6), the probability density function of the range $R = X_{(n)} - X_{(1)}$ is given by

$$\begin{aligned} f_R(r) &= \begin{cases} \int_0^{1-r} n(n-1)(x+r-x)^{n-1}\,\mathrm{d}x & \text{if } 0 < r < 1 \\ 0 & \text{otherwise,} \end{cases} \\ &= \begin{cases} n(n-1)r^{n-1}(1-r) & \text{if } 0 < r < 1 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

It is easy to see by comparing density functions that $X_{(r)} \sim \text{Beta}(r, n-r+1)$ for $1 \leq r \leq n$, and the range $R \sim \text{Beta}(n, 2)$. ■

In general, we may also be interested in the joint distribution of the order statistics. Suppose we have an i.i.d. sample $X_1, X_2, \ldots, X_n$ having distribution X. If X has a probability density function $f : \mathbb{R} \to \mathbb{R}$ then one can show that the order statistic $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ has a joint density $h : \mathbb{R}^n \to \mathbb{R}$ given by

$$h(u_1, u_2, \ldots, u_n) = \begin{cases} n! f(u_1) f(u_2) \ldots f(u_n) & u_1 < u_2 < \ldots < u_n, \\ 0 & \text{otherwise.} \end{cases}$$

The above fact should be intuitively clear: Any ordering $u_1 < u_2 < \ldots < u_n$ has "probability" $f(u_1)f(u_2)\ldots f(u_n)$. Each X_i can assume any of the u_k's. The total number of possible orderings is $n!$. A formal proof involves using the Jacobian method and will be discussed in Appendix B.

8.1.2 χ^2, F and t

χ^2 (pronounced Chi-Square), F and t distributions arise naturally when considering functions of i.i.d. normal random variables $(X_1, X_2, \ldots, X_n)$ for $n \geq 1$. They are useful in estimation and hypothesis testing, which we will study in subsequent chapters. We discuss these distributions via three examples.

The χ^2, F and t distributions arise as functions of Normal random variables. As we will see, they are essentially special cases of distributions we have already encountered, but they are studied separately because they come up naturally when considering the distribution of sample variances obtained from collections of i.i.d. normal random variables. In this section, we discuss these distributions via three examples, before discussing their connection to the sample variance in Section 8.1.3.

Example 8.1.5. For $n \geq 1$, let $(X_1, X_2, \ldots, X_n)$ be a collection of independent Normal random variables with mean 0 and variance 1. Then their joint density is given by

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i) = \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2},$$

for $x_i \in \mathbb{R}$ and $1 \leq i \leq n$. We are interested in the distribution of $Z = \sum_{i=1}^{n} X_i^2$.

We shall find this distribution in two steps. Clearly, the range of X_1^2 is non-negative. The distribution function for X_1^2 at $z \geq 0$ is given by

$$\begin{aligned} F_1(z) &= P(X_1^2 \leq z) \\ &= P(X_1 \leq \sqrt{z}) \\ &= \int_0^{\sqrt{z}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, \mathrm{d}x \\ &= \int_0^{z} \frac{1}{2\sqrt{2\pi}} e^{-\frac{u}{2}} u^{-\frac{1}{2}} \, \mathrm{d}u \end{aligned}$$

Comparing it with the Gamma (α, λ) random variable defined in Definition 5.5.5 and using Exercise 5.5.9, we see that X_1^2 is distributed as a Gamma $(\frac{1}{2}, \frac{1}{2})$ random variable. From the calculation done in Example 5.5.6 for $n = 2$, it follows by using induction that $Z = \sum_{i=1}^{n} X_i^2$ has the Gamma $\left(\frac{n}{2}, \frac{1}{2}\right)$ distribution. This distribution is referred to as χ^2 with n degrees of freedom. We define it precisely next. ■

Definition 8.1.6. *(χ^2_n (i.e. chi-square with n degrees of freedom)) A random variable X whose distribution is Gamma $\left(\frac{n}{2}, \frac{1}{2}\right)$ is said to have the chi-square distribution with n degrees of freedom, denoted $X \sim \chi^2_n$. The density of X is given by*

$$f(x) = \frac{2^{-\frac{n}{2}}}{\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} = \begin{cases} \frac{2^{-\frac{n}{2}}}{(\frac{n}{2}-1)!} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \text{when } n \text{ is even} \\ \frac{2^{n-\frac{n}{2}-1}(\frac{n-1}{2})!}{(n-1)!\sqrt{\pi}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \text{when } n \text{ is odd} \end{cases}$$

for $x > 0$.

We show in Section 8.1.3 that the sample variance obtained from a Normal sample follows a (scaled) χ^2 random variable. The F distribution arises as the ratio of the sample variances of two independent Normal samples, or in other words, as the ratio of two independent (scaled) χ^2 random variables, as we see in the next example.

Example 8.1.7. (F distribution) Let $X_1, X_2, \ldots, X_{n_1}$ be an i.i.d. random sample from the Normal $\left(0, \sigma_1^2\right)$ population, and $Y_1, Y_2, \ldots, Y_{n_2}$ be an *independent* i.i.d. random sample from a Normal $\left(0, \sigma_2^2\right)$ population. It follows from Example 8.1.5 that $U = \sum_{i=1}^{n_1} \left(\frac{X_i}{\sigma_1}\right)^2$ has the $\chi^2_{n_1}$ distribution, and $V = \sum_{i=1}^{n_2} \left(\frac{Y_i}{\sigma_2}\right)^2$ has the $\chi^2_{n_2}$ distribution. Further, by Theorem 8.1.2 U and V are independent because the X_i and Y_j random variables are independent. Let $Z = \frac{U}{n_1} / \frac{V}{n_2} = \frac{U}{V} \cdot \frac{n_2}{n_1}$. It follows from Example 5.5.10 that the density of $W = \frac{U}{V}$ for $w > 0$ is given by

$$f_W(w) = \frac{w^{\frac{n_1}{2}-1}}{(1+w)^{\frac{n_1+n_2}{2}}} \frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})}$$

Therefore, for $z > 0$,

$$F_Z(z) = P(Z \le z) = P\left(W \le \frac{n_1}{n_2} z\right) = \int_{-\infty}^{\frac{n_1}{n_2} z} f_W(w) dw$$

Therefore the density of Z, for $z > 0$ is given by

$$f(z) = \frac{n_1}{n_2} f_W(\frac{n_1}{n_2} z) = \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \frac{z^{\frac{n_1}{2}-1}}{(1+\frac{n_1}{n_2} z)^{\frac{n_1+n_2}{2}}} \frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})}.$$

Z is said to have the F distribution with degrees freedom parameters n_1 and n_2, denoted $Z \sim F_{n_1,n_2}$. ■

Remark 8.1.8. In the previous example, the F distribution essentially arises as the ratio U/V where U, V are independent χ^2 random variables. As the χ^2 distribution is a special case of the Gamma distribution, it follows by Example 5.5.11 that $\frac{U}{U+V}$ is distributed as a Beta random

variable. Further, $\frac{U+V}{U} = 1 + \frac{V}{U}$, so the two distributions are simple transformations of each other. In that sense, the F distribution is not a new distribution either, and it is studied separately mainly for its natural definition as the ratio of sample variances.

The distribution of the ratio of sample mean and sample variance plays an important role in estimation and hypothesis testing. This forms the motivation for the next example where the t distribution arises naturally.

Example 8.1.9. (t distribution) Let X_1 be a Normal $(0,1)$ random variable, and let X_2 be an independent χ^2_n random variable. We wish to find the density of Z, where

$$Z = \frac{X_1}{\sqrt{X_2/n}}.$$

Observe that $U = Z^2$ is given by $\frac{X_1^2}{X_2/n}$. Now, X_1^2 has χ^2_1 distribution (see Example 8.1.5), so applying Example 8.1.7 with $n_1 = 1$ and $n_2 = n$, we find that U has $F_{1,n}$ distribution. The density of U is given by

$$\begin{aligned} f_U(u) &= \left(\frac{1}{n}\right)^{\frac{1}{2}} \frac{u^{\frac{1}{2}-1}}{(1+\frac{1}{n}u)^{\frac{n+1}{2}}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n}{2})} \\ &= \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \frac{u^{-\frac{1}{2}}}{(1+\frac{u}{n})^{\frac{n+1}{2}}}. \end{aligned}$$

As X_1 is a symmetric random variable and $\sqrt{X_2/n}$ is positive valued, we conclude that Z is a symmetric random variable (Exercise 8.1.11). So, for $u > 0$,

$$\begin{aligned} P(U \le u) &= P(Z^2 \le u) \\ &= P(-\sqrt{u} \le Z \le \sqrt{u}) \\ &= P(Z \le \sqrt{u}) - P(Z \le -\sqrt{u}) \\ &= P(Z \le \sqrt{u}) - P(Z \ge \sqrt{u}) \\ &= 2P(Z \le \sqrt{u}) - 1 \end{aligned}$$

Therefore, if $f_Z(\cdot)$ is the density of Z then

$$f_U(u) = \frac{1}{\sqrt{u}}(f_Z(\sqrt{u})).$$

Hence for any $z \in \mathbb{R}$ the density of Z is given by

$$\begin{aligned} f_Z(z) &= |z|\, f_U(z^2) \\ &= |z| \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \frac{z^{2-\frac{1}{2}}}{(1+\frac{u}{n})^{\frac{n+1}{2}}} \\ &= \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1+\frac{z^2}{n}\right)^{-\frac{n+1}{2}}. \end{aligned}$$

Z is said to have the t distribution with n degrees of freedom, denoted $Z \sim t_n$. ■

8.1.3 *Distribution of Sampling Statistics from a Normal Population*

For $n \geq 2$, let $X_1, X_2, \ldots, X_n$ be an i.i.d. random sample from an arbitrary population having mean μ and variance σ^2. Consider the sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

We have already seen in Theorem 7.1.4 that $E[\overline{X}] = \mu$ and in Theorem 7.1.6 that $E[S^2] = \sigma^2$. It is unreasonable to expect that we would be able to precisely describe the distribution of $\overline{X}$ or S^2 unless the distribution of the population is known. It turns out that even in that case, it is not easy to derive these distributions in general. However, when the population is Normal, we can obtain the joint distribution of $\overline{X}$ and S^2 completely. The main result of this section is the following.

Theorem 8.1.10. *For $n \geq 2$, let $X_1, X_2, \ldots, X_n$ be an i.i.d. random sample with distribution $X \sim$ Normal(μ, σ^2). Let $\overline{X}$ and S^2 be defined as above. Then,*

(a) $\overline{X}$ is a Normal random variable with mean μ and variance $\frac{\sigma^2}{n}$.

(b) $(n-1)\frac{S^2}{\sigma^2}$ has the χ^2_{n-1} distribution.

(c) $\overline{X}$ and S^2 are independent.

Proof. (a) follows from Theorem 6.3.13. There are several proofs for (b) and (c), with the most common ones requiring some knowledge of Linear Algebra (e.g., see [Rao73]). Here we will follow Kruskal's proof as illustrated in [Stig84]. The proof is by the method of induction on the sample size n. To implement the inductive step, we shall replace $\overline{X}$ and S^2 with $\overline{X}_n$ and S^2_n for the rest of the proof. This notation also emphasizes that the distributions of $\overline{X}_n$ and S^2_n depend on n, and that as functions defined on the underlying sample space, they are in fact different random variables.
Step 1: (Proof for $n = 2$) Here

$$\overline{X}_2 = \frac{X_1 + X_2}{2} \text{ and } S^2_2 = \left(X_1 - \frac{X_1 + X_2}{2}\right)^2 + \left(X_2 - \frac{X_1 + X_2}{2}\right)^2 = \frac{(X_1 - X_2)^2}{2}. \quad (8.1.7)$$

As X_1 and X_2 are independent Normal random variables with mean μ and variance σ^2, by Theorem 6.3.13, $\frac{(X_1 - X_2)}{\sigma\sqrt{2}}$ is a Normal random variable with mean 0 and variance 1. Using Example 8.1.5, we know that $\frac{S^2_2}{\sigma^2}$ has χ^2_1 distribution and this proves (b).

From (8.1.7), $\overline{X}_2$ is a function of $X_1 + X_2$ and S^2_2 is a function of $X_1 - X_2$. Theorem 8.1.2 will imply that $\overline{X}_2$ and S^2_2 are independent if we show $X_1 + X_2$ and $X_1 - X_2$ are independent.

Let $\alpha, \beta \in \mathbb{R}$. Then using Theorem 6.3.13 again we have that $\alpha(X_1 + X_2) + \beta(X_1 - X_2) = (\alpha+\beta)X_1 + (\alpha-\beta)X_2$ is normally distributed. As this is true for any $\alpha, \beta \in \mathbb{R}$, $(X_1 + X_2, X_1 - X_2)$ has a bivariate Normal distribution by Definition 6.4.1. Using Theorem 6.2.2 (f) and (g), along with the fact that X_1 and X_2 are independent Normal random variables with mean μ and variance σ^2, we have

$$Cov[X_1 + X_2, X_1 - X_2] = Var[X_1] + Cov[X_2, X_1] - Cov[X_1, X_2] - Var[X_2] = 0.$$

Theorem 6.4.3 then implies that $X_1 + X_2$ and $X_1 - X_2$ are independent.

Step 2: (inductive hypothesis) Let us inductively assume that (a),(b), and (c) are true when $n = k$ for some $k \in \mathbb{N}$.

Step 3: (Proof for $n = k + 1$) We shall rewrite $\overline{X}_{k+1}$ and S^2_{k+1} using some elementary algebra.

$$\overline{X}_k - \overline{X}_{k+1} = \overline{X}_k - \frac{1}{k+1}\sum_{i=1}^{k+1} X_i = \left(1 - \frac{k}{k+1}\right)\overline{X}_k - \frac{1}{k+1}X_{k+1} = \frac{1}{k+1}(\overline{X}_k - X_{k+1}). \quad (8.1.8)$$

Adding and subtracting $\overline{X}_k$ inside the summand of S^2_{k+1}, we have

$$\begin{aligned}
S^2_{k+1} &= \frac{1}{k}\sum_{i=1}^{k+1}(X_i - \overline{X}_{k+1})^2 = \frac{1}{k}\sum_{i=1}^{k+1}(X_i - \overline{X}_k + \overline{X}_k - \overline{X}_{k+1})^2 \\
&= \frac{1}{k}\sum_{i=1}^{k+1}(X_i - \overline{X}_k)^2 + 2(X_i - \overline{X}_k)(\overline{X}_k - \overline{X}_{k+1}) + (\overline{X}_k - \overline{X}_{k+1})^2 \\
&= \frac{k-1}{k}S^2_k + \frac{1}{k}(X_{k+1} - \overline{X}_k)^2 + \frac{1}{k}\left(2(X_{k+1} - \overline{X}_k)(\overline{X}_k - \overline{X}_{k+1}) + (k+1)(\overline{X}_k - \overline{X}_{k+1})^2\right) \\
&= \frac{k-1}{k}S^2_k + \frac{1}{k}(X_{k+1} - \overline{X}_k)^2 - \frac{1}{k}\left(2(X_{k+1} - \overline{X}_k)\frac{(X_{k+1} - \overline{X}_k)}{k+1} + \frac{(X_{k+1} - \overline{X}_k)^2}{k+1}\right) \\
&= \frac{k-1}{k}S^2_k + \frac{1}{k+1}(X_{k+1} - \overline{X}_k)^2,
\end{aligned}$$

where we have used (8.1.8) in the second last inequality. Dividing thoughout by σ^2 and multiplying by k we have

$$\frac{k}{\sigma^2}S^2_{k+1} = \frac{k-1}{\sigma^2}S^2_k + \frac{k}{\sigma^2(k+1)}(X_{k+1} - \overline{X}_k)^2. \quad (8.1.9)$$

Part (a) follows again from Theorem 6.3.13. To prove (b), it is enough to show that

$$\left(\sqrt{\frac{k}{(k+1)\sigma^2}}\right)(X_{k+1} - \overline{X}_k) \sim \text{Normal}\,(0,1) \text{ and is independent of } \frac{(k-1)}{\sigma^2}S^2_k.$$

This is so because $\frac{k}{\sigma^2(k+1)}(X_{k+1} - \overline{X}_k)^2$ then has the χ^2_1 distribution by Example 8.1.5, and is independent of $\frac{(k-1)}{\sigma^2}S^2_k$ by Theorem 8.1.2; by the induction hypothesis $\frac{(k-1)}{\sigma^2}S^2_k$ has the χ^2_{k-1}

distribution, so using (8.1.9) along with Example 5.5.6 will imply that $\frac{k}{\sigma^2}S_{k+1}^2$ has the χ_k^2 distribution. It is a routine calculation using Theorem 6.3.13 to verify the above distribution by noting that

$$\left(\sqrt{\frac{k}{(k+1)\sigma^2}}\right)(X_{k+1}-\overline{X}_k)=\left(\sqrt{\frac{(k+1)\sigma^2}{k}}\right)X_{k+1}-\sum_{i=1}^{k}\frac{1}{k}\left(\sqrt{\frac{k}{(k+1)\sigma^2}}\right)X_i.$$

By the induction hypothesis, $\overline{X}_k$ and $\frac{k-1}{\sigma^2}S_k^2$ are independent. As $X_1,\ldots,X_k,X_{k+1}$ are mutually independent, Theorem 8.1.2 implies that X_{k+1} is independent of $\overline{X}_k$ and $\frac{k-1}{\sigma^2}S_k^2$. Therefore,

$$\overline{X}_k,\quad \frac{k-1}{\sigma^2}S_k^2,\quad X_{k+1}\qquad \text{are mutually independent random variables.} \tag{8.1.10}$$

Consequently, another application of Theorem 8.1.2 will then imply that $\frac{k}{\sigma^2(k+1)}(X_{k+1}-\overline{X}_k)^2$ and $\frac{(k-1)}{\sigma^2}S_k^2$ are independent random variables.

To prove (c), it is enough to show that $\overline{X}_{k+1}$ and $X_{k+1}-\overline{X}_k$ are independent. The reason is the following:

(i) Theorem 8.1.2 then implies $\overline{X}_{k+1}$ is independent of $\frac{k}{\sigma^2(k+1)}(X_{k+1}-\overline{X}_k)^2$;

(ii) $\overline{X}_{k+1}$ is a function of X_{k+1} and $\overline{X}_k$. So (8.1.10) and Theorem 8.1.2 will then imply $\overline{X}_{k+1}$ is independent of $\frac{(k-1)}{\sigma^2}S_k^2$ and also $\frac{k}{\sigma^2(k+1)}(X_{k+1}-\overline{X}_k)^2$ is independent of $\frac{(k-1)}{\sigma^2}S_k^2$;

(iii) Using (i) and (ii) we can conclude that $\overline{X}_{k+1}$, $\frac{(k-1)}{\sigma^2}S_k^2$, and $\frac{k}{\sigma^2(k+1)}(X_{k+1}-\overline{X}_k)^2$ are mutually independent; and

(iv) finally S_{k+1}^2 is a function $\frac{(k-1)}{\sigma^2}S_k^2$, and $\frac{k}{\sigma^2(k+1)}(X_{k+1}-\overline{X}_k)^2$ by (8.1.9). Then (iii) and Theorem 8.1.2 will imply that S_{k+1}^2 and $\overline{X}_{k+1}$ are independent.

Let $\alpha,\beta\in\mathbb{R}$. We have

$$\alpha(\overline{X}_{k+1})+\beta(X_{k+1}-\overline{X}_k)=\sum_{i=1}^{k}\left(\frac{\alpha}{k+1}-\frac{\beta}{k}\right)X_i+\left(\frac{\alpha}{k+1}-\beta\right)X_{k+1}.$$

Theorem 6.3.13 will imply that $\alpha(\overline{X}_{k+1})+\beta(X_{k+1}-\overline{X}_k)$ is is normally distributed random variable for any $\alpha,\beta\in\mathbb{R}$. So by Definition 6.4.1 $(\overline{X}_{k+1},X_{k+1}-\overline{X}_k)$ is a bivariate normal random variable. Further, from Theorem 6.2.2 (f) and (g), we have

$$\begin{aligned}
Cov[\overline{X}_{k+1},X_{k+1}-\overline{X}_k] &= Cov[\frac{k\overline{X}_k+X_{k+1}}{k+1},X_{k+1}-\overline{X}_k]\\
&= \frac{1}{k+1}Var[X_{k+1}]-Cov[\overline{X}_k,X_{k+1}]-\frac{k}{k+1}Var[\overline{X}_k]\\
&= \frac{1}{k+1}\sigma^2+0+-\frac{k}{k+1}\frac{\sigma^2}{k}=0,
\end{aligned}$$

where we have used (8.1.10) in the last line. From Theorem 6.4.3 we conclude that $\overline{X}_{k+1},X_{k+1}-\overline{X}_k$ are independent. ■

The following important Corollary connects the sampling distributions of $\overline{X}$ and S^2 to the t distribution, and will be important in the context of confidence intervals, which we discuss in Chapter 9.

Corollary 8.1.11. *For $n \geq 2$, let $X_1, X_2, \ldots, X_n$ be an i.i.d. random sample with distribution $X \sim Normal(\mu, \sigma^2)$. Let $\overline{X}$ and S^2 be as above. Then*

$$\frac{\sqrt{n}(\overline{X} - \mu)}{S}$$

has the t_{n-1} distribution.

Proof. From Theorem 8.1.10 it is clear that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}\,(0,1) \quad \text{and} \quad \frac{(n-1)}{\sigma^2} S^2 \sim \chi^2_{n-1}.$$

Noting that

$$\frac{\sqrt{n}(\overline{X} - \mu)}{S} = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{1}{n-1}\frac{(n-1)S^2}{\sigma^2}}},$$

the result follows by Example 8.1.9. ∎

EXERCISES

Ex. 8.1.1. Let $n \geq 1$. F be the joint distribution function of real valued discrete random variables $X_1, X_2, \ldots, X_n$ as in (8.1.1).

(a) Suppose $n = 2$. Show that for $(s,t) \in \mathbb{R}^2$,

$$P(X_1 = s, X_2 = t) = \lim_{u \downarrow s, v \downarrow t} F(u,v) - F(s,t)$$

(b) Reformulate and prove, part (a) for general $n \geq 1$.

Ex. 8.1.2. Verify that each of $f : \mathbb{R}^3 \to \mathbb{R}$ are density functions on $\mathbb{R}^3$.

(a) $f(x_1,x_2,x_3) = \begin{cases} \frac{2}{3}(x_1+x_2+x_3) & \text{if } 0 < x_i < 1, i = 1,2,3. \\ 0 & \text{otherwise} \end{cases}$

(b) $f(x_1,x_2,x_3) = \begin{cases} \frac{1}{8}(x_1^2+x_2^2+x_3^2) & \text{if } 0 < x_i < 2, i = 1,2,3. \\ 0 & \text{otherwise} \end{cases}$

(c) $f(x_1,x_2,x_3) = \begin{cases} \frac{2}{81}x_1x_2x_3 & \text{if } 0 < x_i < 3, i = 1,2,3. \\ 0 & \text{otherwise} \end{cases}$

(d) $f(x_1,x_2,x_3) = \begin{cases} \frac{3}{4}(x_1x_2+x_1x_3+x_2x_3) & \text{if } 0 < x_i < 1, i = 1,2,3. \\ 0 & \text{otherwise} \end{cases}$

Ex. 8.1.3. Suppose (X_1, X_2, X_3) have a joint density $f : \mathbb{R}^3 \to \mathbb{R}$ given by

$$f(x_1, x_2, x_3) = \begin{cases} \frac{4}{3}(x_1^3 + x_2^3 + x_3^3) & \text{if } 0 < x_i < 1, i = 1, 2, 3. \\ 0 & \text{otherwise} \end{cases}$$

(a) Find $P(X_1 < \frac{1}{2}, X_3 > \frac{1}{2})$.

(b) Find the joint density of (X_1, X_2),(X_1, X_3), (X_2, X_3).

(c) Find the marginal densities of X_1, X_2, and X_3.

Ex. 8.1.4. Let D be a set in $\mathbb{R}^3$ with a well defined volume. (X_1, X_2, X_3) are said be uniform on a set D if they have a joint density given by

$$f(x_1, x_2, x_3) = \begin{cases} \frac{1}{\text{Volume}(D)} & \text{if } x \in D \\ 0 & \text{otherwise.} \end{cases}$$

Suppose D is a cube of dimension R.

(a) Find the joint density (X_1, X_2, X_3) which is uniform on D.

(b) Find the marginal density of X_1, X_2, X_3.

(c) Find the joint density of (X_1, X_2),(X_1, X_3),(X_3, X_2).

Ex. 8.1.5. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having a common distribution function $F : \mathbb{R} \to [0, 1]$ and probability density function $f : \mathbb{R} \to \mathbb{R}$. Let $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ be the corresponding order statistics. Show that for $1 \leq i < j \leq n$, $(X_{(i)}, X_{(j)})$ has a joint density function given by

$$f_{X_{(i)}, X_{(j)}}(x, y) = \frac{n!}{(i-1)!(j-1-i)!(n-j)!} f(x) f(y) [F(x)]^{i-1} [F(y) - F(x)]^{j-1-i} [1 - F(y)]^{n-j},$$

for $-\infty < x < y < \infty$.

Ex. 8.1.6. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having a common distribution $X \sim \text{Uniform}(0, 1)$. Let $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ be the corresponding order statistics. Show that $\frac{X_{(1)}}{X_{(n)}}$ and $X_{(n)}$ are independent random variables.

Ex. 8.1.7. Let $\{U_i : i \geq 1\}$ be a sequence of i.i.d. Uniform$(0, 1)$ random variables, and let $N \sim \text{Poisson}(\lambda)$. Find the distribution of $V = \min\{U_1, U_2, \ldots, U_{N+1}\}$.

Ex. 8.1.8. Let $-\infty < a < b < \infty$. Let $X_1, X_2, \ldots, X_n$ i.i.d. $X \sim \text{Uniform}(a, b)$. Find the probability density function of $M = \frac{X_{(1)} + X_{(n)}}{2}$.

Ex. 8.1.9. Let X_1, X_2 be two independent standard Normal random variables. Find the distribution of $Z = X_{(1)}^2$.

Ex. 8.1.10. Let $X_1, X_2, \ldots, X_n$ be i.i.d. Uniform$(0, 1)$ random variables.

(a) Find the conditional distribution of $X_{(n)} \mid X_{(1)} = x$ for some $0 < x < 1$.

(b) Find $E[X_{(n)} \mid X_{(1)} = x]$ and $Var[X_{(n)} \mid X_{(1)} = x]$.

Ex. 8.1.11. Suppose X is a symmetric continuous random variable. Let Y be a continuous random variable such that $P(Y > 0) = 1$. Show that $\frac{X}{Y}$ is symmetric.

Ex. 8.1.12. Verify (8.1.4).

Ex. 8.1.13. Suppose $X_1, X_2, \ldots$ are i.i.d. Cauchy$(0, 1)$ random variables.

(a) Fix $z \in \mathbb{R}$. Find a, b, c, d such that

$$\frac{1}{1+x^2}\frac{1}{1+(z-x)^2} = \frac{ax+b}{1+x^2} + \frac{cx+d}{1+(z-x)^2},$$

for all $x \in \mathbb{R}$.

(b) Show that $X_1 + X_2 \sim$ Cauchy $(0, 2)$.

(c) Use induction to show that $X_1 + X_2 + \ldots + X_n \sim$ Cauchy $(0, n)$.

(d) Use Lemma 5.3.2 to show that $\overline{X}_n \sim$ Cauchy $(0, 1)$.

Ex. 8.1.14. Suppose U, V are independent random variables with χ^2_m and χ^2_n respectively. Then show that that $Z = \frac{U}{U+V}$ is distributed as Beta$(\frac{m}{2}, \frac{n}{2})$

8.2 WEAK LAW OF LARGE NUMBERS

For $n \geq 1$, let $X_1, X_2, \ldots, X_n$ be an i.i.d. random sample from a population whose distribution is given by a random variable X which has mean μ. In Chapter 7 we considered the sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and showed in Theorem 7.1.4 that $E[\overline{X}] = \mu$. We also discussed that $\overline{X}$ could be considered as an estimate for μ. The following result makes this precise, and is referred to as the Weak law of large numbers. To emphasise the dependence of the sample mean and its behaviour on n, we will denote $\overline{X}$ by $\overline{X}_n$.

Theorem 8.2.1. (Weak Law of Large Numbers) *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables. Assume that X_1 has finite mean μ and finite variance σ^2. Then for any $\epsilon > 0$*

$$\lim_{n\to\infty} P(\mid \overline{X}_n - \mu \mid > \epsilon) = 0, \tag{8.2.1}$$

Proof. Let $\epsilon > 0$ be given. We note that

$$E[\overline{X}_n] = E[\frac{1}{n}\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n}\frac{1}{n}E[X_i] = \frac{n\mu}{n} = \mu.$$

Using Theorem 4.2.4, Theorem 4.2.6 and Exercise 6.2.16 we have

$$\begin{aligned}
\text{Var}[\overline{X}_n] &= \text{Var}[\frac{1}{n}\sum_{i=1}^{n} X_i] \\
&= \frac{1}{n^2}\text{Var}[\sum_{i=1}^{n} X_i] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}[X_i] \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

So we have shown that the random variable $\overline{X}_n$ has finite expectation and variance. By Chebychev's inequality (apply Theorem 6.1.13 (a) with $k = \frac{\epsilon}{\sigma}$), we have

$$P(|\overline{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Therefore as $0 \leq P(|\overline{X}_n - \mu| > \epsilon)$ for all $n \geq 1$ and $\frac{\sigma^2}{n\epsilon^2} \to 0$ as $n \to \infty$, by standard results in real analysis we conclude that

$$\lim_{n\to\infty} P(|\ \overline{X}_n - \mu\ | > \epsilon) = 0.$$

■

Remark 8.2.2. The convergence of $\overline{X}_n$ to μ actually happens with probability one. That is, if we consider the event $A = \left\{\lim_{n\to\infty} \overline{X}_n = \mu\right\}$, then $P(A) = 1$. This result is referred to as the Strong Law of large numbers. We state and prove it in Appendix C (see Theorem A.2.1).

We are often interested in using asymptotic results such as this as approximations when n is a large, but finite number. To develop a sense about the reliability of such approximations, we devote the remainder of this section to simulations of such behavior.

The Weak Law of large numbers is most commonly used to estimate probabilities of events. However, before exploring applications, we first look at some simpler examples where we estimate the mean of a distribution.

Example 8.2.3. Suppose $X \sim \text{Uniform}(0, 1)$. What is $E\left[\log\left(\frac{X}{1-X}\right)\right]$? It is easy to argue, using symmetry, that the answer should be 0 (see Exercise 8.2.1). To verify this using simulation, we can simply generate a large number of Uniform$(0, 1)$ random variables, transform them, and compute their sample mean. If the Weak Law gives a good approximation for finite samples, this should be "close" to the true expectation.

```
u <- runif(10000)
mean(log(u / (1-u)))
```

```
[1] 0.006766486
```

Of course, however good an approximation, this estimate is still random, so we should replicate it several times to get an idea of its general behaviour.

```
replicate(10, {
    x <- runif(10000)
    mean(log(x / (1 - x)))
})
```

```
 [1]  0.006112412 -0.009243912  0.019540336  0.022273225  0.003628815
 [6]  0.016525060 -0.017777608  0.011400835 -0.002747576 -0.020943686
```

These ten replications suggest that the approximation is usually correct only up to the first decimal place, even though the value of $n = 10000$ might normally be considered large. Not surprisingly, the approximation gets worse for $n = 100$.

```
replicate(10, {
    x <- runif(100)
    mean(log(x / (1 - x)))
})
```

```
 [1] -0.021428316  0.003939597  0.171115165  0.122717112  0.235109230
 [6] -0.007103688  0.082358211  0.097755404 -0.185062653 -0.108332961
```

To get a sense of how the approximation improves with n, it is common to plot the cumulative or partial means as a function of n. For example, Figure 8.1 is created using

```
N <- 10000
i <- seq(1, N) # to be used as denominator
x <- runif(N)
m <- cumsum(log(x / (1 - x))) / i
```

```
xyplot(m ~ i, xlab = "Index", ylab = "Partial Mean", type = "l")
```

This plot suggests that the estimate gets close to zero for fairly small n, and after that improvement is not substantial. This plot, however, only tells us about the behaviour of one particular sequence of random variables, whose partial means are guaranteed to converge to the true mean as $n \to 0$ according to the Strong Law, which we have stated but not proved. The Weak Law, on the other hand, states a result about the *distribution* of the sample mean. To assess whether it holds, we look at independent replications of the experiment and plot the resulting paths taken by the partial means together. We omit the code used to do this, but show the result of one such simulation experiment in Figure 8.2. One can observe that there is a reduction in the variance of the partial means as n increases, which was the essential requirement in the proof of the Weak Law. ■

Example 8.2.4. We modify the previous example as follows. Suppose $U, V \sim \text{Uniform}(0, 1)$ are independent, and $X = \max(U, V)$. What is $E(\log \frac{X}{1-X})$?

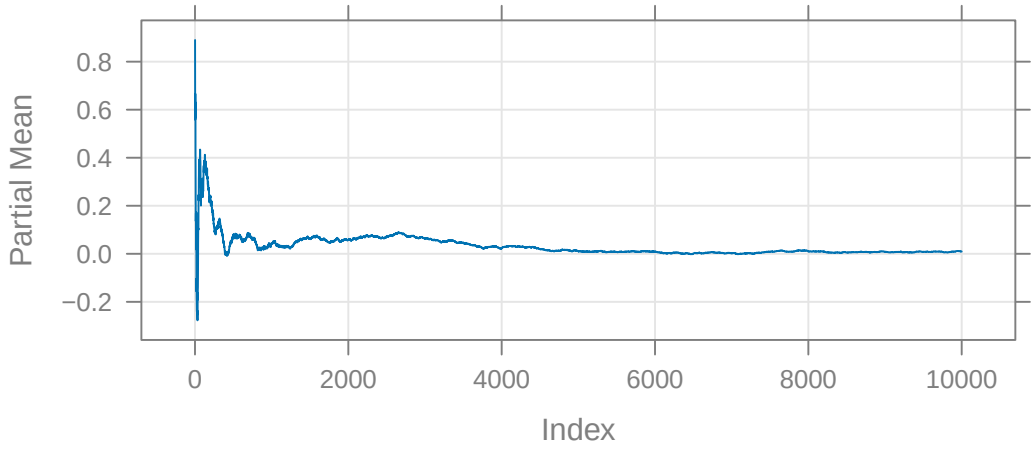


Figure 8.1: Cumulative or partial means computed from 10000 random samples from the population $\log \frac{X}{1-X}$, where X follows Uniform$(0, 1)$.

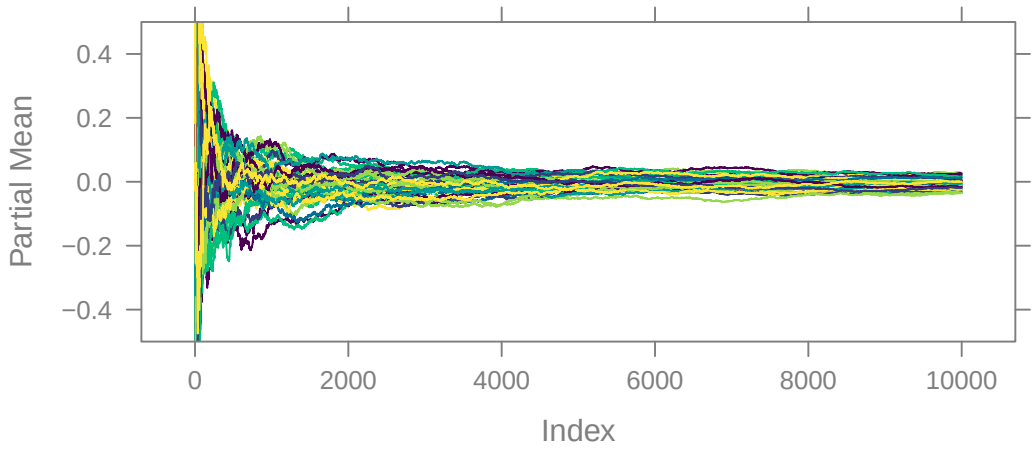


Figure 8.2: Results of the same experiment that is shown in Figure 8.1, replicated 50 times. For each replication, cumulative or partial means computed from 10000 random samples are shown. The underlying population is $\log \frac{X}{1-X}$, where X follows Uniform$(0, 1)$.

The answer is not as obvious in this case. An approximate answer is easy to obtain by invoking the Weak Law of large numbers.

```
replicate(10, {
    u <- runif(10000)
    v <- runif(10000)
    x <- pmax(u, v)
    mean(log(x / (1 - x)))
})
```

```
 [1] 0.9981781 0.9868826 1.0355666 1.0128563 1.0030963 1.0011916
 [7] 0.9939234 0.9900825 0.9905087 0.9863563
```

These results suggest that the expectation is 1, a fact that can be verified by explicit computation (See Exercise 8.2.1). ■

Example 8.2.5. Suppose that U and V are independent Uniform$(0, 1)$, and interpret them as coordinates of a point in $\mathbb{R}^2$. Suppose we want to calculate the expected norm of (U, V). In other words, if $Z = \sqrt{U^2 + V^2}$, we want to calculate $E[Z]$.

As before, we can estimate the expectation by simulating the experiment a large number of times.

```
replicate(10, {
    u <- runif(10000)
    v <- runif(10000)
    z <- sqrt(u^2 + v^2)
    mean(z)
})
```

```
 [1] 0.7642686 0.7674234 0.7660785 0.7681380 0.7608846 0.7651592
 [7] 0.7621663 0.7715568 0.7613604 0.7678146
```

See Exercise 8.2.2 for explicit computation of $E[Z]$. ■

Theorem 8.2.1 states that for any $\epsilon > 0$, the probability $P(|\overline{X}_n - \mu| > \epsilon)$ goes to zero as $n \to \infty$. This mode of convergence of the sample mean $\overline{X}_n$ to the true mean μ is called "convergence in probability" . We define it precisely below.

Definition 8.2.6. *A sequence $X_1, X_2, \ldots$ is said to converge in probability to a random variable X if for any $\epsilon > 0$*

$$\lim_{n\to\infty} P(|\, X_n - X \,| > \epsilon) = 0, \tag{8.2.2}$$

The notation

$$X_n \xrightarrow{p} X$$

is typically used to convey that the sequence $X_1, X_2, \ldots$ converges in probability to X.

Note that in the above definition the limit is allowed to be a non-trivial random variable X, although in most examples we will consider, X will be a constant.

Example 8.2.7. Let $X_1, X_2, \ldots$ be i.i.d. random variables from the Uniform$(0,1)$ distribution. We already know by the law of large numbers that $\overline{X}$ converges to $E(X_1) = \frac{1}{2}$ in probability. Often we are interested in other functionals (i.e. $f(X_1, X_2, \ldots, X_n)$ for some suitable f and $n \geq 1$) of the sample and their convergence properties. As an example, consider the n-th order statistic $X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$. Intuitively, as n increases, it is more and more likely that $X_{(n)}$ will get closer to its maximum possible value 1. To see this formally, first note that for $\epsilon > 1$,

$$P\left(\left|X_{(n)} - 1\right| \geq \epsilon\right) = P\left(X_{(n)} \leq 1 - \epsilon\right) + P\left(X_{(n)} \geq 1 + \epsilon\right) = 0.$$

For any $0 < \epsilon < 1$,

$$\begin{aligned} P\left(\left|X_{(n)} - 1\right| \geq \epsilon\right) &= P\left(X_{(n)} \leq 1 - \epsilon\right) + P\left(X_{(n)} \geq 1 + \epsilon\right) \\ &= P\left(X_{(n)} \leq 1 - \epsilon\right) + 0 \\ &= P\left(\bigcap_{i=1}^{n} \{X_i \leq 1 - \epsilon\}\right) \\ &= \left(1 - \epsilon\right)^n. \end{aligned}$$

As $\lim_{n\to\infty} (1-\epsilon)^n = 0$ for $0 < \epsilon < 1$, it follows from definition 8.2.6 that $X_{(n)} \xrightarrow{p} 1$ as $n \to \infty$. ■

An important application of the Weak Law of large numbers follows by noting that the sample proportion discussed in Section 7.1.2 is the sample mean of Bernoulli random variables.

Example 8.2.8. Suppose we are interested in an event A and want to estimate $p = P(X \in A)$. We consider a sample $X_1, X_2, \ldots, X_n$ which is i.i.d. X. We define a sequence of random variables $\{Y_n\}_{n\geq 1}$ by

$$Y_n = \begin{cases} 1 & \text{if } X_n \in A \\ 0 & \text{if } X_n \notin A. \end{cases}$$

Clearly Y_n are independent (as the X_n are), and further they are identically distributed with $P(Y_n = 1) = P(X_n \in A) = p$. In particular, $\{Y_n\}$ is an i.i.d. Bernoulli(p) sequence of random variables. We readily observe (as done in Chapter 7) that

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\#\{X_i \in A\} = \hat{p}.$$

Hence the Weak Law of large numbers (applied to the sequence Y_n) implies that the sample proportion will converge to the true proportion p in probability. This provides legitimacy, as discussed earlier, to the relationship between probability and relative frequency. ■

EXERCISES

Ex. 8.2.1. Find $E(\log \frac{X}{1-X})$ when

1. $X \sim$ Uniform$(0, 1)$.
2. $X = \max(U, V)$ when $U, V \sim$ Uniform$(0, 1)$ and are independent.

Ex. 8.2.2. Let $U, V \sim$ Uniform(0,1) and be independent.

1. Find the distribution of U^2
2. Find the distribution of $U^2 + V^2$
3. Find the distribution of norm of $Z = (U, V)$ and $E[Z]$

Ex. 8.2.3. Let $(U, V) \sim$ Uniform(D) where $D = \{(x, y) : x^2 + y^2 = 1\}$. Find the distribution of norm of $Z = (U, V)$ and $E[Z]$.

Ex. 8.2.4. Let X, X_1, X_2, ... be i.i.d. random variables that are uniformly distributed over the interval $(0, 1)$. Consider the first order statistic $X_{(1)} = \min\{X_1, \cdots, X_n\}$. Show that $X_{(1)}$ converges to 0 in probability.

Ex. 8.2.5. Let $X_1, X_2, \ldots, X_n, \ldots$ be i.i.d. random variables with finite mean and variance. Define

$$Y_n = \frac{2}{n(n+1)}\sum_{i=1}^{n} iX_i.$$

Show that $Y_n \xrightarrow{p} E(X_1)$ as $n \to \infty$.

Ex. 8.2.6. Let $\{X_i : i \geq 1\}$ be a sequence of i.i.d. Normal $(0, 1)$ random variables. Let $S_n = \sum_{1}^{n} X_i$. Design a suitable R-code as in Example 7.1.9 that will provide an estimate of the probability that $S_1, \ldots, S_{100}$ all have the same sign.

Ex. 8.2.7. Suppose X_n and X are random variables such that $X_n \xrightarrow{p} X$ as $n \to \infty$. Suppose $h : \mathbb{R} \to \mathbb{R}$ is a continuous function. Then show that $h(X_n) \xrightarrow{p} h(X)$ as $n \to \infty$.

8.3 CONVERGENCE IN DISTRIBUTION

When discussing a collection of random variables it makes sense to think of them as a sequence of objects, and as with any sequence in calculus we may ask whether the sequence converges in any way. We have already seen "convergence in probability" in the previous section. Here we will be interested in what is known as "convergence in distribution". This type of convergence plays an important role in understanding the limiting distribution of the sample mean, as we will see later, particularly in the Central Limit Theorem, Theorem 8.4.1.

Definition 8.3.1. *A sequence $X_1, X_2, \ldots$ is said to converge in distribution to a random variable X if $F_{X_n}(x)$ converges to $F_X(x)$ at every point x for which F_X is continuous. The following notation*

$$X_n \xrightarrow{d} X$$

is typically used to convey that the sequence $X_1, X_2, \ldots$ converges in distribution to X.

EXAMPLE 8.3.2. Let $X_n \sim \text{Uniform}(0, \frac{1}{n})$ so that the distribution function is

$$F_{X_n}(x) = \begin{cases} 0 & \text{if } 0 \leq x \\ nx & \text{if } 0 < x < \frac{1}{n} \\ 1 & \text{if } x \geq \frac{1}{n} \end{cases}$$

and it is then easy to see that $F_{X_n}(x)$ converges to

$$F(x) = \begin{cases} 0 & \text{if } 0 \leq x \\ 1 & \text{if } x > 0 \end{cases}$$

If X is the constant random variable for which $P(X = 0) = 1$, then X has distribution function

$$F_X(x) = \begin{cases} 0 & \text{if } 0 < x \\ 1 & \text{if } x \geq 0 \end{cases}$$

It is not true that $F_X(x) = F(x)$, but the two are equal at points where they are continuous. Therefore the sequence $X_1, X_2, \ldots$ converges in distribution to the constant random variable 0. ■

Note that this form of convergence does not generally guarantee that probabilities associated with X can be derived as limits of probabilities associated with X_n. For instance, in the example above $P(X_n = 0) = 0$ for all n while $P(X = 0) = 1$. However, with a few additional assumptions a stronger claim may be made.

Theorem 8.3.3. *Let $f_{X_1}, f_{X_2}, \ldots$ be the respective densities of continuous random variables $X_1, X_2, \ldots$. Suppose they converge in distribution to a continuous random variable X with density f_X. Then for every interval A we have $P(X_n \in A) \to P(X \in A)$.*

Proof. As X is a continuous random variable $F_X(x)$ is the integral of a density, and thus a continuous function. Therefore convergence in distribution guarantees that $F_{X_n}(x)$ converges to $F_X(x)$ everywhere. Let $A = (a, b)$ (and note that whether or not endpoints are included does not matter as all random variables are taken to be continuous). Then

$$\begin{aligned} P(X_n \in A) &= \int_a^b f_{X_n}(x)\, \mathrm{d}x \\ &= F_{X_n}(b) - F_{X_n}(a) \\ &\to F_X(b) - F_X(a) \\ &= \int_a^b f_X(x)\, \mathrm{d}x = P(X \in A). \end{aligned}$$

■

When a sequence $Y_1, Y_2, \ldots$ of random variables converges in probability to a constant c, one often then tries to understand how the distribution of suitably *scaled* versions of the fluctuations $Y_n - c$ behave in the limit. In many cases, we are able to identify the correct scaling at which the scaled fluctuations converge in *distribution* to a non-constant random variable. The most well known example of this is the Central Limit Theorem, to be studied in the next section, which states that the fluctuations of the sample mean of n i.i.d. random variables scaled by $\sqrt{n}$ converges in distribution to standard Normal, under a finite second moment hypothesis. We shall now discuss another fundamental example.

EXAMPLE 8.3.4. Let $X_1, X_2, \ldots$ be i.i.d. Uniform$(0, 1)$ random variables. Consider $M_n = \min(X_1, X_2, \ldots, X_n)$, the minimum value among the first n observations. Normally, we would denote M_n simply by $X_{(1)}$, but here we use a different notation to emphasize that the minimum can change with n.

We saw earlier in Exercise 8.2.4 that that $M_n \xrightarrow{p} 0$ and in Example 8.2.3. that convergence in probability is at a certain rate. To understand the fluctuations around the limit we shall try to identify the correct scaling. To see this, first note that $E(M_n) = 1/(n+1)$, so $(n+1)M_n$ has expected value 1 for all n. Thus we could use a scaling factor of "n" or "$n+1$". So, for $x \in \mathbb{R}$,

$$\begin{aligned} P\left(nM_n > x\right) &= P\left(M_n > \frac{x}{n}\right) \\ &= P\left(X_1 > \frac{x}{n}, \ldots, X_n > \frac{x}{n}\right) \\ &= \left(P\left(X_1 > \frac{x}{n}\right)\right)^n \text{ by independence of } X_1, \ldots, X_n \\ &= \left(1 - \frac{x}{n}\right)^n \to e^{-x} \text{ as } n \to \infty. \end{aligned}$$

In other words, if Z is exponentially distributed with mean 1, then we have shown that $P(nM_n \leq x) \to P(Z \leq x)$ for all x. So we have $M_n \xrightarrow{p} 0$ and $n(M_n - 0) \xrightarrow{d} Z$. ■

Establishing convergence in distribution using the definition, as done in the previous example, is not always possible. There are three key results that we will use in the book. These provide sufficient conditions that are intuitive and often easier to check. The first result deals with the

case of convergence in distribution for continuous random variables, and states that pointwise convergence of densities implies convergence in distribution.

Theorem 8.3.5. *(Scheffé's Lemma) Let $f_{X_1}, f_{X_2}, \ldots$ be the respective densities of continuous random variables $X_1, X_2, \ldots$, and let f_X be the density of a continuous random variable X. Suppose $f_{X_n}(x) \to f_X(x)$ as $n \to \infty$ for all $x \in \mathbb{R}$. Then, $X_n \xrightarrow{d} X$ as $n \to \infty$.*

This is a deceptively simple result. After all, one could argue that if $f_{X_n}(\cdot)$ converges to $f_X(\cdot)$ pointwise as $n \to \infty$, then so should

$$\int_{-\infty}^{a} f_{X_n}(u)du \to \int_{-\infty}^{a} f_X(u)du$$

as $n \to \infty$ for any $a \in \mathbb{R}$. However, such interchanging of limits and integrals is not always valid. The result that permits it in this particular situation, known as the "dominated convergence theorem", is beyond the scope of this book.

Example 8.3.6. Suppose $X_n \sim \text{Normal}\left(\frac{1}{n}, 1\right)$. Then it is intuitively clear that $X_n \xrightarrow{d} Z$, where $Z \sim \text{Normal}\,(0,1)$. This follows from an elementary application of Scheffé's Theorem, as

$$f_{X_n}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(x-\frac{1}{n}\right)^2} \to \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} = f_Z(x) \text{ for all } x \in \mathbb{R}.$$

A direct proof is also simple in this case. As $Y_n = X_n - \frac{1}{n}$ has the Normal $(0,1)$ distribution, we have

$$F_{X_n}(x) = P\left(X_n \le x\right) = P\left(Y_n \le x - 1/n\right) = F_Z\left(x - 1/n\right) \to F_Z(x)$$

as $n \to \infty$ for all $x \in \mathbb{R}$ because F_Z is continuous everywhere. ■

Example 8.3.7. Let $X_n \sim t_n$ distribution. Then

$$f_{X_n}(t) = \frac{\Gamma(\frac{n}{2} + \frac{1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{1}{2}(n+1)}$$

for $t \in \mathbb{R}$. It is straightforward to verify that (see Exercise 8.3.1)

$$f_{X_n}(t) \to \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2}) \tag{8.3.1}$$

as $n \to \infty$. Consequently by Scheffé's Theorem $X_n \xrightarrow{d} Z$ as $n \to \infty$ where $Z \sim$ Normal(0,1). ■

The second result, which works for both discrete and continuous random variables, formalizes the intuition that if all moments of X_n exist and they converge to respective moments of X, then X_n should converge in distribution to X. Unfortunately, a proof of this result is also beyond the scope of this book.

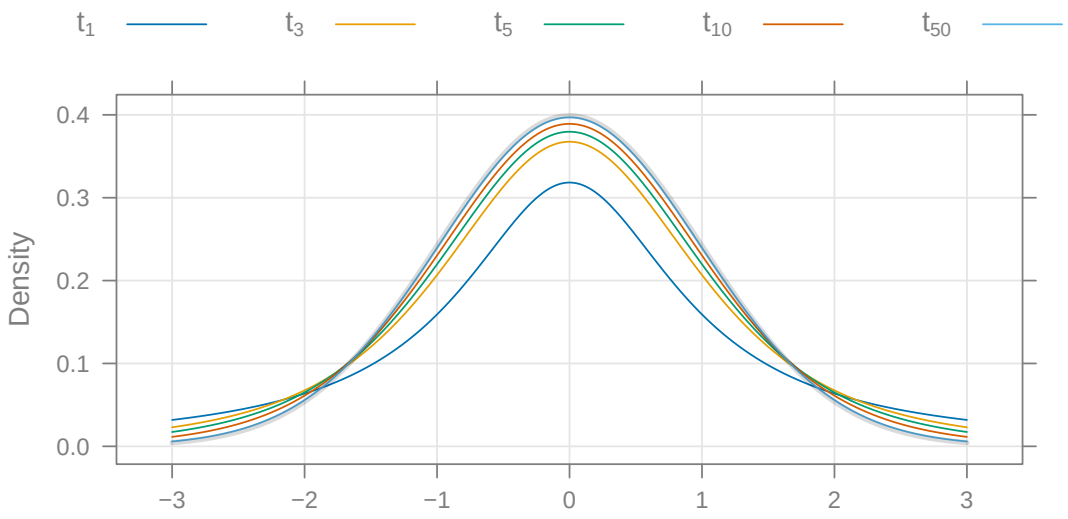


Figure 8.3: Density of the t_n distribution converging to that of standard Normal as $n \to \infty$, illustrated using the parameter values $n = 1, 3, 5, 10$ and 50. The thick grey line, which represents the standard Normal density, is almost indistingushabe from the t_{50} density.

Theorem 8.3.8. *(M.G.F. Convergence Theorem) Let $X_1, X_2, \ldots$ be a sequence of random variables whose moment generating functions $M_n(t)$ exist in an interval containing zero. If $M_n(t) \to M(t)$ on that interval, where $M(t)$ is the moment generating function of a random variable X, then X_n converges to X in distribution.*

To illustrate the use of this result, consider an alternative proof of the limiting relationship between Binomial and Poisson random variables (See Theorem 2.2.2).

Example 8.3.9. Let $X \sim \text{Poisson}(\lambda)$ and let $X_n \sim \text{Binomial}(n, \frac{\lambda}{n})$. Then X_n converges in distribution to X.

The moment generating function of a Binomial variable was already computed in Example 6.3.7. Therefore,

$$\begin{aligned} M_{X_n}(t) &= \left(\frac{\lambda}{n}e^t + \left(1 - \frac{\lambda}{n}\right)\right)^n \\ &= \left(1 + \frac{\lambda(e^t - 1)}{n}\right)^n \end{aligned}$$

Using Exercise 8.4.4, we see that

$$M_{X_n}(t) \to e^{\lambda(e^t - 1)}.$$

On the other hand, the moment generating function of X is

$$\begin{aligned} M_X(t) &= E[e^{tX}] \\ &= \sum_{j=0}^{\infty} e^{tj} P(X=j) \\ &= \sum_{j=0}^{\infty} e^{tj} \frac{\lambda^j e^{-\lambda}}{j!} \\ &= e^{\lambda e^t} \cdot e^{-\lambda} \cdot \sum_{j=0}^{\infty} \frac{(\lambda e^t)^j e^{-\lambda e^t}}{j!} \\ &= e^{\lambda(e^t-1)} \end{aligned}$$

where the series equals 1 since it is simply the sum of the probabilities of a Poisson(λe^t) random variable.

Since $M_{X_n}(t) \to M_X(t)$, by the M.G.F. convergence theorem (Theorem 8.3.8), X_n converges in distribution to X. That is, Binomial(n,p) random variables converge in distribution to a Poisson(λ) distribution when $p = \frac{\lambda}{n}$ and $n \to \infty$. ■

The last result cannot be used to establish convergence in distribution directly. However, if we already know that a sequence $X_n \xrightarrow{d} X$, then this result can often be used to establish the convergence in distribution of small "perturbations" of X_n, as long as the perturbations converge in probability.

Lemma 8.3.10. *(Slutsky's Theorem) Let $\{X_n, Y_n : n \in \mathbb{N}\}$ and X be random variables on a probability space $(\Omega, \mathcal{B}, \mathcal{P})$. Let $X_n \xrightarrow{d} X$, and $Y_n \xrightarrow{p} c$ for some $c \in \mathbb{R}$. Then*

(a) $X_n + Y_n \xrightarrow{d} X + c$

(b) $X_n Y_n \xrightarrow{d} cX$

(c) $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$ *if* $c \neq 0$

Proof. We prove only (a); (b) and (c) can be proved similarly. Let $\epsilon > 0$ be given. Write $F_n = F_{X_n+Y_n}$. Choose t such that $t, t-c+\epsilon, t-c-\epsilon$ are all continuity points of F_X. This is possible as there can be at most countably many points of discontinuity of F_X. Now,

$$\begin{aligned} F_n(t) &\leq P(X_n + Y_n \leq t, |Y_n - c| < \epsilon) + P(|Y_n - c| \geq \epsilon) \\ &\leq P(X_n \leq t - c + \epsilon) + P(|Y_n - c| \geq \epsilon) \end{aligned}$$

and

$$F_n(t) \geq P(X_n < t - c - \epsilon) - P(|Y_n - c| \geq \epsilon)$$

The result follows because we have shown that

$$F_X(t-c-\epsilon) \leq \liminf_{n\to\infty} F_n(t) \leq \limsup_{n\to\infty} F_n(t) \leq F_X(t-c+\epsilon).$$

■

Our primary application of Slutsky's Theorem will come in Section 8.5. However, to illustrate its usefulness, we will show that the result in Example 8.3.6 follows immediately using it below.

Example 8.3.11. Recall the t_n distribution from Example 8.1.9. The convergence of the t_n distribution to the Normal $(0,1)$ distribution, proved in Example 8.3.7, would follow by Lemma 8.3.10 (c) if we could show that the sequence $\sqrt{Y_n/n} \xrightarrow{p} 1$, where Y_n is the χ^2_n random variable in the denominator in the definition of the t_n distribution. This is shown in two steps. Either directly applying Chebychev's inequality (Theorem 6.1.13) on $\frac{Y_n}{n}$, or by an application of the Weak Law of Large Numbers (Theorem 8.2.1) we can show that

$$\frac{Y_n}{n} \xrightarrow{p} 1 \text{ as } n \to \infty. \tag{8.3.2}$$

Indeed, as $Y_n = \sum_{i=1}^{n} X_i$ with X_i i.i.d. χ^2_1 random variables, and $E[X_1^2] = 1 < \infty$, it is immediate by Theorem 8.2.1 that (8.3.2) holds. It then follows from Exercise 8.2.7 that $\sqrt{Y_n/n} \xrightarrow{p} 1$. ■

EXERCISES

Ex. 8.3.1. Show (8.3.1).

Ex. 8.3.2. Let $c \in \mathbb{R}$ and $X_1, X_2, \ldots$ be a sequence of random variables. Show that if $X_n \xrightarrow{d} c$ then $X_n \xrightarrow{p} c$.

Ex. 8.3.3. Let $Y_1, Y_2, \ldots$ be a sequence of χ^2_n random variables.

(a) Show that $\frac{Y_n}{n} \xrightarrow{d} 1$ as $n \to \infty$.

(b) Using Exercise 8.3.2 and Lemma 8.3.10 conclude that $t_n \xrightarrow{d} Z$ as $n \to \infty$ with Z being standard normal random variable.

Ex. 8.3.4. Consider a sequence X_n, $n \geq 1$ of random variables such that $X_n \sim \text{Normal}\left(\frac{1}{n}, 1 + \frac{1}{n}\right)$. Show that $X_n \xrightarrow{d} Z$ as $n \to \infty$ where $Z \sim \text{Normal}(0,1)$.

Ex. 8.3.5. Suppose a sequence X_n, $n \geq 1$ of random variables converges to a random variable X in probability. Show that X_n converges in distribution to X. That is, show that

$$F_{X_n}(x) \to F_X(x) \text{ as } n \to \infty$$

for all continuity points of $F_X : \mathbb{R} \to [0,1]$ with F_{X_n}, F_X being the distribution functions of X_n and X respectively.

Ex. 8.3.6. Let X_1, X_2, ... be i.i.d. Uniform$(0,1)$ random variables. Generalize the definition of $M_n = \min(X_1, X_2, \ldots, X_n)$ in Example 8.3.4 as follows: For fixed $k \geq 1$, define $M_{n,k} = X_{(k+1)} - X_{(k)}$.

(a) Show that $nM_{n,k} = n(X_{(k+1)} - X_{(k)})$ also converges to Exponential(1) random variable in distribution

(b) Show that for any fixed $k \geq 1$, $nX_{(k)}$ converges to the Gamma$(k,1)$ distribution.

8.4 CENTRAL LIMIT THEOREM

For $n \geq 1$, let $X_1, X_2, \ldots, X_n$ be an i.i.d. random sample from a distribution X which has mean μ and variance σ^2, but is otherwise unknown. Consider the sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

As observed in Theorem 7.1.4, $E[\overline{X}] = \mu$ and $\mathrm{SD}[\overline{X}] = \frac{\sigma}{\sqrt{n}}$. As discussed before, we might view this information as indicating that with high probability, $\overline{X}$ is typically close to μ up to an error of $\frac{\sigma}{\sqrt{n}}$. As $n \to \infty$, $\frac{\sigma}{\sqrt{n}} \to 0$ and this indicates that $\overline{X}$ approaches μ. We have already verified that $\overline{X}$ converges in probability to μ courtesy of the Weak Law of large numbers, and noted that in fact it converges with probability 1 by the Strong Law of large numbers.

To get a better understanding of the limiting distribution of $\overline{X}$, we standardize it to have mean 0 and variance 1, and consider

$$Y_n = \frac{(\overline{X} - \mu)}{\sigma/\sqrt{n}} = \sqrt{n}\,\frac{(\overline{X} - \mu)}{\sigma}.$$

Without further information about X, the common distribution of $X_1, X_2, \ldots, X_n$, it is not possible to find the exact probabilities of events connected with Y_n. However, it turns out that one can often find good approximate values, because for a large class of possible distributions X, the distribution of Y_n is close to that of the standard Normal random variable, particularly for large n. This remarkable fact is referred to as the Central Limit Theorem and we prove it next.

As done earlier, we shall denote $\overline{X}$ by $\overline{X}_n$ in the statement and proof of the Theorem below to emphasise its dependence on n.

Theorem 8.4.1. (Central Limit Theorem) *Let $X_1, X_2, \ldots$ be i.i.d. random variables with finite mean μ, finite variance σ^2, and possessing common moment generating function $M_X()$. Then*

$$\sqrt{n}\frac{(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} Z, \tag{8.4.1}$$

where $Z \sim$ Normal $(0, 1)$.

Proof. Let $Y_n = \sqrt{n}\,\frac{(\overline{X}_n - \mu)}{\sigma}$. We will verify that

$$\lim_{n\to\infty} M_{Y_n}(t) = e^{\frac{t^2}{2}}.$$

Now, using the definition of the moment generating function and some elementary algebra we have

$$\begin{aligned} M_{Y_n}(t) &= E\left[\exp(tY_n))\right] = E\left[\exp\left(t\sqrt{n}\frac{(\overline{X}_n-\mu)}{\sigma}\right)\right] \\ &= E\left[\exp\left(\frac{t}{\sigma}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right)\right)\right] = E\left[\exp\left(\sum_{i=1}^{n}\frac{t}{\sigma\sqrt{n}}(X_i-\mu)\right)\right] \\ &= E\left[\prod_{i=1}^{n}\exp\left(\frac{t}{\sigma\sqrt{n}}(X_i-\mu)\right)\right]. \end{aligned} \tag{8.4.2}$$

As $X_1, X_2, \ldots, X_n$ are independent, we can conclude using Theorem 8.1.2 that

$$\exp\left(\frac{t}{\sigma\sqrt{n}}(X_1-\mu)\right), \exp\left(\frac{t}{\sigma\sqrt{n}}(X_2-\mu)\right), \ldots, \exp\left(\frac{t}{\sigma\sqrt{n}}(X_n-\mu)\right)$$

are also independent. From Exercise 7.1.3 and 7.1.4, they also have the same distribution. So from the calculation in (8.4.2) and using Exercise 6.3.4 inductively we have

$$\begin{aligned} M_{Y_n}(t) = E\left[\prod_{i=1}^{n}\exp\left(\frac{t}{\sigma\sqrt{n}}(X_i-\mu)\right)\right] &= \prod_{i=1}^{n}E\left[\exp\left(\frac{t}{\sigma\sqrt{n}}(X_i-\mu)\right)\right] \\ \text{(Using Theorem 6.3.9}(a)\text{)} \\ &= \left(E\left[\exp\left(\frac{t}{\sqrt{n}}\frac{(X-\mu)}{\sigma}\right)\right]\right)^n. \end{aligned} \tag{8.4.3}$$

where X is the common distribution of $X_1, X_2, \ldots$. In other words, $M_{Y_n}(t) = \left(M_U\left(\frac{t}{\sqrt{n}}\right)\right)^n$, where $U = \frac{X-\mu}{\sigma}$. As $E[U] = 0, E[U^2] = 1$ we have that $M'_U(0) = 0$ and $M''_U(0) = 1$. From Exercise 8.4.5, we have that for $t \in \mathbb{R}$

$$M_U(t) = 1 + \frac{t^2}{2} + g(t), \tag{8.4.4}$$

where g satisfies $\lim_{s\to 0}\frac{g(s)}{s^2} = 0$. Thus, we have

$$M_{Y_n}(t)) = \left(M_U\left(\frac{t}{\sqrt{n}}\right)\right)^n = \left[1+\frac{t^2}{2n}+g\left(\frac{t}{\sqrt{n}}\right)\right]^n = \left[1+\frac{1}{n}\left(\frac{t^2}{2}+ng\left(\frac{t}{\sqrt{n}}\right)\right)\right]^n.$$

Using the fact that for any fixed t, $\frac{t^2}{2} + ng\left(\frac{t}{\sqrt{n}}\right) \to \frac{t^2}{2}$ as $n \to \infty$ and Exercise 8.4.4 it follows that,

$$\lim_{n\to\infty} M_{Y_n}(t) = e^{\frac{t^2}{2}}.$$

Theorem 8.3.8 then implies the result as the limit $e^{\frac{t^2}{2}}$ is the moment generating function of the standard Normal distribution. ■

Remark 8.4.2. The existence of moment generating function is not essential for the Central Limit Theorem, and (8.4.1) holds as long as $X_1, X_2, \ldots$ are i.i.d. random variables with finite mean μ and finite variance σ^2. However, the proof of this more general statement is more complicated.

Remark 8.4.3. An equivalent formulation of the Central Limit Theorem is often useful. By definition of $\overline{X}_n$ and elementary algebra we see that $Y_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$, where $S_n = \sum_{i=1}^{n} X_i$. It follows that

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} \text{Normal}\,(0,1)\,. \tag{8.4.5}$$

Remark 8.4.4. The Central Limit Theorem is a remarkable result. But it perhaps bears emphasis that the remarkable part of the result is not the specific statistic, the sample mean, but rather the Normality of the limiting distribution, which arises in many other situations as well. Although most such results are beyond the scope of this book, we show later in this chapter that the sample median, when suitably standardized, also converges to the standard Normal distribution under fairly general conditions.

The Central Limit Theorem for the sample median can be viewed as a refinement of the Weak Law of large numbers. The weak law tells us that the sample mean $\overline{X}$ converges to the expectation μ as $n \to \infty$. However, for any finite n, $\overline{X}$ is still a non-constant random variable, whose distribution we may be interested in. This distribution can be quite complicated in general. The Central Limit Theorem is remarkable because it says that regardless of the underlying distribution, probabilities concerning the sample mean $\overline{X}$ can be well approximated by standard Normal probabilities for large n.

Before looking at uses of such approximations, let us consider the factors that might affect the quality of the approximation. The Central Limit Theorem does not say anything about how well the approximation will be for any given n, but we can guess that it will be better for larger n, and also depend on the distribution giving rise to the data.

Example 8.4.5. As we have seen earlier, an important application of the Weak Law is to estimate probabilities of events by sample proportion. Here the underlying distribution is Bernoulli(p), with the probability p estimated by the sample proportion $\hat{p}_n = S_n/n$. Suppose $X_1, \ldots, X_n$ are independent Bernoulli(p) random variables. Then $S_n \sim \text{Binomial}(n, p)$ and

$$\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} := \frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} Z,$$

where Z is standard Normal. Let us see how the quality of this approximation changes with the choice of n and p.

Instead of simulating X_i-s individually, we can simulate S_n directly using the `rbinom()` function. For a specific choice of p and n, we could simulate standardized S_n values as follows.

```
p <- 0.5
n <- 25
s <- rbinom(1000, size = n, prob = 0.5)
z <- (s - n * p) / sqrt(n * p * (1-p))
mean(z)
```

```
[1] 0.0052
```

```
sd(z)
```

```
[1] 0.9830459
```

The mean and standard deviation of the sample proportion, computed over these 1000 replication, matches what we expect. To see how similar their overall distribution is to the standard Normal distribution, the top panel in Figure 8.4 shows empirical frequency distribution plots obtained from 1000 replications for $p = 0.5$ and $n = 10, 25, 50$, and 100. Similar plots for $p = 0.25$ and $p = 0.05$ shown in the middle and bottom panels. The Normal approximation obtained using the Central Limit Theorem are added for comparison. As the sample spaces differ substantially depending on n, the quantity plotted on the y-axis is not the relative frequency but rather a scaled version, similar to the scaling done in histograms, that makes the scaled quantities comparable with each other and the Normal density. From these plots, we can conclude that the distribution of Binomial proportion is well approximated by Normal when p is close to $\frac{1}{2}$, although for smaller sample sizes the number of ties can become an issue as well. Values of p away from $\frac{1}{2}$ can generate skewed (asymmetric) distributions for which the Normal is not a good approximation. A general convention often used is to consider the approximation valid if both np and $n(1-p)$ are at least 5.

As we saw in Chapter 7, Q-Q plots are often more useful for assessing departure from Normality. Figure 8.5 shows Normal Q-Q plots that are analogous to the empirical frequency distribution plots in Figure 8.4. Each plot represents 1000 replications, for $p = 0.5, 0.25, 0.05$ and $n = 10, 25, 50, 100$. These largely confirm what we already saw from the empirical frequency distribution plots, and suggest, in particular, that the Normal approximation may be unreliable when p is close to 0 or 1. ■

Example 8.4.6. The Central Limit Theorem applies not just to sample proportions but to general discrete and continuous distributions if they have finite expectation and variance. For continuous distributions, ties happen with probability 0, so empirical frequency distribution plots are not useful. We can use histograms as an alternative, but Q-Q plots are more useful when the primary goal is to compare with a Normal distribution.

In Figure 8.6, we show Q-Q plots similar to those in Figure 8.5, but instead of sample proportion, we consider means of random samples from three continuous distributions, namely Uniform$(0, 1)$, Exp(1), and Cauchy, with sample sizes $n = 5, 20, 50, 100$. These plots suggest that even with a shape very different from Normal, the distribution of the Uniform sample mean is well approximated by a Normal distribution even for small n. For the heavily asymmetric Exponential distribution,

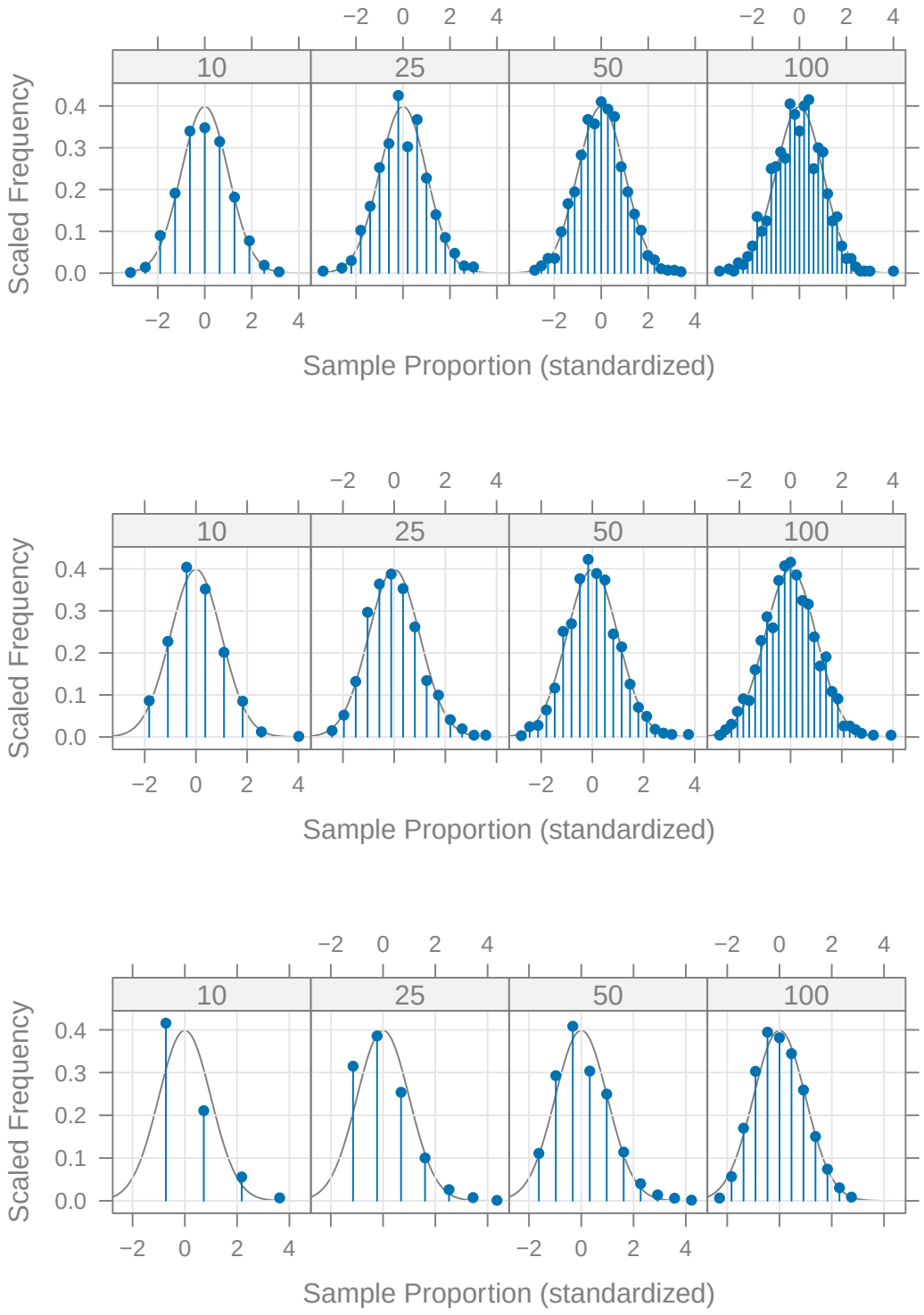


Figure 8.4: Empirical frequency distribution plots of standardized sample proportions when true probability is (top) $p = 0.5$, (middle) $p = 0.25$, and (bottom) $p = 0.05$. In each case, the standard Normal density has been added for comparison. To account for the different sample spaces, the frequencies plotted on the y-axis have been scaled to make them comparable with each other and the Normal density.

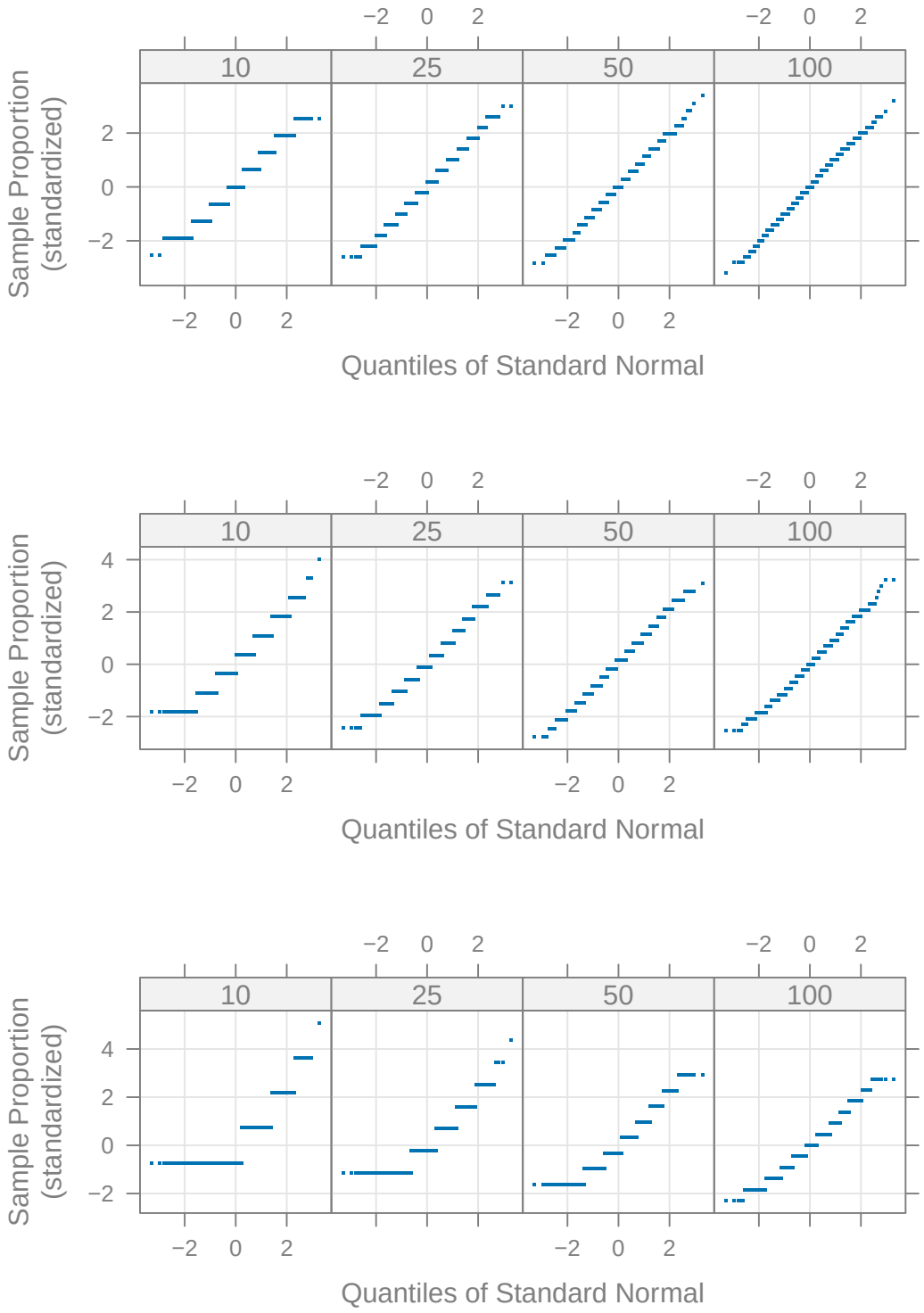


Figure 8.5: Q-Q plot of standardized sample proportions when true probability is (top) $p = 0.5$, (middle) $p = 0.25$, and (bottom) $p = 0.05$.

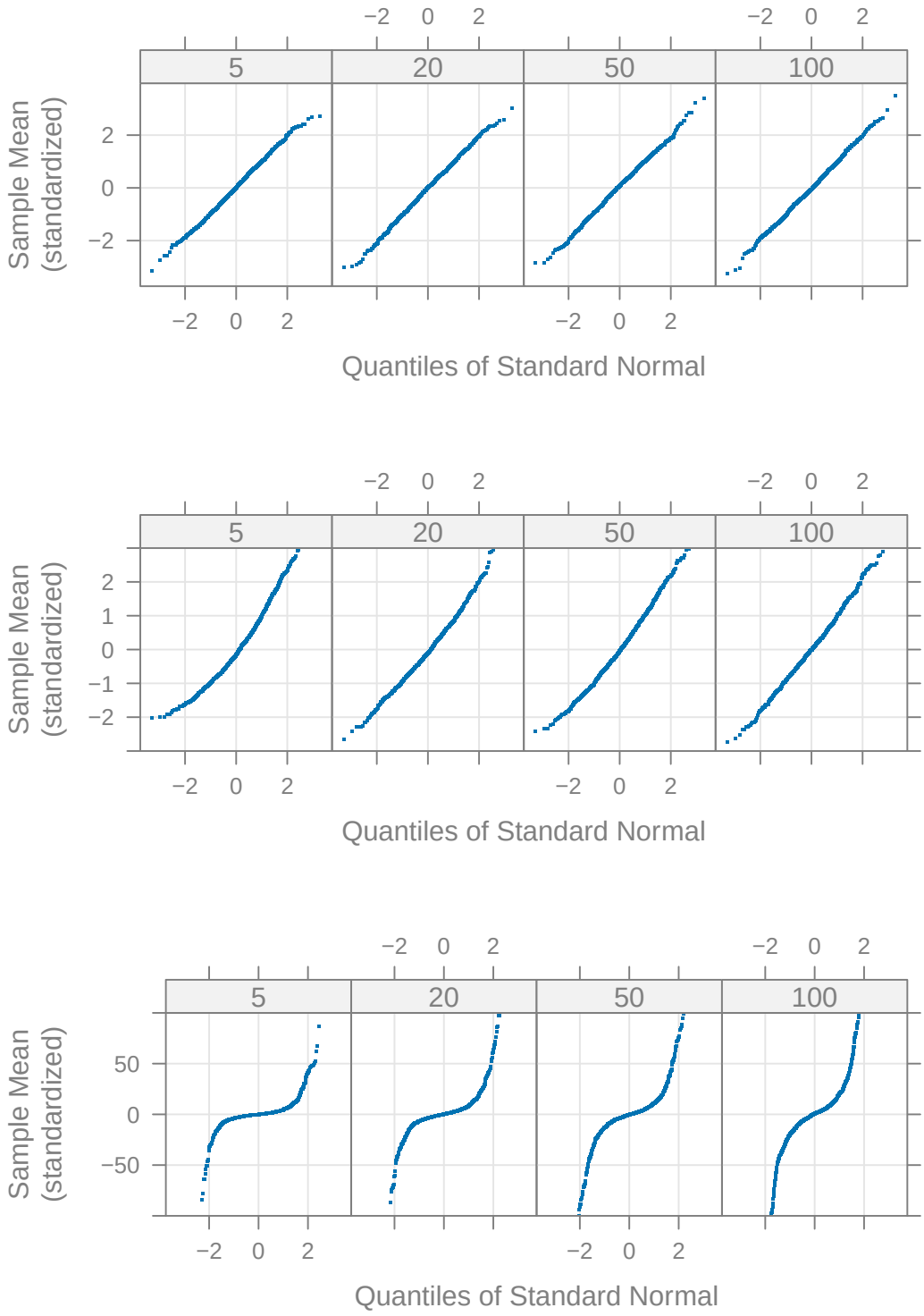


Figure 8.6: Q-Q plot of standardized sample mean of random sample from (top) Uniform$(0, 1)$, (middle) Exponential(1), and (bottom) Cauchy.

this convergence requires a larger sample size. For the Cauchy distribution, which does not have finite mean, the Central Limit Theorem does not hold at all. ■

Before moving on, let us summarize the main conclusions from the last two examples. Although the sample mean converges to the population mean, the convergence is not necessarily immediate. Thus, although we can expect that for large n the sample proportion or sample mean will be "close" to the population proportion or mean, we cannot expect it to be exactly the same. The Central Limit Theorem assures us that under fairly mild assumptions, the difference will behave like a Normal random variable. As we see in the next chapter, this knowledge allows us to make useful statements about the population proportion or mean, when it is unknown, based on what we observe.

8.4.1 *Normal Approximation*

A typical application of the Central Limit Theorem is to find approximate value of the probability of events related to S_n or $\overline{X}$. For instance, suppose we were interested in calculating for any $a, b \in \mathbb{R}$, $P(a < S_n \leq b)$ for large n. We would proceed in the following way. We know from (8.4.5) that

$$P\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \leq x\right) \to P(Z \leq x) \tag{8.4.6}$$

as $n \to \infty$ for all $x \in \mathbb{R}$.

$$\begin{aligned} P(a < S_n \leq b) &= P\left(\frac{a - n\mu}{\sqrt{n}\sigma} < \frac{S_n - n\mu}{\sqrt{n}\sigma} \leq \frac{b - n\mu}{\sqrt{n}\sigma}\right) \\ &= P(\frac{S_n - n\mu}{\sqrt{n}\sigma} \leq \frac{b - n\mu}{\sqrt{n}\sigma}) - P(\frac{S_n - n\mu}{\sqrt{n}\sigma} \leq \frac{a - n\mu}{\sqrt{n}\sigma}) \\ &\quad \text{from (8.4.6) for large enough } n \\ &\approx P(Z \leq \frac{b - n\mu}{\sqrt{n}\sigma}) - P(Z \leq \frac{a - n\mu}{\sqrt{n}\sigma}) \\ &= P(\frac{a - n\mu}{\sqrt{n}\sigma} < Z \leq \frac{b - n\mu}{\sqrt{n}\sigma}), \end{aligned}$$

where in the second last line we have used the notation $\approx$ to indicate that the right hand side is an approximation. Therefore we would conclude that for large n,

$$P(a < S_n \leq b) \approx P\left(\frac{a - n\mu}{\sqrt{n}\sigma} < Z \leq \frac{b - n\mu}{\sqrt{n}\sigma}\right). \tag{8.4.7}$$

We would then use the R function `pnorm()` or Normal Tables (See Table B.1) to compute the right hand side. A similar computation would also yield

$$P\left(a < \overline{X} \leq b\right) \approx P\left(\frac{\sqrt{n}(a - \mu)}{\sigma} < Z \leq \frac{\sqrt{n}(b - \mu)}{\sigma}\right). \tag{8.4.8}$$

Example 8.4.7. Let Y be a random variable distributed as Gamma$(100, 4)$. Suppose we were interested in finding $P(20 < Y \leq 30)$. Suppose $X_1, X_2, \ldots, X_{100}$ are independent Exponential (4)

random variables then Y and $S_{100} = \sum_{i=1}^{100} X_i$ have the same distribution. Therefore, applying the Central Limit Theorem with $\mu = E[X_1] = \frac{1}{4}, \sigma = \text{SD}[X_1] = \frac{1}{4}$, we have

$$\begin{aligned} P(20 < Y \leq 30) &= P(20 < S_{100} \leq 30) \\ &\approx P\left(\frac{20 - 100(0.25)}{\sqrt{100}(0.25)} < Z \leq \frac{30 - 100(0.25)}{\sqrt{100}(0.25)}\right) \text{ by (8.4.7)} \\ &= P(\frac{-5}{2.5} < Z \leq \frac{5}{2.5}) \\ &= P(-2 < Z \leq 2) \\ &= P(Z \leq 2) - P(Z \leq -2) \\ &\quad \text{using symmetry of Normal distribution} \\ &= P(Z \leq 2) - (1 - P(Z \leq 2)) \\ &= 2P(Z \leq 2) - 1 \end{aligned}$$

Looking up Table B.1, we see that this value comes out to be approximately $2 \times 0.9772 - 1 = 0.9544$. A more precise answer is given by R as

```
2 * pnorm(2) - 1
```

```
[1] 0.9544997
```

Using R, we can also compare this with the exact probability that we are approximating.

```
pgamma(30, 100, 4) - pgamma(20, 100, 4)
```

```
[1] 0.9550279
```

In this example, the approximation is correct to three decimal places. ■

8.4.2 *Continuity Correction*

Suppose $X_1, X_2, X_3, \ldots$ are all integer valued random variables. Then $S_n = \sum_{i=1}^{n} X_i$ is also an integer valued random variable. Now, for any integer k, $P(S_n \leq k) = P(S_n \leq k + h)$ for all $0 < h < 1$. However it is easy to see that two distinct values of h will lead to two different answers if we use the Normal approximation provided by the Central Limit Theorem. It is customary to use $h = \frac{1}{2}$ when computing such probabilities using the Normal approximation, as

$$\begin{aligned} P(S_n \leq a) &= P(S_n \leq a + 0.5) \\ &\approx P\left(Z \leq \frac{a + 0.5 - n\mu}{\sqrt{n}\sigma}\right) \end{aligned} \tag{8.4.9}$$

whenever a is a possible value of S_n. This convention is referred to as the "continuity correction".

Example 8.4.8. Two types of coin are produced at a factory: a fair coin and a biased one that comes up heads 55% of the time. Priya is the quality control scientist at the factory. She wants to design an experiment that will test whether a coin is fair or biased. In order to ascertain which type of coin she has, she prescribes the following experiment as a test: *Toss the given coin 1000 times, if the coin comes up heads 525 or more times conclude that it is a biased coin. Otherwise conclude that it is fair.* Factory manager Ayesha is interested in the following question: What is the probability that Priya's test shall reach a false conclusion for a fair coin ?

Let S_{1000} be the number of heads in 1000 tosses of a coin. As discussed in earlier chapters, we know that $S_{1000} = \sum_{i=1}^{1000} X_i$ where each X_i are i.i.d. Bernoulli random variables with parameter p. If the coin is fair, then $p = 0.5$ and $E[X_1] = 0.5, Var[X_1] = 0.25$, and therefore $E[S_{1000}] = 500$ and $SD[S_{1000}] = \sqrt{250} = 15.8114$. We want to approximate

$$P(S_{1000} \geq 525) = 1 - P(S_{1000} \leq 524) = 1 - P(S_{1000} \leq 524.5)$$

Without the continuity correction, we would approximate this probability as

$$1 - P\left(Z \leq \frac{24}{15.8114}\right) = 1 - P\left(Z \leq 1.52\right)$$

which can be computed using Table B.1 as $1 - 0.9357 = 0.0643$, or using R as

```
1 - pnorm(24 / sqrt(250))
```

```
[1] 0.06452065
```

With the continuity correction, the approximation would instead use $z = 24.5/15.8114 = 1.55$, giving $1 - 0.9394 = 0.0606$ using Table B.1 or

```
1 - pnorm(24.5 / sqrt(250))
```

```
[1] 0.06062886
```

in R. We can also compute the exact probability that we are trying to approximate, namely $P(S_{1000} \geq 525)$, in R as

```
1 - pbinom(524, 1000, 0.5)
```

```
[1] 0.06060713
```

As we can see, the continuity correction gives us a slightly better approximation. These calculations tell us that the probability of Priya's test reaching a false conclusion if the coin is fair is approximately 0.061. We shall examine the topic of Hypothesis testing, which is what Priya was trying to do, in more detail in Chapter 10. ■

Example 8.4.9. We return to the Birthday problem. Suppose a small town has 1460 students. What is the probability that five or more students were born on independence day ? Assume that birthrates are constant throughout the year and that each year has 365 days.

The probability that any given student was born on independence day is $\frac{1}{365}$. So the exact probability that five or more students were born on independence day is

$$1-\sum_{k=0}^{4}\binom{1460}{k}\left(\frac{1}{365}\right)^{k}\left(\frac{364}{365}\right)^{1460-k}.$$

In Example 2.2.1 we have used the Poisson approximation with $\lambda = 4$ to estimate the above as

$$\begin{aligned}
&1-\sum_{k=0}^{4}\binom{1460}{k}(\tfrac{1}{365})^{k}(\tfrac{364}{365})^{1460-k}\\
&\approx \quad 1-\left[e^{-4}+4e^{-4}+\frac{4^2}{2}e^{-4}+\frac{1}{6}4^3e^{-4}+\frac{1}{24}4^4e^{-4}\right]\\
&= \quad 0.3711631
\end{aligned}$$

We can do another approximation using Central Limit Theorem, which is typically called the Normal approximation. For $1 \le i \le 1460$, define

$$X_i = \begin{cases} 1 & \text{if } i\text{-th person's birthday is on independence day} \\ 0 & \text{otherwise} \end{cases}$$

Given the assumptions above on birthrates we know X_i are i.i.d. random variables distributed as Bernoulli$(\frac{1}{365})$. Note that $S_{1460} = \sum_{i=1}^{1460} X_i$ is the number of people born on independence day and we are interested in calculating

$$P(S_{1460} \ge 5).$$

Observe that $E(X_1) = \frac{1}{365}, \mathrm{Var}(X_1) = \frac{1}{365}(1-\frac{1}{365}) = \frac{364}{365^2}$. By the Central Limit Theorem, we know that

$$\begin{aligned}
P(S_{1460} \ge 5) &= 1 - P(S_{1460} \le 4) = 1 - P(S_{1460} \le 4.5)\\
&\approx 1 - P\left(Z \le \frac{4.5-(1460)(\frac{1}{365})}{\sqrt{(1460)(\frac{364}{365^2})}}\right)\\
&= 1 - P(Z \le \frac{0.5}{1.9973})\\
&= 0.401.
\end{aligned}$$

Recall from the calculations done in Example 2.2.1 that the exact answer for this problem is 0.3711629. So in this example, the Poisson approximation seems to work better then the Normal approximation. This is due to the fact that more asymmetry in the underlying Bernoulli distribution

worsens the normal approximation, just as it improves the Poisson approximation as we saw in Figure 2.2. ■

EXERCISES

Ex. 8.4.1. Suppose S_n is binomially distributed with parameters $n = 200$ and $p = 0.3$ Use the Central Limit Theorem to find an approximation for $P(99 \leq S_n \leq 101)$.

Ex. 8.4.2. Toss a fair coin 400 times. Use the Central Limit Theorem to

(a) find an approximation for the probability of at most 190 heads.

(b) find an approximation for the probability of at least 70 heads.

(c) find an approximation for the probability of at least 120 heads.

(d) find an approximation for the probability that the number of heads is between 140 and least 160.

Ex. 8.4.3. Suppose that the weight of open packets of daal in a home is uniformly distributed from 200 to 600 gms. In random survey of 64 homes, find the (approximate) probability that the total weight of open boxes is less than 25 kgs.

Ex. 8.4.4. Let $\{a_n\}_{n\geq 1}$ be a sequence of real numbers such that $a_n \to a$ as $n \to \infty$. Show that

$$\lim_{n\to\infty} \left(1 + \frac{a_n}{n}\right)^n = e^a.$$

Ex. 8.4.5. Suppose U is a random variable (discrete or continuous) and $M_U(t) = E(e^{tU})$ exists for all t. Show that

$$M_U(t) = 1 + tM'_U(0) + \frac{t^2}{2}M''_U(0) + g(t),$$

where $\lim_{t\to 0} \frac{g(t)}{t^2} = 0$.

Ex. 8.4.6. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with $X_1 \sim \text{Exp}(1)$. Find

$$\lim_{n\to\infty} P\left(\frac{n}{2} - \frac{\sqrt{n}}{2\sqrt{3}} \leq \sum_{i=1}^{n}[1 - \exp(-X_i)] \leq \frac{n}{2} + \frac{\sqrt{n}}{2\sqrt{3}}\right).$$

Ex. 8.4.7. Let $a_n = \sum_{k=0}^{n} \frac{n^k}{k!}e^{-n}$, $n \geq 1$. Using the Central Limit Theorem, evaluate $\lim_{n\to\infty} a_n$.

Ex. 8.4.8. How many times should you toss a coin:

(a) to be at least 90% sure that your estimate of the P(head) is within 0.1 of its true value ?

(b) to be at least 90% sure that your estimate of the P(head) is within 0.01 of its true value ?

Ex. 8.4.9. To forecast the outcome of the election in which two parties are contesting, an internet poll via Facebook is conducted. How many people should be surveyed to be at least 95% sure that the estimated proportion is within 0.05 of the true value?

Ex. 8.4.10. A medical study is conducted to estimate the proportion of people suffering from April allergies in Bangalore. How many people should be surveyed to be at least 99% sure that the estimate is within 0.02 of the true value?

8.5 DELTA METHOD

In many situations one is interested in knowing whether convergence properties are preserved under transformations. Given random variables $X_1, X_2, \dots$ and Z such that $X_n \xrightarrow{d} Z$, and a function $g : \mathbb{R} \to \mathbb{R}$, we may be interested in knowing the limiting distribution of $g(X_n)$. In earlier chapters, we have learnt techniques to calculate the distribution of $g(X)$ from the distribution of X (see Section 3.3, Section 5.3), which may be helpful in studying this problem. In this section, we discuss the Delta method, which answers this question in a specific situation, where g is a smooth transformation that can be effectively approximated by a linear function in the region of interest. Slutsky's theorem (Lemma 8.3.10) is an important tool in proving the following result.

Theorem 8.5.1. *Let $\mu \in \mathbb{R}$ and $\sigma \neq 0$. Let $g : \mathbb{R} \to \mathbb{R}$ be differentiable at μ, with $g'(\mu) \neq 0$ and $g'(\cdot)$ continuous in a neighbourhood of μ. Suppose $Z \sim$ Normal$(0,1)$ and $X_1, X_2, \dots$ is a sequence of random variables such that*

$$\sqrt{n}\,\frac{(X_n - \mu)}{\sigma} \xrightarrow{d} Z \quad \text{as } n \to \infty.$$

Then

$$\sqrt{n}\,\frac{(g(X_n) - g(\mu))}{\sigma g'(\mu)} \xrightarrow{d} Z \quad \text{as } n \to \infty.$$

Proof. By the fundamental theorem of integral calculus, we have

$$g(x) - g(\mu) = \int_{\mu}^{x} g'(t)\,\mathrm{d}t = (x - \mu)\int_{0}^{1} g'(\mu + s(x - \mu))\,\mathrm{d}s.$$

For $n \geq 1$, using the above, we have

$$\sqrt{n}\,\frac{(g(X_n) - g(\mu))}{\sigma g'(\mu)} = \sqrt{n}\,\frac{(X_n - \mu)}{\sigma} \cdot \frac{1}{g'(\mu)} \int_{0}^{1} g'(\mu + s(X_n - \mu))\,\mathrm{d}s. \tag{8.5.1}$$

By our hypothesis on X_n, we know that $\sqrt{n}\,\frac{(X_n - \mu)}{\sigma} \xrightarrow{d} Z$ as $n \to \infty$. Slutsky's theorem (Lemma 8.3.10) will imply the result if we can show that

$$\frac{1}{g'(\mu)} \int_{0}^{1} g'(\mu + s(X_n - \mu))\,\mathrm{d}s \xrightarrow{p} 1 \quad \text{as } n \to \infty. \tag{8.5.2}$$

Let $\epsilon > 0$ be given. Then

$$\begin{aligned} & P\left(\left|\frac{1}{g'(\mu)}\int_0^1 g'(\mu + s(X_n - \mu))\,\mathrm{d}s - 1\right| > \epsilon\right) \\ = \;& P\left(\left|\int_0^1 g'(\mu + s(X_n - \mu))\,\mathrm{d}s - g'(\mu)\right| > |g'(\mu)|\,\epsilon\right) \\ \leq \;& P\left(\sup_{s\in[0,1]} |g'(\mu + s(X_n - \mu)) - g'(\mu)| > |g'(\mu)|\,\epsilon\right) \end{aligned} \tag{8.5.3}$$

As g' is continuous in a neighbourhood of μ there is a $\delta > 0$ such that for $|x - \mu| < \delta$ we have

$$|g'(x) - g'(\mu)| < |g'(\mu)|\epsilon. \tag{8.5.4}$$

Using (8.5.4) in (8.5.3) we have for all $n \geq 1$,

$$P\left(\left|\frac{1}{g'(\mu)}\int_0^1 g'(\mu + s(X_n - \mu))\,\mathrm{d}s - 1\right| > \epsilon\right) \leq P\left(|X_n - \mu| > \delta\right). \tag{8.5.5}$$

Let $M > 0$ be such that

$$P(|Z| > M) < \epsilon. \tag{8.5.6}$$

Let $N \geq 1$ be such that for all $n \geq N$ we have

$$\frac{\delta\sqrt{n}}{\sigma} < M \qquad \text{and} \qquad P\left(\left|\frac{\sqrt{n}(X_n - \mu)}{\sigma}\right| > M\right) < P(|Z| > M) + \epsilon. \tag{8.5.7}$$

Using (8.5.6) and (8.5.7) we have for all $n \geq N$

$$\begin{aligned} P\left(|X_n - \mu| > \delta\right) &= P\left(\left|\sqrt{n}\,\frac{(X_n - \mu)}{\sigma}\right| > \frac{\delta\sqrt{n}}{|\sigma|}\right) \\ &\leq P\left(\left|\frac{\sqrt{n}(X_n - \mu)}{\sigma}\right| > M\right) \\ &< P(|Z| > M) + \epsilon \\ &< 2\epsilon. \end{aligned} \tag{8.5.8}$$

Therefore substituting (8.5.8) in (8.5.5) we have for all $n \geq N$,

$$P\left(\left|\frac{1}{g'(\mu)}\int_0^1 g'(\mu + s(X_n - \mu))\,\mathrm{d}s - 1\right| > \epsilon\right) < 2\epsilon.$$

Thus we have proved (8.5.2) ■

Remark 8.5.2. The particular transformation g affects the rate of convergence to normality of the sequence $g(X_n)$. Indeed Theorem 8.5.1 shows that

$$\text{if } \sqrt{n}(X_n - \mu) \xrightarrow{d} \text{Normal}\left(0, \sigma^2\right), \text{ then } \sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} \text{Normal}\left(0, \sigma^2 (g'(\mu))^2\right)$$

The value of $g'(\mu)$ determines how large or small the variance of the limiting normal distribution is.

Example 8.5.3. Suppose $\sqrt{n}(X_n - \mu) \xrightarrow{d} \text{Normal}\left(0, \sigma^2\right)$ as $n \to \infty$ for some $\mu \neq 0$ and $\sigma > 0$.

(a) If $g : \mathbb{R} \to \mathbb{R}$ is given by $g(x) = x^2$ then by Theorem 8.5.1 we have that

$$\sqrt{n}(X_n^2 - \mu^2) \xrightarrow{d} \text{Normal}\left(0, 4\sigma^2\mu^2\right) \quad \text{as } n \to \infty.$$

(b) If $g : \mathbb{R} \to \mathbb{R}$ is given by $g(x) = \frac{1}{x}$ for $x \neq 0$ and $g(0) = 0$.then by Theorem 8.5.1 we have that

$$\sqrt{n}\left(\frac{1}{X_n} - \frac{1}{\mu}\right) \xrightarrow{d} \text{Normal}\left(0, \frac{\sigma^2}{\mu^4}\right) \quad \text{as } n \to \infty.$$

Note that the transformed random variables $g(X_n)$ need not have finite expectation for any $n \geq 1$, e.g., take $X_n \sim \text{Normal}\left(\mu, \frac{\sigma}{n}\right)$ with $g(X_n) = \frac{1}{X_n}$ as in (b). ■

8.5.1 *Variance Stabilizing Transformation*

A natural application of the Weak Law of Large Numbers and the Central Limit Theorem is to estimate the unknown mean μ of a population by the sample mean $\overline{X}_n$, provided we have a random sample from the population. Here the assumption is that $E[X_i] = \mu$ and $Var[X_i] = \sigma^2$ for each $1 \leq i \leq n$, so by the Central Limit Theorem,

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \text{Normal}\left(0, \sigma^2\right).$$

In many applications, $\sigma^2 \equiv \sigma^2(\mu)$ is a function of μ, but calculations become more convenient if the variance of the limiting normal distribution does not depend on the parameter of interest μ. This can often be achieved by carefully choosing a transformation $g(\overline{X}_n)$ and applying the Delta method. More precisely, suppose we find a $g : \mathbb{R} \to \mathbb{R}$ such that $g'(\mu)\sigma(\mu) = c$ for some $c \in \mathbb{R}$ then Theorem 8.5.1 will imply that

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) \xrightarrow{d} \text{Normal}\left(0, c^2\right).$$

Such a transformation is called a *variance stabilizing transformation*. In the following example we present three applications.

Example 8.5.4. Suppose $X_1, X_2, X_3, \ldots$ are i.i.d. X with $E[X] = p$ and $\text{Var}[X] = \sigma^2$. Let $\hat{p} = \overline{X}_n$.

(a) Suppose $X \sim \text{Bernoulli}(p)$. Then $\sigma^2 = p(1-p)$, and we need a transformation such that $g'(p)\sqrt{p(1-p)} = c$ for some $c \in \mathbb{R}$. Indeed if $g(x) = \arcsin(\sqrt{x})$ then

$$\sqrt{n}(g(\overline{X}_n) - g(p)) \xrightarrow{d} \text{Normal}\left(0, \frac{1}{4}\right).$$

(b) Suppose $X \sim \text{Poisson}(p)$. Then $\sigma^2 = p$, and we need a transformation such that $g'(p)\sqrt{p} = c$ for some $c \in \mathbb{R}$. Indeed if $g(x) = \sqrt{x}$ then

$$\sqrt{n}(g(\overline{X}_n) - g(p)) \xrightarrow{d} \text{Normal}\left(0, \frac{1}{4}\right).$$

(c) Suppose $X \sim \text{Normal}\left(0, \sigma^2\right)$ and we are interested in estimating σ^2. We will then use $\frac{1}{n}\sum_{i=1}^{n} X_i^2$ as the estimate for σ^2 and the Central Limit Theorem implies that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \sigma^2\right) \xrightarrow{d} \text{Normal}\left(0, 2\sigma^4\right).$$

Thus we need a transformation such that $g'(\sigma^2)\sqrt{2}\sigma = c$ for some $c \in \mathbb{R}$. Indeed if $g(x) = \log x$ then

$$\sqrt{n}\left(g\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - g(\sigma^2)\right) \xrightarrow{d} \text{N}(0, 2).$$

■

EXERCISES

Ex. 8.5.1. Prove Lemma 8.3.10 (a) and (b).

Ex. 8.5.2. Supppose $\{X_n\}_{n\geq 1}, X$ are a sequence of random variables and $\{a_n\}_{n\geq 1}, a$ are a sequence of real numbers such that $X_n \xrightarrow{d} X$ and $a_n \to a$ then show that $a_n X_n \xrightarrow{d} aX$. *Hint: Use Lemma 8.3.10.*

Ex. 8.5.3. Supppose $\{X_n\}_{n\geq 1}, X$ and $\{Y_n\}_{n\geq 1}, Y$ are a sequence of random variables such that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$. Show that $\lambda X_n + (1-\lambda)Y_n \xrightarrow{d} \lambda X + (1-\lambda)Y$. *Hint: Use Lemma 8.3.10.*

Ex. 8.5.4. Suppose $X_n \xrightarrow{d} X$. Show that $X_n^2 \xrightarrow{d} X^2$.

Ex. 8.5.5. Let $\alpha, \mu > 0$. Let $\{X_i\}_{i\geq 1}$ be i.i.d. random variables following Pareto (α, μ) distribution. That is, the probability density function of X_i, for any $i \geq 1$, is given by

$$f_{(\alpha,\mu)}(x) = \begin{cases} \frac{\alpha\mu^\alpha}{x^{\alpha+1}} & x \geq \mu, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Let $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{X_i}{\mu}\right)$. Show that

$$\sqrt{n}\left(\frac{1}{\overline{Y}_n} - \alpha\right) \xrightarrow{d} \text{Normal}\left(0, \alpha^2\right).$$

(b) Let $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n} \log(X_i)$ and $M_n = \max\{X_1, X_2, \ldots, X_n\}$.

(i) Show that $\sqrt{n}(\log(M_n) - \log(\mu)) \xrightarrow{p} 0$.

(ii) Using (a) and Lemma 8.3.10 show that

$$\sqrt{n}\left(\overline{Z}_n - \log(M_n) - \frac{1}{\log(\mu)}\right) \xrightarrow{d} \text{Normal}\left(0, \frac{1}{\alpha^2}\right).$$

(iii) Show that

$$\sqrt{n}\left(\frac{1}{\overline{Z}_n - \log(M_n)} - \alpha\right) \xrightarrow{d} \text{Normal}\left(0, \alpha^2\right).$$

Ex. 8.5.6. For $\alpha, \mu > 0$, let $\{X_i\}_{i\geq 1}$ be i.i.d. random variables with the probability density function of X_i, for any $i \geq 1$, given by

$$f_{(\alpha,\mu)}(x) = \begin{cases} \mu e^{-\mu(x-\alpha)} & x \geq \alpha, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $M_n = \max\{X_1, X_2, \ldots, X_n\}$.

(a) Show that $E[X_1] = \alpha + \frac{1}{\mu}$ and $\text{Var}[X_1] = \frac{1}{\mu^2}$.

(b) Show that $\sqrt{n}(M_n - \alpha) \xrightarrow{p} 0$.

(c) Show that $\sqrt{n}(\overline{X}_n - M_n - \frac{1}{\mu}) \xrightarrow{d} \text{Normal}\left(0, \frac{1}{\mu^2}\right)$.

(d) Show that $\sqrt{n}(\frac{1}{\overline{X}_n - M_n - \mu}) \xrightarrow{d} \text{Normal}\left(0, \mu^2\right)$.

Ex. 8.5.7. Let $X_i, i \geq 1$ be i.i.d. Bernoulli(p) random variables. Let $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$. Show that

$$\sqrt{n}\left(\frac{\overline{X}_n}{1-\overline{X}_n} - \frac{p}{1-p}\right) \xrightarrow{d} \text{Normal}\left(0, \frac{p}{(1-p)^3}\right).$$

The statistic $\frac{\overline{X}_n}{1-\overline{X}_n}$ is typically used to estimate the odds ratio $\frac{p}{1-p}$.

Ex. 8.5.8. Let $X_i, i \geq 1$ be i.i.d. Bernoulli(p) random variables. Let $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$. Show that

$$\sqrt{n}\left(\overline{X}_n(1-\overline{X}_n) - p(1-p)\right) \xrightarrow{d} \text{Normal}\left(0, \frac{p(1-p)}{(1-2p)^2}\right),$$

for $p \neq \frac{1}{2}$. The statistic $\overline{X}_n(1-\overline{X}_n)$ is typically used to estimate the variance $p(1-p)$.

Ex. 8.5.9. Let $X_i, i \geq 1$ be i.i.d. Exp(λ) random variables. Let $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$. Show that

$$\sqrt{n}\left(\frac{1}{\overline{X}_n} - \frac{1}{\lambda}\right) \xrightarrow{d} \text{Normal}\left(0, \frac{1}{\lambda^2}\right).$$

Ex. 8.5.10. Consider the same set up as in Example 8.5.3 with $\mu = 0$. Then show that

$$\sqrt{n}X_n^2 \xrightarrow{p} 0$$

as $n \to \infty$, and that the correct scaling is n as opposed to $\sqrt{n}$, that is, $n\frac{X_n^2}{\sigma^2} \xrightarrow{d} \chi_1^2$ as $n \to \infty$.

8.6 LIMITING DISTRIBUTION OF SAMPLE MEDIAN

The sample median is a natural alternative to the sample mean as a measure of centrality. For continuous symmetric distributions, the median is the same as the mean when the mean exists. The median of a distribution always exists, even if the mean does not. It is invariant under monotone transformations, making it a more appealing measure of centrality for skewed distributions. The asymptotic distribution of the sample median is therefore of natural interest.

The Central Limit Theorem establishes the Normal distribution as the limiting distribution of the standardized sample mean for all distributions that have finite second moment. This is a universality result which says that all sums and averages from a random sample are asymptotically Normal. However, from a sample one can derive many other summary statistics, each with different sampling distributions, and it is natural to ask about their asymptotic behaviour. In this section, we show that the limiting distribution of the sample median is also Normal. We do this in two stages. First we prove it for Uniform$(0,1)$ random variables, and then use the Delta method to prove it for more general distributions.

Lemma 8.6.1. *Suppose that $U_1, U_2, \ldots$ are i.i.d. Uniform$(0,1)$ random variables, and let $\widetilde{U}_n$ be the sample median obtained from $U_1, \ldots, U_n$. Then,*

$$2\sqrt{n}\left(\widetilde{U}_n - \frac{1}{2}\right) \xrightarrow{d} Z, \tag{8.6.1}$$

where Z has a standard Normal distribution.

Proof. To begin with and to keep the definition of the median unambiguous, we consider odd samples sizes such that $n = 2k-1$ for some positive integer k and we shall let $k \to \infty$. In this case the median $\widetilde{U}_n = U_{(k)}$.

As seen in Example 8.1.4, $U_{(k)}$ has the Beta(k,k) distribution with density

$$f_k(u) = \begin{cases} \frac{k}{2}\binom{2k}{k} u^{k-1}(1-u)^{k-1} & 0 < u < 1, \\ 0 & \text{otherwise.} \end{cases}$$

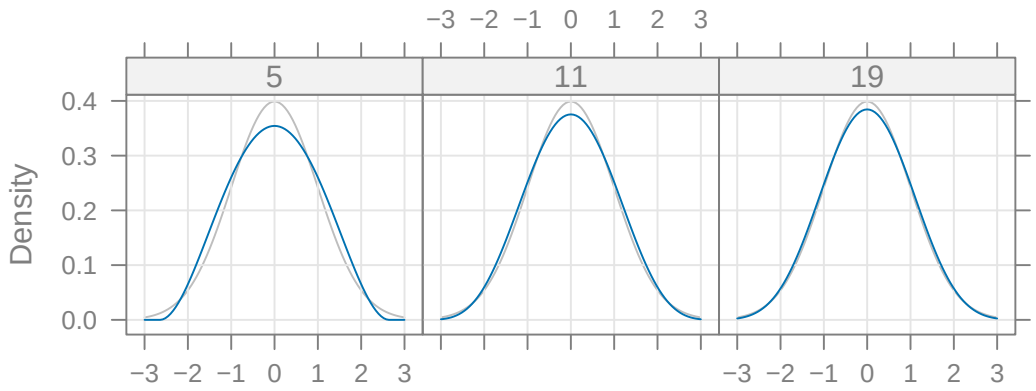


Figure 8.7: Density of standardized sample median for a Uniform$(0,1)$ population, for sample sizes 5, 11, and 19. The grey curve represents the standard Normal density. As the sample size increases, the density function of the standardized median converges to the standard Normal density.

It is easy to verify that $E[U_{(k)}] = \frac{1}{2}$ and $\text{Var}[U_{(k)}] = \frac{1}{4(2k+1)}$. It follows that the density of the standardized median

$$Z_k = 2\sqrt{2k+1}\left(U_{(k)} - \frac{1}{2}\right)$$

is given, after simplification, by

$$g_k(z) = \begin{cases} \dfrac{k}{4^k\sqrt{2k+1}}\dfrac{2k!}{k!^2}\left[1 - \dfrac{z^2}{2k+1}\right]^{k-1} & \text{if } -\sqrt{2k+1} < z < \sqrt{2k+1}, \\ 0 & \text{otherwise.} \end{cases} \tag{8.6.2}$$

Figure 8.7, which plots $g_k(z)$ for $k = 3, 6, 10$ which correspond to $n = 5, 11, 19$ shows that the Normal approximation is good even for small values of n.

By Stirling's approximation for factorials we have that for all $k \geq 1$

$$k^k\sqrt{2\pi k}\exp(-k) < k! < k^k\sqrt{2\pi k}\exp\left(-k + \frac{1}{12k}\right). \tag{8.6.3}$$

Therefore, using the bounds provided in (8.6.3) in (8.6.2) we have

$$e^{-\frac{1}{6k}}\frac{\sqrt{k}}{\sqrt{k+\frac{1}{2}}}\frac{1}{\sqrt{2\pi}}\left[1 - \frac{z^2}{2k+1}\right]^{k-1} \leq g_k(z) \leq e^{\frac{1}{6k}}\frac{\sqrt{k}}{\sqrt{k+\frac{1}{2}}}\frac{1}{\sqrt{2\pi}}\left[1 - \frac{z^2}{2k+1}\right]^{k-1}. \tag{8.6.4}$$

Using the facts that

(a) $\lim_{k\to\infty}\frac{\sqrt{k}}{\sqrt{k+\frac{1}{2}}} = 1$

(b) $\lim_{k\to\infty} \exp(\frac{1}{6k}) = \lim_{k\to\infty} \exp(-\frac{1}{6k}) = 1$

(c) for all $z \in \mathbb{R}$, $\lim_{k\to\infty} \left[1 - \frac{z^2}{2k+1}\right]^{k-1} = e^{-\frac{1}{2}z^2}$

in (8.6.4) we obtain for all $z \in \mathbb{R}$

$$g_k(z) \to \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \text{as } k \to \infty. \tag{8.6.5}$$

Theorem 8.3.5 then implies that

$$Z_k \xrightarrow{d} \text{Normal}\,(0,1) \text{ as } k \to \infty. \tag{8.6.6}$$

As $\sqrt{n}(\widetilde{U}_n - \frac{1}{2}) = \frac{\sqrt{n}}{\sqrt{n+2}} Z_{\frac{n+1}{2}}$, an application of Lemma 8.3.10 (see Exercise 8.5.2) yields the result when n is odd.

For $n = 2k$ even, the median may be defined as any value between $U_{(k)}$ and $U_{(k+1)}$, i.e., any convex combination of $U_{(k)}$ and $U_{(k+1)}$. For example, we could take $\frac{U_{(k)}+U_{(k+1)}}{2}$. Imitating the above argument in n odd case one can prove the result in this case as well (see Exercise 8.6.1). ■

Having established the result for the special case of Uniform$(0,1)$, we next wish to prove that a similar result holds for a much wider class of distributions. For this, we will use the Delta method, generalizing the result to all continuous distributions which have strictly positive density at the population median.

Theorem 8.6.2. *Let $X_1, X_2, \ldots$ be i.i.d. random variables with probability density function f. Let $\widetilde{X}_n$ be the sample median obtained from $X_1, X_2, \ldots, X_n$. Assume that $f(\mu) > 0$, where μ denotes the median of the random variable X. Then,*

$$2\sqrt{n} f(\mu) \left(\widetilde{X}_n - \mu\right) \xrightarrow{d} Z, \tag{8.6.7}$$

where Z has a standard Normal distribution.

Proof. Define $U_i = F(X_i)$ for $i \geq 1$. By Lemma 5.3.7, Exercise 5.3.12 and Theorem 8.1.2, $U_1, U_2, \ldots$ are i.i.d. Uniform$(0,1)$. By Lemma 8.6.1

$$2\sqrt{n}\left(\widetilde{U}_n - \frac{1}{2}\right) \xrightarrow{d} Z, \tag{8.6.8}$$

where Z is standard Normal and $\widetilde{U}_n$ is the median of $U_1, \ldots, U_n$. Let F be the distribution function of X. Now define $G : [0,1] \to \mathbb{R} \cup \{-\infty\} \cup \{\infty\}$ as

$$G(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}.$$

Recall from Exercise 5.3.12 that G is the generalised inverse of F. First note, since X_i are sampled from f, $X_i = G(U_i)$ for $i \geq 1$ with probability 1. By definition G is increasing, and so $\widetilde{X}_n = G(\widetilde{U}_n)$.

Further, since $f(\mu) > 0$ then F is strictly monotone and F' (exists, is continuous) is strictly positive in a neighbourhood of μ. This will imply $G(\frac{1}{2}) = \mu$, G is differentiable at μ with $G'(\frac{1}{2}) = \frac{1}{f(\mu)} > 0$ and $G'(\cdot)$ is continuous in a neighbourhood of μ.

As G satisfies the hypothesis of Theorem 8.5.1, using (8.6.8) we have

$$\sqrt{n}\,\frac{G(\widetilde{U}_n) - G(\frac{1}{2})}{\frac{1}{2}G'(\frac{1}{2})} \xrightarrow{d} Z,$$

with Z being standard Normal. The result follows. ■

We conclude this section with a couple of examples.

Example 8.6.3. Suppose that $X_1, X_2, \ldots$ are i.i.d. Normal(μ, σ^2) random variables, and let $\widetilde{X}_n$ be the sample median obtained from $X_1, \ldots, X_n$. Then, it follows from Theorem 8.6.2 that

$$\sqrt{n}\sqrt{\frac{2}{\pi}}\,\frac{(\widetilde{X}_n - \mu)}{\sigma} \xrightarrow{d} Z, \tag{8.6.9}$$

where Z has a standard Normal distribution.

Note that although the sample mean $\overline{X}_n \sim \text{Normal}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, the distribution of the sample median $\widetilde{X}_n$, which can be calculated from (8.1.3), is not Normal. Asymptotically however, the standardized sample median also has a limiting Normal distribution. One can compare their *asymptotic efficiency* in terms of limiting variances. As noted, we have

$$\sqrt{n}\frac{(\widetilde{X}_n - \mu)}{\sigma} \xrightarrow{d} \sqrt{\frac{2}{\pi}}Z \qquad \text{and} \qquad \sqrt{n}\frac{(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} Z \tag{8.6.10}$$

The asymptotic efficiency of $\widetilde{X}_n$ over $\overline{X}_n$ is defined as the inverse of the ratio of the limiting variances, i.e., $\frac{2}{\pi} \approx 0.64$. This number can be interpreted in the following manner: If one uses $\widetilde{X}_n$ to estimate μ, then one could instead have used use $\overline{X}_m$ with sample size $m \approx 0.64n$ to get an estimator with the same variance. ■

The previous example suggests that one should use the sample mean rather than the sample median when the population is Normal. In practice, however, the underlying distribution can be rarely known with certainty, and we will see in Chapter 9 that under even fairly mild departures from Normality, the sample median may become much more useful than the sample mean in the sense of asymptotic efficiency as defined above. An extreme case is the following example, where the sample mean does not have finite variance, and hence the asymptotic efficiency of the sample mean over the sample median is zero.

Example 8.6.4. Suppose that $X_1, X_2, \ldots$ are i.i.d. Cauchy(θ, α^2) random variables. It is easy to see (by symmetry) that θ is the median. As the Cauchy distribution has no finite moments, one way to estimate θ is using the sample median. We can apply Theorem 8.6.2 to get

$$2\sqrt{n}\frac{1}{\pi\alpha}(\widetilde{X}_n - \theta) \to Z \quad \text{as } n \to \infty,$$

where Z is standard Normal. ■

EXERCISES

Ex. 8.6.1. Suppose that $U_1, U_2, \ldots$ are i.i.d. Uniform$(0,1)$ random variables, and let $\widetilde{U}_n$ be the sample median obtained from $U_1, \ldots, U_n$, with $n = 2k$ for some $k \geq 1$.

(a) Using Example 8.1.4 find the distribution of $U_{(k)}$ and $U_{(k+1)}$ and Compute $E[U_{(k)}]$, $\text{Var}[U_{(k)}]$, $E[U_{(k+1)}]$ and $\text{Var}[U_{(k+1)}]$.

(b) As in the proof of Lemma 8.6.1 find a_k and b_k such that both $Z_k = a_k(U_{(k)} - \frac{1}{2})$, $\widetilde{Z}_k = b_k(U_{(k+1)} - \frac{1}{2})$ converge in distribution to Normal $(0,1)$ as $k \to \infty$.

(c) Using Lemma 8.3.10 For $0 < \lambda < 1$ show that $\sqrt{n}(\lambda U_k + (1-\lambda)U_{k+1} - \frac{1}{2}) \xrightarrow{d}$ Normal $(0,1)$ as $k \to \infty$.

Ex. 8.6.2. Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. Exp(λ) random variables. Find the distribution of sample median $\widetilde{X}_n$ and also identify the standardization required to obtain a Normal distribution as the limiting distribution of the standardized sample median.

Ex. 8.6.3. Under the conditions of Theorem 8.6.2, show that the sample median converges to the population median in probability. Thus, the Weak Law of large numbers holds for the sample median.

9

ESTIMATION

In Chapter 7 we discussed how an i.i.d. sample $X_1, X_2, \ldots, X_n$ from an unknown distribution may be used to estimate aspects of that distribution. In Chapter 8 we saw how some sample statistics behave asymptotically. In this chapter we look at some specific examples where various parameters of the distribution such as the mean μ and the standard deviation σ are unknown, and the sample statistics that are used to estimate these parameters.

For instance, suppose there is a coin which we assume has a probability p of showing heads each time it is flipped. To gather information about p the coin is flipped 100 times. The results of these flips are viewed as i.i.d. random variables $X_1, X_2, \ldots, X_{100}$ with a Bernoulli(p) distribution. Suppose $\sum_{n=1}^{100} X_n = 67$, meaning that 67 of the 100 flips showed heads. How might we use this information to infer something about the value of p?

The first two topics we will consider are the "method of moments" and the "method of maximum likelihood". Both of these are direct forms of estimation in the sense that they produce a single-value estimate for p. A benefit of such methods is that they produce a unique prediction, but a downside is that the prediction they make is most likely not exactly correct. These methods amount to a statement like "Since 67 of the 100 flips came up heads, we predict that the coin should come up heads 67% of the time in the long run". In some sense the 67% prediction may be the most reasonable one given what was observed in the 100 flips, but it should also be recognised that 0.67 is unlikely to be the true value of p.

Another potential approach is that of the "confidence interval". In using this method we recognise that a specific estimate is unreasonable, and instead produce a range of values which is expected to contain the true value of the parameter. This approach could yield a statement such as this: "With 90% confidence, the actual probability that the coin will show heads is between 0.59 and 0.75". Of course, the true p is not random and will either lie between 0.59 and 0.75 or it will not; there is nothing probabilistic about that event. The 90% confidence here refers to our professed confidence in the *procedure*, in the sense that we believe that the procedure produces a "correct" interval with probability 0.9.

Yet another approach is based on the idea of a "hypothesis test." In this case we make a conjecture about the value of the parameter and carry out a computation to test the credibility of the conjecture. There is an obvious link between hypothesis tests and confidence intervals: we can define a confidence interval to consist of values of the parameter that are credible according to a test, and vice versa. We will discuss the hypothesis testing approach in the next chapter.

For the purposes of this chapter, we will assume that the sample $X_1, X_2, \ldots, X_n$ are i.i.d. copies of a random variable X with a probability mass function or probability density function $f(x)$. For brevity, we shall often refer to the distribution X, by which we will mean the distribution of the random variable X. We shall further assume that $f(x)$ depends on one or more unknown parameters $p_1, p_2, \ldots, p_d$ and emphasise this using the notation $f(x \mid p_1, p_2, \ldots, p_d)$. We may abbreviate this

as $f(x \mid p)$, where $p = (p_1, p_2, \ldots, p_d)$ represents the vector of all the parameters. We will assume that the set $\mathcal{P}$ of all possible values p can take is known, where $\mathcal{P} \subset \mathbb{R}^d$ for some $d \geq 1$. The set $\mathcal{P}$ may be all of $\mathbb{R}^d$ or some proper subset depending on the nature of the parameters.

We now fix some notations and terminology for estimators.

Definition 9.0.1. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a population with distribution $f(x \mid p)$. Suppose we are interested in estimating $\theta(p)$ for some $\theta : \mathcal{P} \to \mathbb{R}$. Then $g(X_1, X_2, \ldots, X_n)$ for any $g : \mathbb{R}^n \to \mathbb{R}$ can be considered as a "point estimator" of $\theta(p)$, and its value from a particular realisation is called an "estimate".*

In practice the function g is chosen keeping in mind the parameter $\theta(p)$ of interest. We have seen the following in Chapter 7.

EXAMPLE 9.0.2. Let $\mu = E[X]$. Let $g : \mathbb{R}^n \to \mathbb{R}$ be given by

$$g(x) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Then $g(X_1, X_2, \ldots, X_n)$ is the (now familiar) sample mean and it is an estimator for μ. Further, $E[g(X_1, X_2, \ldots, X_n)] = \mu$ regardless of the true value of μ. We called such an estimator an unbiased estimator. Finally we also know by the strong law of large numbers, Theorem 8.2.1, that $g(X_1, X_2, \ldots, X_n) \xrightarrow{p} \mu$ as $n \to \infty$. ■

Recall from Chapter 6 that $E[X]$ is the first moment of X. As noted in Chapter 7, we can thus view the sample mean, which is the first moment of the empirical distribution based on a sample, as estimating the first moment of the underlying distribution. A generalization of this method is known as the method of moments.

9.1 METHOD OF MOMENTS

For $n \geq 1$, let $X_1, X_2, \ldots, X_n$ be a sample from a population with distribution X. Assume that X has either probability mass function or probability density function $f(x \mid p)$ depending on parameter(s) $p = (p_1, p_2, \ldots, p_d)$, for some $d \geq 1$. Let for $k \geq 1$, $m_k : \mathbb{R}^n \to \mathbb{R}$ be given by

$$m_k(x) = \frac{1}{n} \sum_{i=1}^{n} x_i^k.$$

Notice that $m_k(X_1, X_2, \ldots, X_n)$ is the k-th moment of the empirical distribution based on the sample $X_1, X_2, \ldots, X_n$, which we will refer to simply as the k-th sample moment.

Let $\mu_k = E[X^k]$, the k-th moment of the distribution X. As the distribution of X depends on $(p_1, p_2, \ldots, p_d)$ one can view $\mu_k \equiv \mu_k(p_1, p_2, \ldots, p_d)$ as a function of p. The method of moments

estimator for $(p_1, p_2, \ldots, p_d)$ is obtained by equating the first d sample moments to the corresponding moments of the distribution. Specifically, it requires solving the d equations in d unknowns given by

$$\mu_k(p_1, p_2, \ldots, p_d) = m_k(X_1, X_2, \ldots, X_n), \quad k = 1, 2, \ldots, d.$$

for $p_1, p_2, \ldots, p_d$. There is no guarantee in general that these equations have a unique solution or that it can be computed, but in practice it is often possible to do so. The solution will be denoted by $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_d$ which will be writen in terms of the realised values for $m_k, k = 1, 2, \ldots, d$. We will now explore this method for two examples.

Example 9.1.1. Suppose $X_1, X_2, \ldots, X_{10}$ is an i.i.d. sample with distribution Binomial(N, p) where neither N nor p is known. Suppose the empirical realisation of these variables is 8, 7, 6, 11, 8, 5, 3, 7, 6, 9. One can check that the average of these values is $m_1 = 7$ while the average of their squares is $m_2 = 53.4$. Since $X \sim$ Binomial (N, p) the probability mass function is given by

$$f(k \mid N, p) = \binom{N}{k} p^k (1-p)^{N-k}, k = 0, 1, \ldots, N.$$

We have previously shown that

$$E[X] = Np \text{ and } E[X^2] = \text{Var}[X] + E[X]^2 = Np(1-p) + N^2p^2.$$

Thus, the method of moments estimator for (N, p) is obtained by solving

$$7 = m_1 = \hat{N}\hat{p} \text{ and } 53.4 = m_2 = \hat{N}\hat{p}(1-\hat{p}) + \hat{N}^2\hat{p}^2.$$

Using elementary algebra we see that

$$\begin{aligned} \hat{N} &= \frac{m_1^2}{m_1 - (m_2 - m_1^2)} \approx 19 \\ \hat{p} &= \frac{m_1 - (m_2 - m_1^2)}{m_1} \approx 0.371. \end{aligned}$$

Thus, according to the method of moments, the distribution from which the sample came from is estimated to be the Binomial$(19, 0.371)$ distribution. In practice, we usually wish to restrict the estimates of the parameters based on the context of the problem. Since the N value is surely some integer, the estimate of $\hat{N}$ was rounded to the nearest meaningful value in this case. ■

Example 9.1.2. Suppose our distribution of interest X has a Normal (μ, σ^2) distribution. Therefore our probability density function is given by

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \ x \in \mathbb{R}.$$

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the distribution X. We have shown that

$$E[X] = \mu \text{ and } E[X^2] = \text{Var}[X] + E[X]^2 = \mu^2 + \sigma^2.$$

The method of moments estimator for μ, σ is found by solving

$$m_1 = \mu \text{ and } m_2 = \mu^2 + \sigma^2.$$

from which

$$\begin{aligned} \hat{\mu} &= m_1 = \overline{X} \text{ and} \\ \hat{\sigma}^2 &= m_2 - m_1^2 = \left(\frac{1}{n} \sum_{i=1}^{n} X_i^2 \right) - \overline{X}^2 = \frac{n-1}{n} S^2. \end{aligned}$$

Here $\overline{X}$ and S^2 are, respectively, the sample mean and sample variance defined in Chapter 7. ■

The method of moment estimators may not always be very reliable, in the sense that it might give implausible estimates. For instance, in Example 9.1.1 above, you can check that the estimate for p would be negative if the sample mean $\overline{X}$ happened to be smaller than $\frac{n-1}{n} S^2$. Such defects can be somewhat rectified using moment matching and other techniques (see [CasBer90]).

EXERCISES

Ex. 9.1.1. Suppose $X_1, \ldots, X_5$ is an i.i.d. sample with Uniform(a, b) distribution for some unknown a and b. Suppose the empirical realisation of these variables is $3.5, 2.1, 5.7, 4.8, 3.9$. Use the method of moments to estimate a and b.

Ex. 9.1.2. Suppose $X_1, X_2, \ldots, X_n$ is an i.i.d. sample with Uniform(a, b) distribution for some unknown a and b. Let m_1 and m_2 be the empirical realisation of the first and second moments of the $X_1, X_2, \ldots, X_n$ data. Find an expression for the estimates of a and b given by the method of moments in terms of the quantities m_1 and m_2.

Ex. 9.1.3. Suppose $X_1, X_2, \ldots, X_n$ is an i.i.d. sample with Uniform(a, b) distribution for some unknown a and b. Prove that the method of moments produces estimates $\hat{a}$ and $\hat{b}$ such that $\hat{a} = \hat{b}$ if and only if every data point in the empirical realisation has exactly the same value.

Ex. 9.1.4. Suppose $X_1, \ldots, X_4$ is an i.i.d. sample with Binomial(N, p) distribution for some unknown N and p. Suppose the empirical realisation of these variables is $1, 2, 5, 12$. Show that the method of moments for estimating N and p gives negative (and therefore meaningless) results.

Ex. 9.1.5. Suppose $X_1, X_2, \ldots, X_n$ is an i.i.d. sample with Binomial(N, p) distribution for some unknown N and p. Prove that the method of moments will produce a negative estimate for p if an only if it also produces a negative estimate for N.

Ex. 9.1.6. Suppose $X_1, \ldots, X_6$ is an i.i.d. sample with Gamma(α, λ) distribution for some unknown α and λ. Suppose the empirical realisation of these variables is $5.3, 2.4, 2.8, 7.6, 6.9, 4.2$. Use the method of moments to estimate α and λ.

Ex. 9.1.7. Suppose $X_1, X_2, \ldots, X_n$ is an i.i.d. sample with Gamma(α, λ) distribution for some unknown α and λ. Let m_1 and m_2 be the empirical realisations of the first and second moments of $X_1, X_2, \ldots, X_n$.

(a) Find an expression for the estimates of α and λ given by the method of moments in terms of the quantities m_1 and m_2.

(b) Show that the estimates of α and λ from part (a) can never be negative.

Ex. 9.1.8. The following code simulates 100 samples from a population with distribution Binomial $(20, 0.4)$ and computes the method of moments estimate for the sample size parameter n and the success probability p ($n = 20$ and $p = 0.4$ in the simulation).

```
n <- 100
N <- 20
p <- 0.4
x <- rbinom(n, N, p)
m1 <- mean(x)
m2 <- mean(x^2)
Nhat <- m1^2 / (m1 -(m2-m1^2))
phat <- (m1 - (m2-m1^2)) / m1
```

(a) Run the code in `R` and compute `phat - p` and `Nhat - N`.

(b) Change the code suitably to simulate 1000 samples from Binomial $(20, 0.4)$ and see if the answer to (a) changes.

(c) Change the code suitably to simulate samples from Binomial $(10, 0.1)$ and Binomial $(10, 0.9)$, and repeat (a) and (b).

Ex. 9.1.9. Using the method from Exercise 9.1.2, and by suitably modifying the `R`-code in Exercise 9.1.8, write `R`-code that computes the method of moments estimate for a and b in Uniform(a, b) when $a = 3$ and $b = 5$ by generating 100 samples from Uniform$(3, 5)$

Ex. 9.1.10. Using the method from Example 9.1.2, and by suitably modifying the `R`-code in Exercise 9.1.8, write `R` code that computes the method of moments estimate for μ and σ^2 in Normal(μ, σ^2) when $\mu = 4$ and $\sigma^2 = 10$ by generating 100 samples from Normal$(4, 10)$

Ex. 9.1.11. Using the method from Exercise 9.1.6, and by suitably modifying the `R` code in Exercise 9.1.8, write `R` code that computes the method of moments estimate for a and b in Gamma (a, b) when $a = 10$ and $b = 0.5$ by generating 100 samples from Gamma $(10, 0.5)$

9.2 MAXIMUM LIKELIHOOD

For $n \geq 1$, let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the distribution X. Assume that X has either probability mass function or probability density function denoted by $f(x \mid p)$ depending on parameter(s) $p = (p_1, p_2, \ldots, p_d) \in \mathcal{P} \subset \mathbb{R}^d$.

Definition 9.2.1. *The "likelihood function" for the sample $X_1, X_2, \ldots, X_n$ is the function $L : \mathcal{P} \times \mathbb{R}^n \to \mathbb{R}$ given by*

$$L(p; X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} f(X_i \mid p).$$

For a given sample $X_1, X_2, \ldots, X_n$, suppose $\hat{p} \equiv \hat{p}(X_1, X_2, \ldots, X_n)$ is a point at which $L(p; X_1, X_2, \ldots, X_n)$ attains its maximum as a function of p. Then $\hat{p}$ is called a "maximum likelihood estimator" of p (abbreviated as MLE of p) given the sample $X_1, X_2, \ldots, X_n$.

One observes readily that the likelihood function is the joint density or joint mass function of $(X_1, X_2, \ldots, X_n)$ when p is fixed. Assuming that the MLE $\hat{p}$ as defined above is unique, it can be thought of as the most "likely" value of the parameter p for the given realisation of $X_1, X_2, \ldots, X_n$, as for any other parameter value $\widetilde{p}$, the corresponding joint density or joint mass function has a lower value at $(X_1, X_2, \ldots, X_n)$. If $\hat{p}$ is not unique, the same is true for any $\widetilde{p}$ which is not an MLE.

Example 9.2.2. Let $p \in \mathbb{R}$ and $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a population distributed as Normal with mean p and variance 1. Then the likelihood function is given by

$$L(p; X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i - p)^2}{2}} = \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\sum_{i=1}^{n}(X_i - p)^2}.$$

To find the MLE, treating the given the realisation $X_1, X_2, \ldots, X_n$ as fixed, one needs to maximise L as a function of p. Noting that maximising L is equivalent to maximising $\log_e L$ (as logarithm is an increasing function), the problem is then to find the minimum of $g : \mathbb{R} \to \mathbb{R}$ given by

$$g(p) = \sum_{i=1}^{n} (X_i - p)^2.$$

Method 1: Since $g(p) = \sum_{i=1}^{n} (X_i - \overline{X})^2 + (\overline{X} - p)^2$ (see Exercise 9.2.2) and the first term does not depend on p, the minimum of g will occur at $\hat{p} = \overline{X}$.

Method 2: An alternative approach is to use differential calculus. As g is a quadratic function of p, it is differentiable at all p, and

$$g'(p) = -2\sum_{i=1}^{n} (X_i - p) \qquad \text{and } g''(p) = 2n.$$

As $g''(\cdot) > 0$, the minimum will occur when $g'(p) = 0$. This occurs when p is equal to $\frac{1}{n}\sum_{i=1}^{n} X_i$. So the MLE of p is given by $\hat{p} = \overline{X}$. ■

Example 9.2.3. Let $p \in (0,1)$ and $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a population distributed as Bernoulli(p). The probability mass function f can be written as

$$f(x \mid p) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \\ 0 & \text{otherwise.} \end{cases} = \begin{cases} p^x(1-p)^{1-x} & \text{if } x \in \{0,1\} \\ 0 & \text{otherwise.} \end{cases}$$

Then the likelihood function is given by

$$L(p\,; X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\left(\sum_{i=1}^{n} X_i\right)}(1-p)^{\left(n-\sum_{i=1}^{n} X_i\right)}.$$

To find the MLE, treating the given the realisation $X_1, X_2, \ldots, X_n$ as fixed, one needs to maximise L as a function of p. We can use calculus to do this, but differentiating L is cumbersome, so as before we look at $\log_e L$, which is called the log likelihood function.

$$\begin{aligned} \ell(p\,; X_1, X_2, \ldots, X_n) &= \log_e L(p\,; X_1, X_2, \ldots, X_n) \\ &= \begin{cases} \log_e\left(\frac{p}{1-p}\right)\left(\sum_{i=1}^{n} X_i\right) + n\log_e(1-p) & \text{if } 0 < \sum_{i=1}^{n} X_i < n \\ n\log_e(1-p) & \text{if } \sum_{i=1}^{n} X_i = 0 \\ n\log_e(p) & \text{if } \sum_{i=1}^{n} X_i = n \end{cases} \end{aligned}$$

As $\sum_{i=1}^{n} X_i$ is fixed for the purpose of this maximisation problem, we can approach the problem separately for the three cases above.

In the first case with $0 < \sum_{i=1}^{n} X_i < n$, we can re write

$$\ell(p\,; X_1, X_2, \ldots, X_n) = \log_e(p)\left(\sum_{i=1}^{n} X_i\right) + \left(n - \sum_{i=1}^{n} X_i\right)\log_e(1-p).$$

It is easy to see that

$$\ell'(p\,; X_1, X_2, \ldots, X_n) = \frac{1}{p}\left(\sum_{i=1}^{n} X_i\right) - \frac{1}{1-p}\left(n - \sum_{i=1}^{n} X_i\right)$$

and

$$\ell''(p\,; X_1, X_2, \ldots, X_n) = -\frac{1}{p^2}\left(\sum_{i=1}^{n} X_i\right) + \frac{1}{(1-p)^2}\left(n - \sum_{i=1}^{n} X_i\right).$$

As $0 < \sum_{i=1}^{n} X_i < n$, $\ell''(p\,; X_1, X_2, \ldots, X_n) < 0$ for all p. So the global maximum will occur at the point where $\ell'(p\,; X_1, X_2, \ldots, X_n) = 0$. This happens when $p = \frac{1}{n}\sum_{i=1}^{n} X_i$.

In the second case with $\sum_{i=1}^{n} X_i = 0$, ℓ is a decreasing function of p and the maximum occurs at $p = 0$ which can be trivially re-written as $p = \frac{1}{n}\sum_{i=1}^{n} X_i$ in this case.

In the third case with $\sum_{i=1}^{n} X_i = n$, ℓ is an increasing function of p and maximum occurs at $p = 1$ which can be trivially re-written as $p = \frac{1}{n}\sum_{i=1}^{n} X_i$ in this case.

Combining the three cases, we can conclude that the MLE of p, $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$. ■

At times we may wish to maximize the likelihood as a function of a parameter that takes values in a discrete set. Consider a collection of empirical measurements of waiting times. Suppose we know that each waiting time is the sum of some fixed number of i.i.d. Exp(λ) distributions, but we are not certain how many such distributions are in each sum. We might let m represent that unknown number, and attempt to find the m which maximizes the likelihood. As we have previously seen, such sums have a Gamma (m, λ) distribution. In the example below we will assume λ is known.

Example 9.2.4. Let $\lambda > 0$ and let m be an unknown positive integer. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a population distributed as Gamma(m, λ). Then the likelihood function is given by

$$L(m) = L(m\,; X_1, \ldots, X_n) = \prod_{i=1}^{n} \frac{\lambda^m}{(m-1)!} X_i^{m-1} e^{-\lambda X_i}.$$

Now consider the ratio

$$\frac{L(m+1)}{L(m)} = \frac{\lambda^n}{m^n} \prod_{i=1}^{n} X_i.$$

This ratio is a decreasing function of m, so $L(m)$ is maximimized at the smallest value of m for which this ratio is less than 1. Therefore the maximum likelihood estimate for m is the smallest integer which is larger than $\lambda \left(\prod_{i=1}^{n} X_i \right)^{1/n}$. The quantity $\left(\prod_{i=1}^{n} X_i \right)^{1/n}$ is known as the "geometric mean" of the $X_1, X_2, \ldots, X_n$ values. ■

As a final example, let us revisit Example 9.1.1, where we considered a Binomial distribution with both parameters unknown.

Example 9.2.5. Suppose $X_1, X_2, \ldots, X_n$ is an i.i.d. sample with distribution Binomial(N, p) where neither N nor p is known. The likelihood function is given by

$$L(N, p\,; X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} \binom{N}{X_i} p^{X_i} (1-p)^{N - X_i}.$$

To obtain MLEs for N and p, we need to maximise this expression as a function of N and p, for a fixed set of empirical observations $X_1, X_2, \ldots, X_n$.

This is unfortunately not easy to do explicitly. We can simplify the problem by observing that we have already calculated an estimate of p if N is known. In that case, $\sum_{i=1}^{n} X_i$ is the sum of Nn

independent Bernoulli(p) random variables, for which the MLE of p is (see Example 9.2.3 and Exercise 9.2.9)

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{Nn}.$$

By plugging in this estimator in the expression for $L(N, p; X_1, X_2, \ldots, X_n)$, we obtain the so called "profiled likelihood function"

$$\widetilde{L}(N) \equiv L(N, \hat{p}; X_1, X_2, \ldots, X_n),$$

which can now be viewed as a function of N only. Such profiled likelihood functions, where some parameters in the likelihood function are replaced by estimators that depend on the remaining parameters, are useful because they reduce the number of parameters over which the maximization problem needs to be solved. It is easy to see that maximizing the profiled likelihood is equivalent to maximizing the original likelihood function.

Unfortunately, further theoretical analysis of this function is difficult. Numerically, however, this problem is not difficult to solve. Consider again the empirical realisations given in Example 9.1.1, with $n = 10$ and observations 8, 7, 6, 11, 8, 5, 3, 7, 6, 9. Clearly, N must be at least 11, the largest of the observations. Let us use R to compute the logarithm of the profiled likelihood for values of N from 11 to 50.

```
x <- c(8, 7, 6, 11, 8, 5, 3, 7, 6, 9)
n <- length(x)
N <- 11:50
phat <- sum(x) / (N * n)
logL <- rep(0, length(N))
for (i in seq(1, length(N)))
    logL[i]  <- sum(dbinom(x, size = N[i],
                           prob = phat[i], log = TRUE))
d <- data.frame(N = N,
                phat = phat,
                logL = logL)
```

We can now plot these log-likelihood values, as we do in Figure 9.1, or look at them directly.

```
head(d, 15)
```

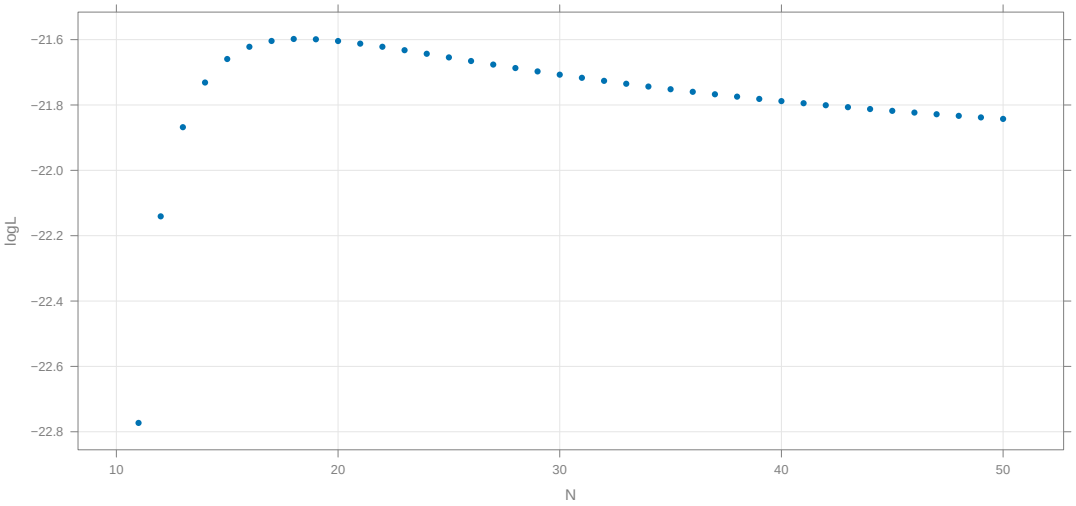


Figure 9.1: Profiled log-likelihood for the Binomial(N, p) distribution as a function of N.

N	phat	logL
11	0.6363636	-22.77295
12	0.5833333	-22.14081
13	0.5384615	-21.86794
14	0.5000000	-21.73123
15	0.4666667	-21.65932
16	0.4375000	-21.62190
17	0.4117647	-21.60412
18	0.3888889	-21.59799
19	0.3684211	-21.59894
20	0.3500000	-21.60426
21	0.3333333	-21.61225
22	0.3181818	-21.62184
23	0.3043478	-21.63233
24	0.2916667	-21.64328
25	0.2800000	-21.65438

By inspecting the first few rows of the table, we see that the likelihood is maximized at $\hat{N} = 18$ and $\hat{p} = 0.389$. These estimates are not very different from the ones we obtained using the method of moments. ■

Similar numerical methods are required to compute the maximum likelihood estimate in many other examples (See Exercise 9.2.10).

Remark 9.2.6. In general, one may not be able to compute the sampling distribution (i.e. probability distribution of the random-sample-based statistic) of the maximum likelihood estimate. However there is a well-understood theory of limiting distributions of maximum likelihood estimators, whether

they are available in closed form solutions or not. They are "better" than other estimators in terms of variance, and follow a normal distribution, asymptotically. A detailed discussion and proof of these results are beyond the scope of this book.

Sampling distributions of these estimates do play an important role in obtaining confidence interval (see Section 9.3) and test of hypotheses (see Chapter 10). For this one must understand the limiting behaviour of the sampling distributions.

EXERCISES

Ex. 9.2.1. In the examples above we have used the fact that an exponential with negative exponent may be maximized by minimizing the exponent. Prove that this is generally true. Suppose $f(x)$ is a function which achieves a minimum when $x = a$. Let $g(x) = e^{-f(x)}$. Prove that $g(x)$ achieves a maximum when $x = a$.

Ex. 9.2.2. Show that for any real numbers $p, x_1, x_2, \ldots, x_n$,

$$\sum_{i=1}^{n}(x_i - p)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 + (\overline{x} - p)^2.$$

Ex. 9.2.3. Let $\lambda > 0$ and $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a population with Exponential(λ) distribution.

(a) Find the likelihood function $L(\lambda\,; X_1, X_2, \ldots, X_n)$.

(b) Prove that the maximum likelihood estimate for λ is $1/\overline{X}$.

Ex. 9.2.4. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a population with Poisson(λ) distribution, where λ is known to be strictly positive.

(a) Find the likelihood function $L(\lambda\,; X_1, X_2, \ldots, X_n)$.

(b) Prove that if at least one of the X_j values in non-zero, then the maximum likelihood estimate for λ is $\overline{X}$.

(c) Prove that if all of the X_j values are zero, then $L(\lambda\,; X_1, X_2, \ldots, X_n)$ has no maximum value for $\lambda > 0$.

Ex. 9.2.5. Let $0 < p < 1$ and $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a population with Geometric(p) distribution.

(a) Find the likelihood function $L(p\,; X_1, X_2, \ldots, X_n)$.

(b) Let $\ell(p\,; X_1, X_2, \ldots, X_n) = \log_e(L(p\,; X_1, X_2, \ldots, X_n))$. Find the value of p for which $\ell(p\,; X_1, X_2, \ldots, X_n)$ is maximized.

(c) Prove that the maximum likelihood estimate for p is $1/\overline{X}$.

Ex. 9.2.6. Let $\sigma > 0$ and $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a population with Normal($0, \sigma^2$) distribution.

(a) Find the likelihood function $L(\sigma\,; X_1, X_2, \ldots, X_n)$.

(b) Prove that the maximum likelihood estimate for σ^2 is $\frac{1}{n}\sum_{i=1}^{n} X_i^2$.

(c) Prove that this maximum likelihood estimate is also an unbiased estimator for σ^2 in this case.

Ex. 9.2.7. Let $\mu \in \mathbb{R}, \sigma > 0$ and $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a population with Normal(μ, σ^2) distribution.

(a) Find the likelihood function $L(\mu, \sigma\,; X_1, X_2, \ldots, X_n)$.

(b) Find the maximum likelihood estimators of μ and σ^2.

Ex. 9.2.8. Let $a < b$ and $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a population with Uniform(a, b) distribution. Prove that the maximum likelihood estimates for a and b are $min\{X_1, X_2, \ldots, X_n\}$ and $max\{X_1, X_2, \ldots, X_n\}$ respectively.

Ex. 9.2.9. Let m be a known integer and $p \in (0, 1)$ be an unknown parameter. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a population with Binomial(m, p) distribution.

(a) Find the likelihood function $L(p, X_1, X_2, \ldots, X_n)$.

(b) Let $\ell(p\,; X_1, X_2, \ldots, X_n) = \log_e(L(p\,; X_1, X_2, \ldots, X_n))$. Find the value of p that maximizes $\ell(p\,; X_1, X_2, \ldots, X_n)$.

(c) Prove that the maximum likelihood estimate for p is $\overline{X}/m$.

Ex. 9.2.10. Let $\theta > 0$ be unknown and $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a population with Cauchy$(\theta, 1)$ distribution.

(a) Find the likelihood function $L(\theta, X_1, X_2, \ldots, X_n)$.

(b) Let $\ell(\theta\,; X_1, X_2, \ldots, X_n) = \log_e(L(\theta\,; X_1, X_2, \ldots, X_n))$. Decide that one cannot explicitly compute the critical points on $\ell(\cdot)$.

(c) Modify the R code provided in Example 9.2.5 and try to find the maximum likelihood estimate for θ.

Ex. 9.2.11. Suppose we have a sample of size n from Multinomial distribution with parameters $k, (p_1, p_2, \ldots, p_k)$. Let $X_j, 1 \le j \le k$ represent the number of samples that correspond to j-th outcome. Show that for each $1 \le j \le k$ the MLE for p_j, $\hat{p}_j = \frac{X_j}{n}$.

9.3 CONFIDENCE INTERVALS

In the previous sections, we have considered data $X_1, X_2, \ldots, X_n$ whose distributions are governed by parameters and described two general methods (namely the methods of moments and of maximum likelihood) to estimate the parameters of the model from this data. In this section, we will try to understand how we can quantify the accuracy of estimates.

We will start with the simple model considered in Example 9.2.2 where data are distributed as Normal with unknown mean but known variance.

Example 9.3.1. Let $p \in \mathbb{R}$ and $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a population distributed as Normal with mean p and variance 1. Both the method of moments and the method of maximum likelihood give $\hat{p} = \overline{X}$ as the estimator of p.

We know from Chapter 8 that $\hat{p}$ has a normal distribution with mean p and variance $1/n$. This tells us that $\hat{p}$ is more "likely" to be close to the true mean p for larger sample size n. It is conventional to present this information in a slightly different way, in the form of a *confidence interval*. The idea of a confidence interval is that instead of a point estimate for a (scalar) parameter p, we report an interval, depending on the observed data, that is "likely" to contain the unknown p.

As $\hat{p} \sim \text{Normal}(p, 1/n)$, we have $\sqrt{n}(\hat{p} - p) \sim \text{Normal}(0, 1)$. So we can write, for instance,

$$P(-3 \leq \sqrt{n}(\hat{p} - p) \leq 3) = \Phi(3) - \Phi(-3) = 0.9973, \tag{9.3.1}$$

where the probability value on the right hand side can be computed using R.

```
pnorm(3) - pnorm(-3)
```

```
[1] 0.9973002
```

By manipulating the inequalities, we can write the equation above in its more standard form, namely,

$$P\left(\hat{p} - \frac{3}{\sqrt{n}} \leq p \leq \hat{p} + \frac{3}{\sqrt{n}}\right) = 0.9973.$$

or

$$P\left(p \in \left(\hat{p} - \frac{3}{\sqrt{n}}, \hat{p} + \frac{3}{\sqrt{n}}\right)\right) = 0.9973.$$

It is common practice to express this interval more concisely as

$$\hat{p} \pm \frac{3}{\sqrt{n}}.$$

When viewed as a random interval, the probability statement above implies that the interval contains p with probability 0.9973. This property is usually conveyed by the statement that the interval is a 99.73% "confidence interval." The factor of 3 used above, which led to the confidence level of 99.73%, is arbitrary. It is more common to specify a desired confidence level and then calculate the factor accordingly. For example, to get a confidence level of 95%, we want a factor z such that

$$P(Z > z) = P(Z < -z) = 0.05/2 = 0.025,$$

where $Z \sim \text{Normal}(0, 1)$. Such a z is easily calculated as

```
-qnorm(0.025)
```

```
[1] 1.959964
```

Similarly, for a confidence level of 80%, the factor can be obtained as

```
-qnorm(0.10)
```

```
[1] 1.281552
```

If we consider a specific empirical realization of $X_1, X_2, \ldots, X_n$, this process leads to a specific interval. For example, consider the following $n = 15$ data points: 11.22, 9.56, 10.06, 10.21, 10.95, 10.03, 10.75, 10.71, 11.42, 9.61, 10.91, 8.14, 8.95, 10.57, 11.1. Here, $\hat{p} = 10.279$, so the 99.73% confidence interval is given by

$$10.279 \pm \frac{3}{\sqrt{15}} = [9.5047, 11.054]$$

Of course, this *specific* interval may or may not contain the true mean. Our "confidence" is in the procedure which was used to produce the interval, in the sense that it will yield an interval containing the true mean p with probability 99.73%, as long as the data are from the postulated Normal model. ■

We could simulate data using R to verify that these confidence intervals do in fact contain the true parameter as often as expected. It would be also interesting to evaluate the statistical properties of the intervals when the data are from a different distribution. We do this in Section 9.3.2

9.3.1 *Pivotal Quantity approach*

The key observation that allows us to write the probability statement (9.3.1) in Example 9.3.1 is that $\sqrt{n}(\hat{p} - p) \sim \text{Normal}(0, 1)$. In other words, we have found a function of the data and the parameter of interest, namely,

$$T(X_1, X_2, \ldots, X_n, p) = \sqrt{n}(\hat{p} - p),$$

so that regardless of the value of p, $T(X_1, X_2, \ldots, X_n, p)$ has a completely known distribution. Such functions are sometimes called pivotal quantities. The derived confidence interval is completely specified, in principle, once we choose an interval in the support of the known distribution that has the desired probability. For instance, in the previous example, to get a confidence interval with coverage probability $\beta = 0.95$, we require an interval $[a, b]$ such that $P(Z \in [a, b]) = \beta$ for a standard Normal random variable Z. There are many such intervals, but intuitively, the interval $[-1.959964, 1.959964] \approx [-1.96, 1.96]$ is a good choice because it is the shortest such interval, as the density of Z is symmetric and decreases away from 0.

In general, we could choose any interval that has probability β. Popular alternative choices in the standard normal case are $(-\infty, \Phi^{-1}(\beta)]$ and $[\Phi^{-1}(1-\beta), \infty)$, which give "one-sided" confidence intervals. Once we choose a suitable interval $[a, b]$, the confidence "interval" for p is given by the set

$$\{p : T(X_1, X_2, \ldots, X_n, p) \in [a, b]\}.$$

This set is random as it depends on the random sample $X_1, X_2, \ldots, X_n$, but will have a specific realisation for any particular empirical sample. We will try to follow this approach in a few other situations.

Example 9.3.2. Let $\lambda > 0$ and $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from an Exponential population with rate λ. We are interested in obtaining a confidence interval for λ. Take

$$T(X_1, X_2, \ldots, X_n, \lambda) = n\lambda\overline{X}.$$

It is easy to see that $T(X_1, X_2, \ldots, X_n, \lambda)$ follows a Gamma$(n, 1)$ distribution. Thus, to obtain a confidence interval for λ with coverage probability β, we need to find an interval $[a, b] \subset [0, \infty)$ which has probability β under the Gamma$(n, 1)$ distribution.

Such an interval will depend on n. For example, with $n = 15$ and $\beta = 0.95$, one-sided intervals can be obtained as follows.

```
beta <- 0.95
c(0, qgamma(beta, shape = 15))
```

```
[1]  0.00000 21.88649
```

```
c(qgamma(1 - beta, shape = 15), Inf)
```

```
[1] 9.24633     Inf
```

As the Gamma distribution is not symmetric, the choice of a two-sided interval is not obvious. A simple choice is given by taking the exclusion probabilities on both tails to be equal, as follows.

```
alpha <- 1 - beta
c(qgamma(alpha / 2, shape = 15), qgamma(1 - alpha / 2, shape = 15))
```

```
[1]  8.395386 23.489621
```

This will typically not be the shortest interval. The shortest interval cannot be obtained in closed form, but can be computed numerically by varying the left and right tail exclusion probabilites together so that they add up to $1 - \beta$, and choosing the one giving the shortest interval. ■

Example 9.3.3. Let $\theta > 0$ and $X_1, X_2, \ldots, X_n$ be from the the Uniform$(0, \theta)$ distribution. We are interested in a confidence interval for θ. Take

$$T(X_1, X_2, \ldots, X_n, \theta) = \frac{X_{(n)}}{\theta}.$$

From Example 8.1.4, we know that $T(X_1, X_2, \ldots, X_n, \theta)$ has a Beta$(n, 1)$ distribution, which has an increasing density supported on $(0, 1)$. Thus the shortest interval of probability β will have right

endpoint 1. The left endpoint will depend on n. For example, with $n = 15$ and $\beta = 0.95$, the left endpoint is given by

```
qbeta(0.95, 15, 1)
```

```
[1] 0.9965863
```

The corresponding confidence interval for θ is therefore given by

$$\left\{\theta : 0.9965863 \leq \frac{X_{(n)}}{\theta} \leq 1\right\} = \left\{\theta : X_{(n)} \leq \theta \leq \frac{X_{(n)}}{0.9965863}\right\}. \qquad \blacksquare$$

In the next example we discuss two situations where the pivotal quantity approach will not work.

Example 9.3.4. Suppose we want to find a procedure to obtain confidence intervals for the mean parameter when the underlying distribution is Bernoulli or Poisson.

(a) Let $p > 0$ and $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the Bernoulli(p) distribution. We are interested in a confidence interval for p. Unfortunately, there is no obvious pivotal quantity $T(X_1, X_2, \ldots, X_n, p)$ in this example, so this approach does not work.

(b) Let $\lambda > 0$ and $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the Poisson(λ) distribution. We are interested in a confidence interval for λ. We see immediately that $T(X_1, \ldots, X_n, \lambda) = X_1 + \cdots + X_n$ has a Poisson($n\lambda$) distribution and consequently depends on λ. We could thus take $n = 1$ without loss of generality. However, there is again no obvious pivotal quantity $T(X_1, \lambda)$ in this example, so this approach does not work.

In both these examples, it is common practice to obtain approximate confidence intervals using the Central Limit Theorem. We discuss this approach in the next section. $\blacksquare$

We conclude this section with a final example, returning to the Normal distribution in Example 9.3.1, where we assume that both mean and variance are unknown.

Example 9.3.5. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a population distributed as Normal with unknown mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$. Suppose we are interested in estimating only the mean parameter μ.

If σ^2 were known, then a natural candidate for our pivotal quantity would have been

$$T(X_1, X_2, \ldots, X_n, \mu, \sigma) = \sqrt{n}\,\frac{(\overline{X} - \mu)}{\sigma}.$$

With a factor of 1.96 for a 95% coverage probability, the corresponding confidence interval for μ would then be

$$\overline{X} \pm 1.96\,\frac{\sigma}{\sqrt{n}},$$

which agrees with our intuition. Unfortunately, σ^2 is not known, so this approach is not valid. However, as we do have a natural estimator S^2 for σ^2 (recall Theorem 7.1.6), we can try replacing σ^2 by this estimate and arrive at

$$T(X_1, X_2, \ldots, X_n, \mu) = \sqrt{n}\,\frac{(\overline{X} - \mu)}{S}.$$

Recall from Corollary 8.1.11 that the exact distribution of $T(X_1, X_2, \ldots, X_n, \mu)$ is t_{n-1} and thus is a pivotal quantity (see also Exercise 9.3.1). We can now proceed as we did in Example 9.3.1. For the standard normal, the shortest interval with probability 0.95 was the symmetric interval $[-1.96, 1.96]$. For the t_{n-1} distribution, similar quantiles can be computed using the `qt()` function. For example, with $n = 15$, the right endpoint is given by

```
qt(0.975, df = 14)
```

```
[1] 2.144787
```

Considering again the $n = 15$ data points 11.22, 9.56, 10.06, 10.21, 10.95, 10.03, 10.75, 10.71, 11.42, 9.61, 10.91, 8.14, 8.95, 10.57, 11.1, we have $\hat{\mu} = \overline{X} = 10.279$ and $\hat{\sigma} = S = 0.9112$, so the confidence interval for μ is given by

$$\hat{\mu} \pm 2.145\,\frac{\hat{\sigma}}{\sqrt{n}} = 10.279 \pm 2.145\,\frac{0.9112}{\sqrt{15}} = [9.775, 10.784].$$

The ability to derive these kinds of confidence intervals, where we can control the confidence level exactly as long as the model assumptions hold, is one of the main reasons for studying the t distribution. ■

9.3.2 *Empirical Coverage Probability of Confidence Intervals*

The theoretical calculations in the previous section guarantee that the confidence intervals we derived will satisfy the target coverage, as long as the data come from the distribution assumed. It is still a good idea to verify this using simulation. Doing so will also allow us test how the coverage probabilities change when the data *do not* come from the postulated distribution.

We can use R to construct confidence intervals using the formulas derived above. For example, we can generate a random sample from the Normal distribution and then construct 80% and 95% confidence intervals for the mean using the following code.

```
x <- rnorm(20, mean = 10, sd = 3)
n <- length(x)
m <- mean(x)
s <- sd(x)
q80 <- qt(0.9, df = n - 1)
q95 <- qt(0.975, df = n - 1)
ci80 <- c(m + c(-1, 1) * q80 * s / sqrt(n))
ci95 <- c(m + c(-1, 1) * q95 * s / sqrt(n))
ci80
```

```
[1]  8.230541 10.388727
```

```
ci95
```

```
[1]  7.608558 11.010710
```

We can now repeat this process using `replicate()`, but to do so, it would be convenient to put the code above into an R function. Functions in R encapsulate repetitive calculations as code that can be run with different values of certain variables that are provided to the function as *arguments*. In this case, we would like to repeat the above process many times, but with different data, and possibly different confidence levels. So, we can create a function that takes two arguments, `x` and `level`, and repeats the calculations above.

```
normalMeanCI <- function(x, level)
{
    n <- length(x)
    m <- mean(x)
    s <- sd(x)
    alpha <- 1 - level
    qlevel <- qt(1 - alpha / 2, df = n - 1)
    ci <- c(m + c(-1, 1) * qlevel * s / sqrt(n))
    ci
}
```

When the function is called with these two arguments, the value of the last expression in the function is returned as its value. So, we can now repeat our earlier calculations using this more general function as follows.

```
normalMeanCI(x, level = 0.8)
```

```
[1]  8.230541 10.388727
```

```
normalMeanCI(x, level = 0.95)
```

```
[1]  7.608558 11.010710
```

We can also repeat this experiment multiple times using `replicate()`.

```
t(replicate(10, normalMeanCI(rnorm(20, mean = 10, sd = 3), level = 0.8)))
```

```
           [,1]     [,2]
 [1,]  8.843671 10.18292
 [2,]  8.803891 10.46128
 [3,]  9.327093 10.80719
 [4,]  9.176958 10.63078
 [5,]  9.625395 11.23960
 [6,]  8.745298 10.96407
 [7,] 10.237023 12.58567
 [8,] 10.148164 11.66396
 [9,]  9.636105 11.04266
[10,]  9.099662 10.42231
```

Of course, to estimate the coverage probability we need to repeat this process a much larger number of times and check how many times the interval contains the true mean. We can do this for 10000 replications as follows.

```
## each experiment samples 20 independent Normal random variables, and
## produces 80% Confidence interval
cirep <- t(replicate(10000, # repeat experiment 1000 times.
                     normalMeanCI(rnorm(20, mean = 10, sd = 3),  level = 0.8)))
## count proportion of confidence interval that contain the true mean, 10.
mean(cirep[,1] <= 10 & cirep[,2] >= 10)
```

```
[1] 0.795
```

We leave it to the reader to verify that the coverage probabilities seen in simulation match the target level regardless of the mean, standard deviation or sample sizes.

A related question is how the length of the confidence intervals, which are random quanities, behave. The empirical distribution of the lengths obtained in the 10000 replications above can be summarized as follows.

```
summary(cirep[,2] - cirep[,1])
```

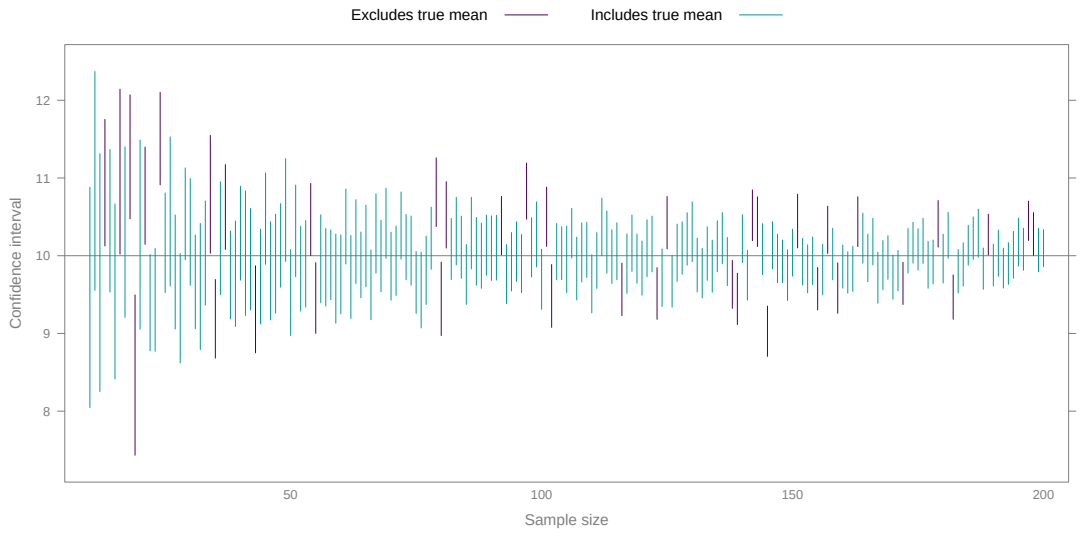


Figure 9.2: Simulated 80% confidence intervals for the mean of a Normal population with mean 10 and variance 3^2, computed assuming mean and variance are both unknown. The sample sizes used are from $10, 11, \ldots, 200$. The interval widths decrease on average as sample size increases. We expect roughly 1 in 5 intervals to exclude the true mean (shown in a different color) regardless of sample size as data are generated from a Normal distribution.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.7437  1.5547  1.7458  1.7562  1.9491  3.0018
```

From the formula it is clear that the length is inversely proportional to $\sqrt{n}$ and proportional to S. As S^2/σ^2 follows a χ^2_{n-1} distribution, the square of the length will have a scaled χ^2_{n-1} distribution, where the scaling factor depends on σ and n. Figure 9.2 plots simulated confidence intervals obtained from random samples generated from the normal distribution, with varying sample sizes but fixed mean and varince.

In real life, we cannot expect data to follow a normal distribution exactly. If we knew the true distribution, we could perhaps derive an appropriate confidence interval, but this is rarely the case. Even if we knew the true distribution, a suitable confidence interval may not be easy to derive. Thus, an important practical question is how the normal confidence interval performs as a confidence interval for the population mean when the data do not arise from the normal distribution.

The answer obviously depends on the distribution from which the data *do* arise. Below we consider four specific examples that represent different types of departures from normality

Example 9.3.6. Consider data $X_1, X_2, \ldots, X_n$ from the Bernoulli(p) distribution, for some unknown $p \in (0, 1)$. We saw earlier that the pivotal quantity approach did not work for this model. However, as the parameter of interest p is the mean, we could blindly apply the normal confidence interval formula and expect to get something reasonable.

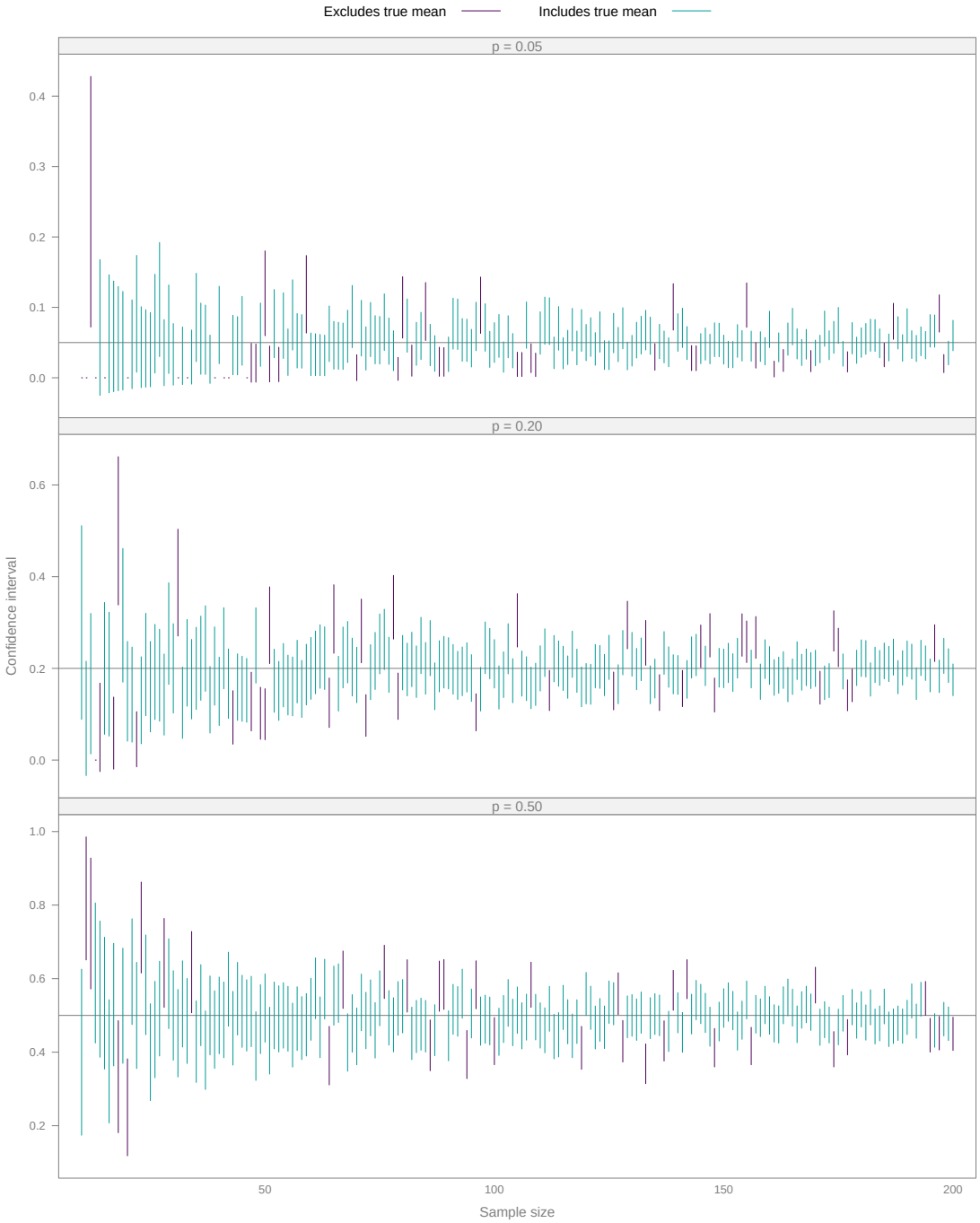


Figure 9.3: Simulated 80% confidence intervals for Bernoulli with probability $p = 0.05$, 0.2, and 0.5, computed using the normal confidence interval formula for the mean when mean and variance are both unknown. The sample sizes used are from $10, 11, \ldots, 200$. Notice that some intervals go below zero, especially for $p = 0.05$, where in addition, some intervals consist of the single point $\{0\}$. This happens when all outcomes are 0, which may happen for small n and small p.

We have already generated one Bernoulli sample above, in the experiment where we computed confidence intervals from normal data 10000 times. If we denote by X_i whether the i-th interval contained the true mean, then the vector of observed X_i values is given by

```
x <- cirep[,1] <= 10 & cirep[,2] >= 10
```

We had previously estimated $P(X_1 = 1)$ by the sample proportion

```
mean(x)
```

```
[1] 0.795
```

which is close to the nominal coverage probability 0.8. However, in view of the current discussion, we would be more reassured if we see that the value 0.8 is included in a reasonable confidence interval. We obtain the following 95% interval by applying the normal confidence interval formula.

```
normalMeanCI(x, level = 0.95)
```

```
[1] 0.7870862 0.8029138
```

This is of course just one example, but we can evaluate the performance of the method using the same techniques as above, replacing the data generating process to simulate Bernoulli data instead of normal. Figure 9.3 plots confidence intervals for data generated using Bernoulli(p) for $p = 0.05$, 0.2, and 0.5, with the setup otherwise similar to Figure 9.2. This plot illustrates some problems with small p, which are also present for large p close to 1, but otherwise suggests reasonable performance.

To estimate coverage probability and average length for specific combinations of p and n, we can use the replication approach.

```
cisummaryBernoulli <- function(p, n, level) {
    cirep <- t(replicate(10000,
                         normalMeanCI(rbinom(n, size = 1, prob = p),
                                      level = level)))
    data.frame(n = n, p = p,
               coverage = mean(cirep[,1] <= p & cirep[,2] >= p),
               avg_length = mean(cirep[,2] - cirep[,1]))
}
```

```
rbind(cisummaryBernoulli(p = 0.05, n = 10,   level = 0.95),
      cisummaryBernoulli(p = 0.05, n = 100,  level = 0.95),
      cisummaryBernoulli(p = 0.05, n = 1000, level = 0.95),
      cisummaryBernoulli(p = 0.20, n = 10,   level = 0.95),
      cisummaryBernoulli(p = 0.20, n = 100,  level = 0.95),
      cisummaryBernoulli(p = 0.20, n = 1000, level = 0.95),
      cisummaryBernoulli(p = 0.50, n = 10,   level = 0.95),
      cisummaryBernoulli(p = 0.50, n = 100,  level = 0.95),
      cisummaryBernoulli(p = 0.50, n = 1000, level = 0.95))
```

n	p	coverage	avg_length
10	0.05	0.4059	0.19823628
100	0.05	0.8781	0.08462128
1000	0.05	0.9463	0.02701204
10	0.20	0.8879	0.53350536
100	0.20	0.9438	0.15823556
1000	0.20	0.9466	0.04963809
10	0.50	0.9778	0.71138220
100	0.50	0.9400	0.19840636
1000	0.50	0.9471	0.06205443

It is reassuring to see that for large n the coverage probability is close to the target of 95%, with the average interval length decreasing with n. However, for small n and small p, the observed coverage probability is substantially smaller than the target. In the next section, we will compare these results with other approximate confidence intervals obtained using asymptotic results. ■

Example 9.3.7. Consider data $X_1, X_2, \ldots, X_n$ from the Exponential(λ) distribution, for some unknown $\lambda > 0$. Here we have an exact confidence interval for λ based on the pivotal quantity $n\lambda\overline{X}$ which follows a Gamma$(n, 1)$ distribution. However, as λ is the mean, we can try using the normal confidence interval as well. Below we contrast the coverage probability and average interval length for the two methods for $\lambda = 1$.

```
exponentialMeanCI <- function(x, level)
{
    alpha <- 1 - level
    qgamma(c(alpha / 2, 1 - alpha / 2), shape = length(x)) / sum(x)
}
```

```
cisummaryExponential <- function(n, level) {
    cirepNorm <- t(replicate(10000, normalMeanCI(rexp(n), level = level)))
    cirepExp <- t(replicate(10000, exponentialMeanCI(rexp(n), level = level)))
    data.frame(n = n,
               coverageNorm = mean(cirepNorm[,1] <= 1 & cirepNorm[,2] >= 1),
               coverageExp  = mean(cirepExp[,1]  <= 1 & cirepExp[,2]  >= 1),
               lengthNorm = mean(cirepNorm[,2] - cirepNorm[,1]),
               lengthExp  = mean(cirepExp[,2]  - cirepExp[,1]),
               negative = mean(cirepNorm[,1] < 0))
}
```

```
rbind(cisummaryExponential(n = 10,   level = 0.95),
      cisummaryExponential(n = 100,  level = 0.95),
      cisummaryExponential(n = 1000, level = 0.95))
```

n	coverageNorm	coverageExp	lengthNorm	lengthExp	negative
10	0.8998	0.9506	1.3237152	1.3655739	0.0425
100	0.9397	0.9507	0.3934190	0.3955935	0.0000
1000	0.9461	0.9505	0.1239441	0.1240712	0.0000

Here again we observe that for $n = 10$, the normal interval has lower than desired coverage probability, possibly stemming from the lower interval length on average. As in the Bernoulli case with small p, the normal interval goes below zero in a small proportion of cases. As n increases, however, the performance of the normal interval becomes comparable with the exponential interval. ■

Example 9.3.8. Consider data $X_1, X_2, \ldots, X_n$ from the Cauchy(θ, α^2) distribution, for some unknown location $\theta \in \mathbb{R}$ that we are interested in estimating and an unknown scale $\alpha > 0$. The Cauchy distribution does not have finite mean, so it is unclear whether it makes sense to use a confidence interval designed for the population mean. However, we can still blindly apply it and see what happens. In the simulation below, we take $\theta = 0$ and $\alpha^2 = 1$.

```
cisummaryCauchy <- function(n, level) {
    cirepNorm <- t(replicate(10000, normalMeanCI(rcauchy(n), level = level)))
    data.frame(n = n,
               coverage = mean(cirepNorm[,1] <= 0 & cirepNorm[,2] >= 0),
               meanLength = mean(cirepNorm[,2] - cirepNorm[,1]),
               medianLength = median(cirepNorm[,2] - cirepNorm[,1]))
}
rbind(cisummaryCauchy(n = 10,   level = 0.95),
      cisummaryCauchy(n = 100,  level = 0.95),
      cisummaryCauchy(n = 1000, level = 0.95))
```

n	coverage	meanLength	medianLength
10	0.9823	25.31168	5.319377
100	0.9789	24.08324	4.687978
1000	0.9809	25.20705	4.681542

Here we observe a phenomenon that we have not seen in the ealier examples. Although the coverage probabilities look reasonable, the average length of the confidence intervals (as measured either by the mean or the median) do not decrease as sample size increases, and in fact can be quite unstable. In other words, we we have no improvement in the accuracy of estimation with increasing sample size. The reason for this seemingly unexpected behaviour is that the sample mean is *not* a reliable estimator of the Cauchy location parameter. We will see an alternative confidence interval with better performance in the next section. ■

Example 9.3.9. Real life data do not always follow theoretical assumptions precisely, but as we have seen in the previous examples, Normal confidence intervals often work well for estimating the population mean regardless, especially for large sample sizes. This is essentially a consequence the Central Limit Theorem, which we will use with more finesse in the next section.

Another kind of departure from theoretical assumptions we sometimes see in real data are outliers or other kinds of contaminations that affect individual data points rather than the full sample. In this example, we will consider data $X_1, X_2, \ldots, X_n$ that mostly come from the Normal$(0, 1)$ distribution, but sometimes, with probability 0.01 or 1%, come from the Normal$(0, 100)$ distribution. We can write an R function to generate such a mixture as follows.

```
rmixnorm <- function(n, prob = 0.01) {
    rnorm(n, mean = 0, sd = ifelse(runif(n) < prob, 100, 1))
}
```

We can now simulate Normal confidence intervals as before, and summarize their properties for various values of n.

```
cisummaryMixture <- function(n, level) {
    cirepNorm <- t(replicate(10000, normalMeanCI(rmixnorm(n), level = level)))
    data.frame(n = n,
               coverage = mean(cirepNorm[,1] <= 0 & cirepNorm[,2] >= 0),
               meanLength = mean(cirepNorm[,2] - cirepNorm[,1]))
}
rbind(cisummaryMixture(n = 10,   level = 0.95),
      cisummaryMixture(n = 100,  level = 0.95),
      cisummaryMixture(n = 1000, level = 0.95))
```

n	coverage	meanLength
10	0.9535	4.817333
100	0.9779	2.771039
1000	0.9593	1.195399

Here again the coverage probabilities are close to the target and the confidence intervals decrease in size as sample size increases. ■

9.3.3 *Approximate Confidence Intervals using CLT*

We have seen in the previous section that the Normal confidence intervals obtained from data $X_1, X_2, \ldots, X_n$ perform quite reasonably even when the underlying distribution is not Normal, with the exception of the Cauchy distribution. This essentially follows from the Central Limit Theorem (Theorem 8.4.1), which can be interpreted as saying that for large n we have the approximation

$$P\left(\left|\sqrt{n}\frac{(\overline{X}-\mu)}{\sigma}\right| \leq 1.96\right) \approx 0.95$$

In the simulation examples in the previous section, we have replaced σ by the sample standard deviation S, and the normal quantile 1.96 by the corresponding t_{n-1} quantile. As $n \to \infty$, the t_{n-1} becomes unnecessary, and it can be shown that

$$\sqrt{n}\frac{(\overline{X}-\mu)}{S} \xrightarrow{d} \text{Normal}(0,1)$$

under reasonable moment conditions on the underlying distribution (see Example 8.3.7). In this section, we will see whether applying the Central Limit Theorem to *specific* models can provide improved confidence intervals.

EXAMPLE 9.3.10. Consider data $X_1, X_2, \ldots, X_n$ from the Bernoulli(p) distribution, for some unknown $p \in (0,1)$. We saw in Example 9.3.6 that the normal confidence interval performs reasonably well for large n but not for small n, especially when p is small. Let us see if the Central Limit Theorem for this specific model leads us to a better interval.

The Normal confidence interval we have been using so far assumes that both mean and variance are unknown. However, the Bernoulli model has only one parameter p, on which both mean and variance depend. Specifically, the Central Limit Theorem (See Example 8.4.5) gives the asymptotic distribution of the sample proportion $\hat{p}$ as

$$T(X_1, X_2, \ldots, X_n, p) = \sqrt{n}\,\frac{(\hat{p}-p)}{\sqrt{p(1-p)}} \xrightarrow{d} \text{Normal}(0,1).$$

This $T(X_1, X_2, \ldots, X_n, p)$ can be viewed as an approximately pivotal quantity, and following our earlier approach, an approximate 95% confidence interval for p is given by

$$\left\{p : -Q \leq \sqrt{n}\,\frac{(\hat{p}-p)}{\sqrt{p(1-p)}} \leq Q\right\}, \tag{9.3.2}$$

where $Q = 1.96$ is the 0.975 quantile of standard Normal. The bounds will be achieved by p that satisfy the quadratic equation

$$n(\hat{p}-p)^2 = \frac{Q^2}{n}\,p(1-p).$$

It it easy to verify that this equation has two real solutions, given by

$$\frac{1}{(C_n+1)}\left(\frac{C_n}{2}+\hat{p}\pm\frac{\sqrt{C_n^2+4C_n\hat{p}(1-\hat{p})}}{2}\right)$$

where $C_n = Q^2/n$. These are called Wilson Confidence intervals [Wil27]. Ignoring the possibility that these solutions may lie outside $[0,1]$, we can compute this confidence interval in R as follows

```
bernoulliQuadraticCI <- function(x, level)
{
    n <- length(x)
    phat <- mean(x)
    Q <- qnorm(1 - (1 - level) / 2)
    C <- Q^2 / n
    ci <- ((C + 2 * phat) + c(-1, 1) *
           sqrt(C^2 + 4 * C * phat * (1-phat))) / (2 * (C + 1))
    ci
}
```

To avoid the complication of solving a quadratic equation, a common alternative is to replace the unknown p in the denominator of (9.3.2) by the sample proportion $\hat{p}$. This is equivalent to using a corollary of the Central Limit Theorem for the sample proportion, which says that (see Exercise 9.3.5)

$$\widetilde{T}(X_1, X_2, \ldots, X_n, p) = \sqrt{n}\,\frac{(\hat{p}-p)}{\sqrt{\hat{p}(1-\hat{p})}} \xrightarrow{d} \text{Normal}(0,1). \tag{9.3.3}$$

It is then trivial to see that the confidence interval for p is given by solving linear equations as

$$\hat{p} \pm Q\sqrt{\hat{p}(1-\hat{p})/n}.$$

This simpler interval can be computed in R as follows.

```
bernoulliLinearCI <- function(x, level)
{
    n <- length(x)
    phat <- mean(x)
    Q <- qnorm(1 - (1 - level) / 2)
    ci <- phat + c(-1, 1) * Q * sqrt(phat * (1 - phat) / n)
    ci
}
```

We can now repeat our earlier simulation experiment from Example 9.3.6, this time with these intervals instead of the normal interval.

```
cisummaryBernoulli <- function(p, n, level) {
    cirepQ <- t(replicate(10000,
                          bernoulliQuadraticCI(rbinom(n, size = 1, prob = p),
                                               level = level)))
    cirepL <- t(replicate(10000,
                          bernoulliLinearCI(rbinom(n, size = 1, prob = p),
                                            level = level)))
    data.frame(n = n, p = p,
               coverageQ = mean(cirepQ[,1] <= p & cirepQ[,2] >= p),
               coverageL = mean(cirepL[,1] <= p & cirepL[,2] >= p),
               meanLengthQ = mean(cirepQ[,2] - cirepQ[,1]),
               meanLengthL = mean(cirepL[,2] - cirepL[,1]))
}
```

```
rbind(cisummaryBernoulli(p = 0.05, n = 10,   level = 0.95),
      cisummaryBernoulli(p = 0.05, n = 100,  level = 0.95),
      cisummaryBernoulli(p = 0.05, n = 1000, level = 0.95),
      cisummaryBernoulli(p = 0.20, n = 10,   level = 0.95),
      cisummaryBernoulli(p = 0.20, n = 100,  level = 0.95),
      cisummaryBernoulli(p = 0.20, n = 1000, level = 0.95),
      cisummaryBernoulli(p = 0.50, n = 10,   level = 0.95),
      cisummaryBernoulli(p = 0.50, n = 100,  level = 0.95),
      cisummaryBernoulli(p = 0.20, n = 1000, level = 0.95))
```

n	p	coverageQ	coverageL	meanLengthQ	meanLengthL
10	0.05	0.9088	0.3992	0.32787000	0.15990182
100	0.05	0.9686	0.8821	0.08827661	0.08290597
1000	0.05	0.9487	0.9408	0.02712679	0.02697052
10	0.20	0.9636	0.8842	0.43447816	0.43615822
100	0.20	0.9375	0.9319	0.15429259	0.15563230
1000	0.20	0.9483	0.9473	0.04948471	0.04953288
10	0.50	0.9801	0.8900	0.50689814	0.58510371
100	0.50	0.9449	0.9452	0.19141843	0.19503036
1000	0.20	0.9456	0.9449	0.04951000	0.04955135

Comparing with the results for the Normal interval, we see that the simpler linear interval has roughly similar performance. However, the more complicated quadratic interval has substantially better performance for small n and small p. It is important to remember that all these intervals are based on asymptotic results, and it is difficult to analyse performance in small samples except through simulation. Nonetheless, it is not surprising that the methods which replace the variance term by an estimate do worse than the method that does not. ■

9.3.4 *Confidence Intervals for the Population Median*

Just as the Central Limit Theorem for the sample mean allows us to formulate approximate confidence intervals for the population mean, we can try to use the limiting distribution of the sample median to obtain a confidence interval for the population median. For symmetric distributions, the population median coincides with the population mean, so the two intervals are for the same population parameter.

Let $X_1, X_2, \dots$ be i.i.d. random variables with probability density function f with median μ. Theorem 8.6.2 tells us that if $f(\mu) > 0$, then

$$2\sqrt{n} f(\mu) \left(\widetilde{X}_n - \mu \right) \xrightarrow{d} Z,$$

where $\widetilde{X}_n$ is the sample median obtained from $X_1, X_2, \dots, X_n$. We cannot use this result directly to get a confidence interval for μ, because $f(\mu)$ is unknown. To estimate it, we shall proceed with the following approximations. First from the fact that $F^{-1}(F(\mu)) = 1$ and an application of chain rule

$$\begin{aligned} f(\mu) &= F'(\mu) \\ &= \frac{1}{(F^{-1})'(F(\mu))} \\ &\approx \frac{2}{\sqrt{n}\left(F^{-1}(\frac{1}{2} + \frac{1}{\sqrt{n}}) - F^{-1}(\frac{1}{2} - \frac{1}{\sqrt{n}}) \right)} \end{aligned}$$

Now if $X_{(r)}$ is the r-th order statistic derived from $X_1, X_2, \dots, X_n$ (see Section 8.1.1) then it is intuitively clear that

$$F^{-1}(\frac{1}{2} + \frac{1}{\sqrt{n}}) \approx X_{([n(\frac{1}{2} + \frac{1}{\sqrt{n}})])} \qquad \text{and} \qquad F^{-1}(\frac{1}{2} - \frac{1}{\sqrt{n}}) \approx X_{([n(\frac{1}{2} - \frac{1}{\sqrt{n}})])}.$$

Thus we arrive at our plausible estimate

$$\hat{f}_n(\mu) = \frac{2}{\sqrt{n}(X_{([n(\frac{1}{2} + \frac{1}{\sqrt{n}})])} - X_{([n(\frac{1}{2} - \frac{1}{\sqrt{n}})])})},$$

for $f(\mu)$. This approximation above can be made rigorous and can be used to provide confidence intervals by using the lemma below. We refer the reader to [Ser09] for a complete treatment of convergences of sample quantiles.

Lemma 9.3.11. *Suppose $X_1, X_2, \dots$ be an i.i.d. sequence from X. Suppose X is a continuous random variable with probability density function $f : \mathbb{R} \to \mathbb{R}$. Suppose $f(\mu) > 0$ where μ is the median of f. Then*

$$2\sqrt{n} \hat{f}_n(\mu) \left(\widetilde{X}_n - \mu \right) \xrightarrow{d} Z, \tag{9.3.4}$$

Proof. From [Ser09, Corolllary, Section 2.5.2] we have

$$\hat{f}_n(\mu) \xrightarrow{p} f(\mu)$$

as $n \to \infty$. An application of the Central Limit Theorem for the median (Theorem 8.6.2) along with Slutsky's Theorem (Lemma 8.3.10) yields the result. ■

Now using (9.3.4) we have the confidence interval

$$\widetilde{X}_n \pm \frac{1.96}{2\sqrt{n}\hat{f}_n(\mu)}$$

for confidence level 0.95, with the numerator on the right hand side adjusted for other confidence levels. We can calculate this confidence interval using R as follows.

```
medianCI <- function(x, level)
{
    n <- length(x)
    m <- median(x)
    delta <- diff(quantile(x, 0.5 + c(-1, 1) / sqrt(n),
                           names = FALSE))
    fhat <- 2 / (delta * sqrt(n))
    qlevel <- qnorm(1 - (1 - level) / 2)
    ci <- c(m + c(-1, 1) * qlevel / (2 * sqrt(n) * fhat))
    ci
}
```

We apply this in three examples below.

Example 9.3.12. Let $X_1, X_2, \ldots, X_n$ be i.i.d. Normal (μ, σ^2). In the simulation study below, we keep the true mean and variance fixed at 0 and 1, but assume that they are both unknown. The sample size is varied to see how well the median approximation works for small samples.

```
cisummaryMedianNormal <- function(n, level) {
    cirepMean <- t(replicate(10000, normalMeanCI(rnorm(n), level = level)))
    cirepMedian <- t(replicate(10000, medianCI(rnorm(n), level = level)))
    data.frame(n = n,
               coverageMean   = mean(cirepMean[,1] <= 0 & cirepMean[,2] >= 0),
               coverageMedian = mean(cirepMedian[,1] <= 0 & cirepMedian[,2] >= 0),
               lengthMean     = mean(cirepMean[,2] - cirepMean[,1]),
               lengthMedian   = mean(cirepMedian[,2]  - cirepMedian[,1]))
}

rbind(cisummaryMedianNormal(n = 10,   level = 0.95),
      cisummaryMedianNormal(n = 100,  level = 0.95),
      cisummaryMedianNormal(n = 1000, level = 0.95))
```

n	coverageMean	coverageMedian	lengthMean	lengthMedian
10	0.9501	0.9252	1.3931430	1.5160993
100	0.9467	0.9396	0.3958827	0.4913522
1000	0.9532	0.9463	0.1240568	0.1554546

The coverage of the median interval is reasonable, although a little lower than the target for smaller n. Observe that on average the length of the median interval is larger than the mean interval, in spite of the lower coverage probability. This is related to the asymptotic efficiency of the median discussed in the previous chapter. ■

Example 9.3.13. Median intervals are potentially more useful when the underlying distribution is not Normal. An extreme case illustrating this is the Cauchy distribution. Continuing Example 9.3.8, consider data $X_1, X_2, \ldots, X_n$ from the Cauchy(θ, α^2) distribution, where we are interested in estimating the unknown location $\theta \in \mathbb{R}$, and the unknown scale $\alpha > 0$ is a nuisance parameter. As we saw in Example 9.3.8, the Normal mean confidence inteval completely fails in this case.

Repeating the experiment with the median interval, we get the following.

```
cisummaryMedianCauchy <- function(n, level) {
    cirepNorm <- t(replicate(10000, medianCI(rcauchy(n), level = level)))
    data.frame(n = n,
               coverage = mean(cirepNorm[,1] <= 0 & cirepNorm[,2] >= 0),
               meanLength = mean(cirepNorm[,2] - cirepNorm[,1]),
               medianLength = median(cirepNorm[,2] - cirepNorm[,1]))
}
rbind(cisummaryMedianCauchy(n = 10,   level = 0.95),
      cisummaryMedianCauchy(n = 100,  level = 0.95),
      cisummaryMedianCauchy(n = 1000, level = 0.95))
```

n	coverage	meanLength	medianLength
10	0.9741	3.6774258	2.8444431
100	0.9504	0.6397202	0.6304624
1000	0.9457	0.1951665	0.1942360

With the median confidence interval, we get reasonable coverage with the average interval lengths decreasing with sample size, as we would expect. ■

Example 9.3.14. Continuing Example 9.3.9, consider the following model that mimics low-probability contamination of the data by outliers. Suppose $X_1, X_2, \ldots, X_n$ come from the Normal$(0, 1)$ distribution with probability 0.99, but with probability 0.01, they arise from the Normal$(0, 100)$ distribution.

Although we have not compared properties of the sample mean and the sample median, one intuitively obvious property of the median is that changing the values of a few extreme data points will not affect it, whereas it might affect the sample mean. We can thus expect the median to be more *stable* under this contamination model, in the sense that it will have lower variance. Although we will not do it here, we can formalize this property in terms of the relative asymptotic efficiency of the sample median over the sample median. In the last chapter, we noted that the sample mean was more efficient than the sample mean when data are obtained from the normal distribution, and this explains the wider median confidence intervals in Example 9.3.12. However, the situation is reversed in this model, where the median is more efficient, and will lead to narrower confidence intervals on average. We can see this in the results of the following simulation.

```
cisummaryMedianMixture <- function(n, level) {
    cirepMean <- t(replicate(10000, normalMeanCI(rmixnorm(n), level = level)))
    cirepMedian <- t(replicate(10000, medianCI(rmixnorm(n), level = level)))
    data.frame(n = n,
               coverageMean   = mean(cirepMean[,1] <= 0 & cirepMean[,2] >= 0),
               coverageMedian = mean(cirepMedian[,1] <= 0 & cirepMedian[,2] >= 0),
               lengthMean = mean(cirepMean[,2] - cirepMean[,1]),
               lengthMedian = mean(cirepMedian[,2] - cirepMedian[,1]))
}
rbind(cisummaryMedianMixture(n = 10,   level = 0.95),
      cisummaryMedianMixture(n = 100,  level = 0.95),
      cisummaryMedianMixture(n = 1000, level = 0.95))
```

n	coverageMean	coverageMedian	lengthMean	lengthMedian
10	0.9563	0.9343	4.957189	1.5930168
100	0.9763	0.9367	2.794230	0.4943884
1000	0.9590	0.9483	1.201950	0.1569595

As we can see, the coverage probabilities in both intervals are reasonable close to the target level of 95%, but the median intervals are much narrower on average. ■

EXERCISES

Ex. 9.3.1. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a Normal population with unknown mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$. Show that the function

$$T(X_1, X_2, \ldots, X_n, \mu) = \sqrt{n}\,\frac{(\overline{X} - \mu)}{S}.$$

is a pivotal quantity. (Hint: define $Z_i = \frac{X_i - \mu}{\sigma}$ and show that $T(X_1, X_2, \ldots, X_n, \mu) = T(Z_1, Z_2, \ldots, Z_n, 0)$. The distribution of $T(Z_1, Z_2, \ldots, Z_n, 0)$ clearly does not depend on μ or σ^2.)

Ex. 9.3.2. $X_1, X_2, \ldots, X_n$ i.i.d. $Uniform(0, \theta)$, Confidence interval for θ (one-sided? shortest length of the form $[aX_{(n)}, bX_{(n)}]$)

Ex. 9.3.3. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a Normal $\left(\mu, \sigma^2\right)$ distribution. Take

$$T(X_1, X_2, \ldots, X_n, \sigma^2) = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2}.$$

It is easy to see from Theorem 8.1.10 that $T(X_1, X_2, \ldots, X_n, \sigma^2)$ has a χ^2_{n-1} distribution. Use this to obtain a confidence interval for σ^2.

Ex. 9.3.4. Let $\sigma > 0$ and $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the Normal$(0, \sigma^2)$ distribution.

(a) Find the distribution of

$$T(X_1, X_2, \ldots, X_n, \sigma^2) = \sum_{i=1}^{n} \frac{X_i^2}{\sigma^2},$$

(b) Using `qchisq()` in R, find a 95% confidence interval for σ^2 with equal tail exclusion probabilites.

(c) Using `qchisq()` in R, find a 95% one-sided confidence interval for σ^2.

Ex. 9.3.5. Consider an i.i.d. sample $X_1, X_2, \ldots$ from Bernoulli(p) distribution. Let

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

(a) Show that $\hat{p} \xrightarrow{p} p$ as $n \to \infty$.

(b) Using Slutsky's theorem (Lemma 8.3.10) show (9.3.3).

Ex. 9.3.6. Suppose $X_1, X_2, X_3, \ldots$ are i.i.d. X with $E[X] = p$ and $\mathrm{Var}[X] = \sigma^2$. Let $\hat{p} = \overline{X}_n$.

(a) Suppose $X \sim$ Bernoulli(p) then using the variance stabilising formula $g(x) = \arcsin(\sqrt{x})$ construct a 95% confidence interval for p.

(b) Construct confidence intervals using the variance stablisling formula from (a), `R`-code below and `cisummaryBernoulli` discussed in Example 9.3.10.

```
bernoulliVSTCI <- function(x, level)
{
    n <- length(x)
    phat <- mean(x)
    Q <- qnorm(1 - (1 - level) / 2)
    tci <- asin(sqrt(phat)) + c(-1, 1) * Q / (2 * sqrt(n))
    ci <- sin(tci)^2
    ci
}
```

(c) Are the intervals larger in comparison to those in Example 9.3.10?

Ex. 9.3.7. Repeat Exercise 9.3.6 when $X \sim \text{Poisson}(10)$ with $g(x) = \sqrt{x}$.

10

HYPOTHESIS TESTING

In Chapter 9 we looked at the problem of estimating the value of unknown parameters in a probability model, where the probability model prescribes a distribution for an attribute or quantity of interest in the population. The estimation is based on a sample of data from the population that we assume have been generated from that model. In many situations however, estimating the parameters is not the primary problem of interest. Rather, we may be interested in whether two or more subpopulations are *different*, or determining whether two or more measured attributes are *independent* of others. Examples of such problems abound:

- Are temperatures on average higher now than they were a hundred years ago?
- Are people with higher blood glucose levels at age 30 more likely to develop diabetes by age 60?
- Does smoking decrease life expectancy? Does eating organic food increase life expectancy?

One way in which we can formulate and answer such questions is by viewing as the "test of a hypothesis", which is a central problem in statistics. A hypothesis test makes a conjecture or hypothesis about the population (e.g., an attribute has the same distribution in two subpopulations, or two attributes are independent) and then carries out a computation to test the credibility of the conjecture. Probability theory is important in hypothesis testing because hypotheses are based on probability models, and the computations required to arrive at a decision are done using probabilistic techniques.

In this chapter we will discuss simplified prototype hypothesis testing problems. These may not be as nuanced as many real life hypothesis testing problems but will convey their essence. The techniques described are in fact practically useful in a wide range of situations.

10.1 INTRODUCTION

We begin our discussion on testing with a simple example.

Example 10.1.1. Recall the example from Chapter 9 of a coin which we assume has a probability p of showing heads each time it is flipped. The results of flipping the coin 100 times are viewed as i.i.d. random variables $X_1, X_2, \ldots, X_{100}$, each with a Bernoulli($p$) distribution. This is a special case of the multinomial distribution, a fact that will become useful in a later example. As there are only two possible outcomes here, there is only one parameter, the probability p of seeing heads.

Suppose $\sum_{n=1}^{100} X_n = 67$ of these flips showed heads. In the previous chapter we considered what inferences we can make about the value of p. However, suppose our primary interest is only in deciding whether $p = 0.5$; that is, whether or not the coin is fair. Testing the conjecture (or hypothesis) that $p = 0.5$ involves making a judgement about whether the observed data is consistent with the conjecture. Given what we have learned in the previous chapter, one simple approach

could be to say "yes" if a suitable confidence interval for p includes the value 0.5. This is a perfectly reasonable approach for this problem, but for now we will take an alternative, more direct approach that is easier to generalize to other testing problems. In this case, for example, we may argue as follows: "If the coin had an equal chance of showing heads or tails, then the probability of observing 67 heads or more in 100 flips is around 0.0004 (this can be verified using R). This number is small enough to suggest that the hypothesis of heads and tails being equally likely is inaccurate". ■

We will try to make this intuition concrete in a way that can be generalized to other situations. In Chapter 9 we introduced the idea of maximum likelihood as a systematic approach to estimation. In this chapter, we will continue using it to develop a unified approach to hypothesis testing. We will see that this approach leads to useful testing procedures in many situations. However, we will also come across situations where it does not. In practice we often resort to ad hoc procedures instead. Such procedures, while important and useful, are beyond the scope of this book.

As before, we assume that the sample $X_1, X_2, \ldots, X_n$ are i.i.d. copies of a random variable X with a probability mass function or probability density function $f(x)$, where $f(x) = f(x \mid p_1, p_2, \ldots, p_d) = f(x \mid p)$ depends on one or more unknown parameters $p = (p_1, p_2, \ldots, p_d) \in \mathcal{P} \subset \mathbb{R}^d$ for some $d \geq 1$. A conjecture or "null hypothesis" about X restricts the possible values that p can take, and is represented by the statement that $p \in \mathcal{P}_0$, where $\mathcal{P}_0$ is a proper subset of $\mathcal{P}$. Just as we computed the maximum likelihood estimator $\hat{p}$ as the value of p that maximizes the likelihood function, we can also compute an estimator $\hat{p}_0$ that maximizes the likelihood function within this smaller subset $\mathcal{P}_0$. If the null hypothesis holds, we expect the likelihoods at $\hat{p}$ and $\hat{p}_0$ to be close to each other, whereas we expect the likelihood at $\hat{p}_0$ to be substantially smaller if the null hypothesis does not hold.

Our goal is to develop this idea into an approach to hypothesis testing. In the formal testing problem described above, this approach will lead to a "test statistic", which is a function of the data $X_1, X_2, \ldots, X_n$. Naturally, the probability distribution of this statistic will depend on the unknown parameter $p \in \mathcal{P}$. To "test" the conjecture that $p \in \mathcal{P}_0$ for a specified $\mathcal{P}_0 \subset \mathcal{P}$, one essentially asks whether the observed value of the test statistic could conceivably have come from a value of $p \in \mathcal{P}_0$, quantifying the degree of this possibility through a probability that is referred to as the "p-value". The only requirement, albeit a very important one, is that this p-value not depend on the unknown parameter p, which is equivalent to saying that the distribution of the test statistic should be fully known when $p \in \mathcal{P}_0$, so that probability calculations involving the test statistic can be performed explicitly leading to a numerical answer.

The precise notion of the p-value is important, but somewhat difficult to convey. We will get to a formal definition only in Section 10.5.3, but loosely speaking, it can be thought of as the probability of the test statistic being "at least as extreme" as the observed statistic if the null hypothesis was true. This makes intuitive sense: an observed value of the test statistic that is "likely" to occur if the null hypothesis were true supports the null hypothesis, so conversely one that is "unlikely" contradicts it. However, only considering the probability of the observed statistic under the null is not enough, and we need to consider the probabilities of "more extreme" outcomes as well. To get a sense of why this is important, let us revisit Example 10.1.1 above.

Example 10.1.1. (Continued) Suppose we use the natural test statistic $S = \sum_{i=1}^{100} X_i$ for this example. Under the null hypothesis that the coin is fair ($\mathcal{P}_0 = \{\frac{1}{2}\}$), $S \sim \text{Binomial}(100, \frac{1}{2})$, so $P(S = 67)$ and $P(S \geq 67)$ are given by

```
dbinom(67, 100, 0.5)
```

```
[1] 0.0002324713
```

```
1 - pbinom(66, 100, 0.5)
```

```
[1] 0.0004368599
```

Both of these probabilities are small, and provide evidence against the null. But consider the situation where the observed sum was 50 rather 67. Clearly, this outcome is the strongest possible evidence in favour of the null hypothesis. Here, the corresponding probabilities $P(S = 50)$ and $P(S \geq 50)$ are

```
dbinom(50, 100, 0.5)
```

```
[1] 0.07958924
```

```
1 - pbinom(49, 100, 0.5)
```

```
[1] 0.5397946
```

Here $P(S = 50)$ is arguably quite small, and if we make a decision based on $P(S = 50)$ alone, we ignore the fact that while small, it is still the highest probability for any individual outcome. The contrast is more extreme if we consider a similar problem with 1000 instead of 100 tosses; in this case, $P(S = 500)$ and $P(S \geq 500)$ are

```
dbinom(500, 1000, 0.5)
```

```
[1] 0.02522502
```

```
1 - pbinom(499, 1000, 0.5)
```

```
[1] 0.5126125
```

In general, the probability of individual outcomes decreases as the size of the sample space "increases", and in fact for continuous distributions, the probability of any singleton outcome is 0. It is thus more natural to define the p-value as $P(S \geq s)$ where s is the observed quantity. ■

We will see later that the above notion of the p-value defined in terms of the probability of the test statistic being "at least as extreme as the observed test statistic" is generally useful regardless of the underlying model.

10.2 THE GOODNESS OF FIT PROBLEM IN THE MULTINOMIAL MODEL

A natural generalization of the previous example is to situations where there are more than one possible outcomes, and we want to "test", based on observed data, whether the probabilities associated with each outcome are as expected. A prototype of this problem is given by the following example.

Example 10.2.1. Suppose we roll a six-sided die n times, and record the outcome of the i-th roll as R_i. Assuming that the rolls are independent, the distribution of $R_1, R_2, \ldots, R_n$ is determined by the vector of probabilities of each outcome, which we denote by $p = (p_1, p_2, \ldots, p_6)$. Formally, the parameter space for the problem is

$$\mathcal{P} = \left\{ p = (p_1, p_2, \ldots, p_6) : 0 \leq p_i \leq 1 \text{ for all } i, \sum_{i=1}^{6} p_i = 1 \right\}.$$

We wish to test whether the die is fair, or in other words, $p_i = 1/6$ for all i. This happens if p belongs to the singleton subset $\mathcal{P}_0 = \{(1/6, 1/6, \ldots, 1/6)\}$ of $\mathcal{P}$. ■

Recall the multinomial distribution discussed in Example 3.2.12. If we define the vector $X = (X_1, X_2, \ldots, X_6)$ as the number of times each outcome occurs, that is,

$$X_j = \sum_{i=1}^{n} 1\{R_i = j\}, j = 1, 2, \ldots, 6$$

then we can identify the distribution of X as the multinomial distribution with size n and probability vector p. Another useful way to represent X is using indicator vectors as follows. For $j = 1, 2, \ldots, 6$, let e_j be the 6-dimensional unit vector with 1 in the j-th position and 0 otherwise. Define the i.i.d. random variables $Y_i = e_j$ if $R_i = j$. Then it is easy to verify that the distribution of Y_i is Multinomial with parameter 1 and probabilities $p = (p_1, p_2 \ldots p_6)$ and $X = \sum_{i=1}^{n} Y_i$ is Multinomial with parameter n and probabilities p. The distribution of individual Y_i's is refered to as the "categorical distribution" which generalizes the Bernoulli distribution to multiple categories.

In the multinomial framework, we have seen in Exercise 9.2.11 that the MLE of $\hat{p}_k = X_k/n$ with $X = \sum R_i$, then

There is no obvious test statistic, although intutively it seems that large values of $\sum (X_i - \frac{n}{6})^2$ would suggest that the equal probability hypothesis does not hold.

In its more general form, when the number of outcomes is some fixed number m (not necessarily 6) and $\mathcal{P}_0$ is a singleton set consisting of a fixed element of $\mathcal{P}$ (not necessarily with all components

equal), this problem is known as the "goodness of fit" problem. Despite the simplicity of the problem, we will see that it is difficult to find a simple solution.

10.3 INDEPENDENCE OF TWO CATEGORICAL ATTRIBUTES

Another simple but important problem that arises in the context of categorical data is to decide whether two categorical attributes are independent. Once suitably framed, this problem can also be formulated in terms of the multinomial distribution, but with a more complicated $\mathcal{P}_0$. As with the goodness of fit problem, there is no simple solution, yet the simplicity and wide applicability of the problem makes it important to study. We use the remainder of this section to formulate the problem precisely. We then move on to a discussion of the testing problem in general and some specfic problems that are simpler to analyze. We will come back to the goodness of fit and independence of categorical attributes problems towards the end of the chapter.

To motivate the problem we want to solve, consider the following example.

Example 10.3.1. Suppose a medical research team has come up with a potential vaccine for a dangerous disease, and we have been asked to design an experiment to determine whether it is effective. There are of course established protocols to design such studies (commonly known as 'clinical trials'), but the following strategy is reasonable as a first attempt. Choose n individuals from a vulnerable population, give the vaccine to n_1 of them (giving the remaining $n_2 = n - n_1$ individuals a placebo as "control"). Then, wait for a reasonable period of time (which may depend on the features of the disease) to see how many individuals are affected by the disease in each group. The result may be summarized in a 2×2 table as follows, where X_{11} denotes the number of vaccinated individuals who were affected, X_{12} denotes the number of vaccinated individuals who were not affected and so on.

	Affected	Not Affected	Total
Vaccine	X_{11}	X_{12}	n_1
Placebo	X_{21}	X_{22}	n_2

If the vaccine is effective, we expect a smaller proportion of the vaccinated group to be affected. If the chance of getting affected does not depend on whether the vaccine was given, then the vaccine is ineffective. The principle of scientific skepticism suggests that we should start by supposing independence (i.e., the vaccine has no effect), unless convinced otherwise by evidence. ■

It is easy to see that the setup in the previous example can apply to a wide range of problems of a similar nature. The essential aspects are that either of two "treatments" are applied to a group of experimental units, and one of two possible "outcomes" is then recorded for each unit. In the example above, treatments are *vaccine* and *placebo*, and outcomes are *affected* and *not affected*, but in general they can represent any binary categorical variable. The number of participants given treatment k who have outcome ℓ is denoted by $X_{k\ell}$ for $k, \ell = 1, 2$. In general, neither the number of possible treatments nor the number of possible outcomes need to be two, and both can be categorical attributes with an arbitrary number of categories.

In the example above, we have not explicitly introduced any randomness. Indeed, many aspects of the experiment could be random, such as the way the participants are selected from the population of interest, the number of participants, the number of participants given each treatment, and so on. We will now consider ways in which parametric probability models can describe such experiments.

To use our usual parametric setup, we need a sample of i.i.d. observations. The natural independent units in Example 10.3.1 are the participants in the study; in general, we may assume that individuals or units are selected randomly from a population of interest. Each such individual has two categorical attributes (coded by the numbers 1 and 2 for convenience), which we will generally refer to as *treatment* and *outcome*, even if they are not actually treatments or outcomes in the literal sense. There are four possible treatment-outcome combinations for each unit, namely $(1,1), (1,2), (2,1)$, and $(2,2)$. If we identify these combinations respectively with the four matrices

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \text{ and } \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

then it is easy to see that the summary table is in fact simply the sum of these matrices over all units in the experiment.

10.3.1 *A Multinomial Model for Two-way Tables*

It is now quite simple to formulate the problem parametrically. The information available for each individual unit (namely, the values of the "treatment" and "outcome" attributes) can be one of four possible 2×2 matrices, which can thus be thought of as the sample space of outcomes in a random experiment. If we assume that these outcomes, say M_i for the i-th individual, are distributed *independently* and *identically* for each individual unit, then we are back in our usual setup where we have n i.i.d. random observations (i.e., a random sample) $M_1, M_2, \ldots, M_n$ from some unknown distribution. The unknown distribution is discrete with four possible outcomes, so the most general parametric model is to assign four unknown probabilities to each outcome, with the only constraint that they must add up to 1. In other words, the parameters in the model are the probabilities p_{11} of seeing the outcome $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, p_{21} of seeing $\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$, p_{12} of seeing $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, and p_{22} of seeing $\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$, with the constraints that $0 \leq p_{11}, p_{21}, p_{12}, p_{22} \leq 1$ and $p_{11} + p_{21} + p_{12} + p_{22} = 1$. This model is precisely the categorical distribution we saw in the previous section, with the number of possible outcomes being four rather than six. As the n random outcomes are independent, the individuals outcomes M_i can be combined and summarized by their sum, giving the 2×2 table $X = \sum_{i=1}^{n} M_i = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}$. The distribution of the table X can thus be identified as a multinomial

distribution, with size n and probability vector $(p_{11}, p_{21}, p_{12}, p_{22})$. The parameter space is a subset of $\mathbb{R}^4$, specifically

$$\mathcal{P} = \left\{ p = (p_{11}, p_{21}, p_{12}, p_{22}) : 0 \leq p_{ij} \leq 1, \sum_{i,j} p_{ij} = 1 \right\}.$$

This multinomial model is appropriate when individuals or units are sampled from the population independently, with the total sample size n fixed in advance. This assumption is usually reasonable in observational studies where both attributes are intrinsic properties of the units being sampled, for example, in a study of college students that record each individual's gender and whether they need corrective lenses. For controlled trials such as the one described in Example 10.3.1, it is not immediately clear whether this model is appropriate because one of the attributes, namely the treatment, is assigned by the experimenter and not an intrinsic property of the individual subjects.

Consider this alternative probability model for such experiments: Suppose individuals are chosen independently at random from the population of interest. Each chosen individual is assigned treatment 1 with probability π_1 and treatment 2 with probability $\pi_2 = 1 - \pi_1$. Let q_{11} be the conditional probability of observing outcome 1 given treatment 1 and $q_{12} = 1 - q_{11}$. Similarly, suppose the conditional probability of outcome 1 given treatment 2 is q_{21} and the conditional probability of outcome 2 given treatment 2 is q_{22}. Then, the unconditional probability that individual i gets treatment 1 and has outcome 1 can be calculated as

$$\begin{aligned} P\left(M_i = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\right) &= P\left(\text{treatment} = 1, \text{outcome} = 1\right) \\ &= P\left(\text{treatment} = 1\right) P\left(\text{outcome} = 1 \mid \text{treatment} = 1\right) \\ &= \pi_1 q_{11} \end{aligned}$$

Similarly, we have

$$\begin{aligned} P\left(M_i = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}\right) &= \pi_2 q_{21}, \\ P\left(M_i = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\right) &= \pi_1 q_{12}, \text{ and} \\ P\left(M_i = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}\right) &= \pi_2 q_{22}. \end{aligned}$$

By construction, $M_1, M_2, \ldots, M_n$ are independent. Thus, the multinomial model is appropriate as a model for this setup as well if we identify the parameter vector $p = (p_{11}, p_{21}, p_{12}, p_{22})$ with $(\pi_1 q_{11}, \pi_2 q_{21}, \pi_1 q_{12}, \pi_2 q_{22})$. It is easy to see that in fact the two formulations are simple reparameterizations of each other.

Another alternative experimental setup, which may be more appropriate depending on how the experiment was conducted, is to fix the sample sizes corresponding to treatment 1 and treatment 2 in advance. We consider this model in Section 10.8.2.

10.3.2 *Independence in the Multinomial Model*

Our current interest is in testing whether the treatment and outcome attributes are independent. We will need to understand how this "hypothesis" translates to restrictions on $\mathcal{P}$ that define $\mathcal{P}_0$, the parameter space under the null hypothesis. This is given by the following lemma.

Lemma 10.3.2. *Let $\{p_{k\ell}, \pi_k$ and $q_{k\ell} : k, \ell = 1, 2\}$ be as in Section 10.3.1. Let $p_{k\circ} = p_{k1} + p_{k2}$ and $p_{\circ\ell} = p_{1\ell} + p_{2\ell}$ for $k, \ell = 1, 2$. For an individual, let T denote the treatment and Y denote the outcome. Then, the following are equivalent.*

(a) T and Y are independent

(b) $p_{k\ell} = p_{k\circ} p_{\circ\ell}$ for $k, \ell = 1, 2$.

(c) $q_{11} = q_{21}$.

Proof. The proof is left as an exercise. ∎

Thus the conjecture or hypothesis of independence that we are interested in testing can be stated as a constraint on the parametric model, specifically, that the true parameter p is in

$$\mathcal{P}_0 = \{p = (p_{11}, p_{21}, p_{12}, p_{22}) \in \mathcal{P} : p_{k\ell} = p_{k\circ} p_{\circ\ell} \text{ for } k, \ell = 1, 2\}.$$

It is easy to verify that the maximum likelihood estimator of p is $\hat{p} = \frac{1}{n} X = \frac{1}{n} \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}$ (see Exercise 9.2.11) It can be similarly verified that the likelihood as a function of $p \in \mathcal{P}_0$ is maximized at $\hat{p}_{k\circ} = \frac{1}{n}(X_{k1} + X_{k2})$ and $\hat{p}_{\circ\ell} = \frac{1}{n}(X_{1\ell} + X_{2\ell})$, or equivalently, $\hat{q}_{11} = \hat{q}_{21} = \frac{1}{n}(X_{11} + X_{21})$.

As we will see later, test statistics are usually a function of maximum likelihood estimators. Unfortunately, unlike in Example 10.1.1, it is not obvious how we can combine these estimators into a summary statistic that can be used for testing of independence. The difference in this case is that even under the null hypothesis, when $p \in \mathcal{P}_0$, the actual value of p is unknown, so the distribution of the MLEs obtained above cannot be fully determined.

There is in fact a widely used 'solution' for this problem: Pearson's χ^2 test of independence. The test statistic used in this test is not difficult to motivate intuitively. However, its exact distribution is unknown and can be only approximately calculated. We will return to this test towards the end of this chapter, after discussing some simpler parametric problems and introducing some general terminology and results.

10.4 TESTING IN THE PARAMETRIC SETUP : THE INTUITIVE APPROACH

So far, in this chapter, we have discussed testing problems with one common theme: the primary attribute we were interested in was a categorical outcome, and could be modeled using a categorical distribution (or the Bernoulli distribution in the simplest case). Although we have not yet been able to derive a satisfactory general approach to testing in this problem, the setup is useful in motivating

the next set of examples, where instead of a categorical outcome we consider a continuous outcome, which we model using a Normal distribution.

In this section, we consider the continuous analog of Example 10.1.1, with one set of univariate i.i.d. observations $X_1, X_2, \ldots, X_n$ from the Normal distribution.

10.4.1 *Finding a Test Statistic*

As we will soon see in Definition 10.4.1, any statistic, or function of the available data, can be a test statistic in principle. To be of practical use, the distribution of a test statistic should be completely known under the null hypothesis. In the Bernoulli coin toss example above, $\mathcal{P}$ is the interval $[0, 1]$, $\mathcal{P}_0$ is the singleton set $\{0.5\}$ and the test statistic is the total number of heads in say n tosses is $S = \sum_{i=1}^{n} X_i$. Although the distribution of S (Binomial) generally involves the unknown p, it is completely known for $p = 0.5$.

Ideally one wants to find the test statistic that is also "best" in some sense, but often the optimality of a test is difficult to establish. In Section 10.5, we will describe *one* principled approach that often, but not always, leads to a test statistic with some optimality properties. Sometimes, however, a suitable test statistic is easy to derive intuitively, as in the coin toss example above. In this section, we consider some specific examples where such tests are available.

Intuitively, a good test statistic should satisfy two main criteria. Firstly, the sampling distribution of the test statistic should be known when $p \in \mathcal{P}_0$. Secondly, it should have the "power" to distinguish whether the conjecture is true or false (i.e., its distribution should vary substantially when $p \in \mathcal{P}_0$ and when it is not). We have seen one important example, namely the Bernoulli model, where an intuitively appealing test statistic was easy to find, but we have also seen closely related multinomial models where such test statistics were not available. We will now investigate a few examples involving the Normal distribution, which has two parameters, to see whether we can come up with similarly simple and intuitive test statistics. Our goal is to get a sense of what test procedures typically look like, before discussing a more systematic but somewhat abstract approach in the next section. Before proceeding, let us formally define the notion of a test statistic in this context.

Definition 10.4.1. *Given observed data $X_1, X_2, \ldots, X_n$, a test statistic is any real-valued function $T(X_1, X_2, \ldots, X_n)$ of $X_1, X_2, \ldots, X_n$.*

10.4.2 *The Normal Distribution: Test for Sample Mean when σ is Known*

Suppose $X \sim Normal(\mu, \sigma^2)$ where μ is an unknown mean, but σ is a *known* standard deviation. Thus, here the parameter of interest $p \equiv \mu$ and parameter space $\mathcal{P} = \mathbb{R}$.

Two-sided test

Suppose we want to test the null hypothesis that $\mu = a$, where a is some known value that has special meaning in the context of the problem. Thus, here $\mathcal{P}_0 = \{a\}$.

EXAMPLE 10.4.2. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the distribution $X \sim \text{Normal}(\mu, 1)$, where $\mu \in \mathcal{P} = \mathbb{R}$. We are interested in testing the null hypothesis that $\mu = 0$, or equivalently, that $\mu \in \mathcal{P}_0 = \{0\}$. Clearly, an intuitively appealing test statistic is the observed sample mean $\overline{X}$. The further $\overline{X}$ is from 0, the less confident we would be that the null hypothesis is true. The question is how far away should it be before we decide that the evidence against the null hypothesis is strong enough.

Note that although we have assumed $\sigma = 1$ and $a = 0$, this does not lead to any loss in generality. Suppose the true distribution of X was $\text{Normal}(\mu, \sigma^2)$, and we were interested in testing the null hypothesis that $\mu = a$. Consider the transformed observations $Z_i = (X_i - a)/\sigma$. Then their common distribution is $Z \sim \text{Normal}((\mu - a)/\sigma, 1)$, and the null hypothesis $\mu = a \iff (\mu - a)/\sigma = 0$. ■

The sample mean $\overline{X}$ can thus be viewed as the test statistic $T(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$. In general, a value of $\overline{X}$ close to a tends to support the conjecture that $\mu = a$, and a value away from a tends to make us suspect it. But the strength of the evidence against the conjecture depends not only on the difference $\overline{X} - a$, but also on σ. It is easy to see from results we have already encountered that *if* the null hypothesis is true, that is, $\mu = a$, then $\overline{X} \sim \text{Normal}(a, \sigma^2/n)$, and so

$$T_a(X_1, X_2, \ldots, X_n) = \sqrt{n}\left(\frac{\overline{X} - a}{\sigma}\right) \sim \text{Normal}(0, 1).$$

The *if* in the last sentence is important, and bears emphasizing. Another way to express the same statement is to say that if $Y_1, Y_2, \ldots, Y_n$ are i.i.d. observations from the $\text{Normal}(a, \sigma^2)$ distribution, then

$$T_a(Y_1, Y_2, \ldots, Y_n) = \sqrt{n}\left(\frac{\overline{Y} - a}{\sigma}\right) \sim \text{Normal}(0, 1).$$

This distinguishes between the *observed* data $X_1, X_2, \ldots, X_n$ whose assumed distribution depends on the unknown μ, and the *hypothetical* random variables $Y_1, Y_2, \ldots, Y_n$ whose distribution is assumed to satisfy the null hypothesis.

The known distribution of the test statistic, calculated from the distribution of the hypothetical data for which the null hypothesis is true, is known as the "null distribution". It then remains to compare the observed value of T_a to this known distribution to obtain a p-value. The formal definition of p-value will be given later, but as noted earlier, we can think of it as the probability that under the null distribution we will see a value of $T_a(Y_1, Y_2, \ldots, Y_n)$ at least as extreme as the value of $T_a(X_1, X_2, \ldots, X_n)$ observed from the data at hand. Here, "more extreme" is to be interpreted as values that are even less likely, and thus would have provided even *more* evidence against the null hypothesis. In this case, large positive and large negative values of T_a are both evidence against the null hypothesis, and the symmetry of the distribution of $T_a(Y_1, Y_2, \ldots, Y_n)$ suggests that the p-value may be computed as

$$2P\big(T_a(Y_1, Y_2, \ldots, Y_n) \geq |T_a(X_1, X_2, \ldots, X_n)|\big) \tag{10.4.1}$$

given the observed value of the test statistic $T_a(X_1, X_2, \ldots, X_n)$. We would reject the null hypothesis if the p-value obtained in (10.4.1) is small enough (say smaller than 0.05).

One-sided test

The test above conjectures that $\mu = a$, and considers deviations on either side to be departures that invalidate the null hypothesis. A common variation of the above test is interested in deviations from the null hypothesis in only one direction. Suppose, as before, that the distribution of X is Normal(μ, σ^2) with known σ^2 and $\mu \in \mathcal{P} = \mathbb{R}$. Further, $X_1, X_2, \ldots, X_n$ is an i.i.d. sample from X, and we are interested in testing the null hypothesis that $\mu \leq a$, or equivalently, that $\mu \in \mathcal{P}_0 = (-\infty, a]$. Intuitively, the data support the null hypothesis if $\overline{X} \leq a$, and oppose it more and more strongly the larger $\overline{X}$ is than a. Here, we would expect the null hypothesis to be rejected if $\overline{X} - a$ is large. It would be natural to compute

$$P\left(T_a(Y_1, Y_2, \ldots, Y_n) \geq T_a(X_1, X_2, \ldots, X_n)\right),$$

where $Y_1, Y_2, \ldots, Y_n \sim$ i.i.d. Normal(μ, σ^2) and reject if this probability is uniformly small for all $\mu \leq a$. We thus define the p-value to be

$$\max\left\{P\left(T_a(Y_1, Y_2, \ldots, Y_n) \geq T_a(X_1, X_2, \ldots, X_n)\right) : Y_1, Y_2, \ldots, Y_n \text{ i.i.d. Normal}(\mu\,\sigma^2),\ \mu \leq a\right\}.$$

One can in fact show that this maximum is achieved for $\mu = a$. Let us look at a specific numerical example to make these ideas concrete.

Example 10.4.3. Students in a probability class were surveyed to obtain the sex and heights of one of their parents, along with those of that parent's oldest sibling, provided they had one. When the siblings compared were both male, the difference in heights (oldest sibling - parent) rounded to the nearest centimeter were as follows: -5 -5 -4 -2 -2 1 2 2 3 4 4 5 7.

Assume that these values are i.i.d. observations from a Normal distribution with unknown mean μ and known variance $\sigma^2 = 3^2$. We want to test the null hypothesis that $\mu = 0$.

The observed value of $\overline{X}$ is 0.769, and thus $T_0(X_1, X_2, \ldots, X_n) = \sqrt{13}(0.769/3) = 0.925$. The probability that the random variable $T_0(Y_1, Y_2, \ldots, Y_n)$, which follows Normal$(0, 1)$, is larger than 0.925 can be calculated in R as follows.

```
pnorm(0.925, mean = 0, sd = 1, lower.tail = FALSE)
```

```
[1] 0.177483
```

This is the p-value for the one-sided null hypothesis $\mu \leq 0$. For the two-sided null hypothesis the p-value would then be $2 \times 0.1775 = 0.354$. As both p-values are larger than 0.05 we do not reject either of the null hypotheses. ■

We will discuss the interpretation of the p-value in more detail in Section 10.5, but the value 0.05 is generally considered a reasonable cutoff between "weak" and "strong" evidence against the null hypothesis.

10.4.3 *The Normal Distribution: Test for Sample Mean when σ is Unknown*

Despite all the elaborate notation, at its core the previous example is not very different from the Binomial example we started with. In both cases, there is only one unknown parameter of interest, and the null hypothesis fixes it to a particular known value, making the distribution of the data essentially known if the null hypothesis holds. Obtaining a test statistic whose distribution would be known under the null hypothesis is then almost trivial, as literally any function of the data would satisfy this requirement.

This is rarely the case in realistic hypothesis testing scenarios. Suppose again that $X \sim \text{Normal}\left(\mu, \sigma^2\right)$, but now both the mean μ and the standard deviation σ are unknown. Thus, here parameter of interest is $p \equiv (\mu, \sigma)$ and parameter space $\mathcal{P} = \mathbb{R} \times (0, \infty)$. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the distribution X. As before, we want to test the null hypothesis that $\mu = a$, where a is some known value that has special meaning in the context of the problem. However, the null hypothesis makes no specific conjecture about the possible values of σ^2. As $\mathcal{P}_0 = \{a\} \times (0, \infty)$ is no longer a singleton set, the distribution of X is not completely specified when the null hypothesis holds. As the null hypothesis makes a conjecture only about μ and none about σ, the latter is often referred to as a "nuisance parameter".

The essential approach remains the same as before. We wish to find a test statistic whose distribution does not depend on the unknown parameters when the null hypothesis is true. The test statistic in the earlier case was

$$T_a(X_1, X_2, \ldots, X_n) = \sqrt{n}\left(\frac{\overline{X} - a}{\sigma}\right) \sim \text{Normal}\,(0, 1)\,.$$

However, this is not a valid test statistic in this case because it cannot even be calculated, as it involves σ, which is unknown. Intuitively, we may expect to fix this problem by replacing σ with an estimate, say the sample standard deviation S (see Definition 7.1.5 and Exercise 7.1.8). This leads to a proper statistic

$$T_a(X_1, X_2, \ldots, X_n) = \sqrt{n}\left(\frac{\overline{X} - a}{S}\right).$$

For this modified statistic $T_a(X_1, X_2, \ldots, X_n)$ to be useful as a test statistic, its distribution under the null hypothesis should be completely known. In general, the distribution of T_a would depend on μ and σ^2. If the null hypothesis $\mu = a$ is true, then μ is known, but σ^2 is still unknown. However,

it is easy to see that the distribution of $T_a(Y_1, Y_2, \ldots, Y_n)$ does *not* depend on σ^2 in this case. To see this, let $Y_1, Y_2, \ldots, Y_n$ be an i.i.d. sample from Normal(a, σ^2) for some arbitrary $\sigma^2 > 0$. Note,

$$\begin{aligned} T_a(Y_1, Y_2, \ldots, Y_n) &= \sqrt{n}\left(\frac{\overline{Y} - a}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})^2}}\right) \\ &= \sqrt{n}\left(\frac{\frac{\overline{Y}-a}{\sigma}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n \left(\frac{Y_i - \overline{Y}}{\sigma}\right)^2}}\right) \end{aligned}$$

Define $Z_i = (Y_i - a)/\sigma$. So from the above we have

$$T_a(Y_1, Y_2, \ldots, Y_n) = \sqrt{n}\left(\frac{(\overline{Z} - 0)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (Z_i - \overline{Z})^2}}\right) = T_0(Z_1, Z_2, \ldots, Z_n).$$

Clearly, $Z_i \sim$ Normal$(0, 1)$ has distribution free of σ^2. Thus the distribution of $T_a(Y_1, Y_2, \ldots, Y_n)$ which is equal to $T_0(Z_1, Z_2, \ldots, Z_n)$ is also free of σ^2. So we can calculate p-values involving it, at least in principle.

Continuing Example 10.4.3, we can obtain the value of $T_a(X_1, X_2, \ldots, X_n)$, with $a = 0$, using the following R code, by first estimating $\overline{X}$ and S^2 from the data itself.

```
a <- 0
x <- c(-5, -5, -4, -2, -2, 1, 2, 2, 3, 4, 4, 5, 7)
n <- length(x)
mu_x <- mean(x)
sigma_x <- sqrt(sum((x - mu_x)^2) / (n-1))
T_x <- sqrt(n) * ((mu_x - a) / sigma_x)
T_x
```

```
[1] 0.6964513
```

The value of $T_a(X_1, X_2, \ldots, X_n)$ is thus 0.6965. To obtain the corresponding p-value for the two-sided null hypothesis $\mu = 0$, we must compute

$$2P\left(T_a(Y_1, Y_2, \ldots, Y_n) \geq |\ 0.6965\ |\ \right)$$

when $Y_1, Y_2, \ldots, Y_n$ are i.i.d. Normal $\left(a, \sigma^2\right)$.

Suppose we did not know the distribution of $T_a(Y_1, Y_2, \ldots, Y_n)$, we could approximately compute this probability by simulating values of $Y_1, Y_2, \ldots, Y_n$. The critical observation here is that it does not matter that we do not know σ^2, because the distribution of $T_a(Y_1, Y_2, \ldots, Y_n)$ does not depend on σ^2. Thus, for simulation, any value of σ^2 is as good as any other. Here we choose to simulate

with $\sigma^2 = 1$. Once $Y_1, Y_2, \ldots, Y_n$ are available, we simply repeat the same calculations as above to calculate $T_a(Y_1, Y_2, \ldots, Y_n)$.

```
n <- length(x)
T_sim <-
    replicate(1000000,
    {
        y <- rnorm(n, mean = 0, sd = 1)
        mu_y <- mean(y)
        sigma_y <- sqrt(sum((y - mu_y)^2) / (n-1))
        T_y <- sqrt(n) * (mu_y / sigma_y)
        T_y
    })
uprob <- mean(T_sim >= abs(T_x))
uprob
```

```
[1] 0.249762
```

As before, the p-value is twice the upper tail probability, which in turn is estimated by `uprob`, the proportion of cases out of a million simulations where the simulated $T_0(Y_1, Y_2, \ldots, Y_n)$ exceeds 0.6965. The estimated p-value is thus $2 \times 0.2498 = 0.4995$. Of course, this estimated p-value will vary every time the simulation is run, but it should be reasonably close to the correct answer. In fact, as this estimate is based on estimating the proportion p from a series of Bernoulli trials, we can use what we learned in Chapter 9 to obtain a confidence interval for the true p-value. Using the `bernoulliQuadraticCI()` and `bernoulliLinearCI()` functions defined in Section 9.3.3, we have the following 95% confidence intervals for the p-value, which are essentially identical, which is not surprising given the large number of replications.

```
2 * bernoulliQuadraticCI(T_sim >= abs(T_x), level = 0.95)
```

```
[1] 0.4978291 0.5012228
```

```
2 * bernoulliLinearCI(T_sim >= abs(T_x), level = 0.95)
```

```
[1] 0.4978272 0.5012208
```

In this somewhat special situation, however, this simulation approach is not necessary because the distribution of $T_a(Y_1, Y_2, \ldots, Y_n)$ is well-studied theoretically, with its density function available in closed form and numerical algorithms available to evaluate its cumulative distribution func-

tion. Recall that we learned about the t distribution in Chapter 8, and note in particular that Corollary 8.1.11 immediately implies that in this case,

$$T_a(Y_1, Y_2, \ldots, Y_n) = \sqrt{n}\,\frac{(\overline{Y} - a)}{S} \sim t_{n-1}.$$

Tail probabilities of the t distribution can be computed in R using the `pt()` function, which is analogous to the `pnorm()` function for the Normal distribution. Thus, the exact p-value in this case can be computed as follows.

```
2 * pt(T_x, df = n - 1, lower.tail = FALSE)
```

```
[1] 0.4994148
```

As the p-values we have computed for this example are all larger than 0.05, we would not have rejected the null hypothesis that $\mu = 0$, irrespective of the approach taken.

10.4.4 *An Alternative Test Based on the Median*

As long as we are considering tests that are intuitively appealing, the test statistics outlined above are not the only possibilities. For example, the parameter μ in the Normal (μ, σ^2) distribution is the median as well as the mean, so it may be reasonable to define a test statistic based on the sample median instead of the sample mean. Specifically, to test the null hypothesis that $\mu = a$, consider the statistic

$$\widetilde{T}_a(X_1, X_2, \ldots, X_n) = \sqrt{n}\left(\frac{\text{median}(X) - a}{\widetilde{S}}\right).$$

where

$$\widetilde{S} = \frac{1}{n}\sum_{i=1}^{n} |X_i - \text{median}(X)|\,.$$

Essentially, $\widetilde{T}_a(X_1, X_2, \ldots, X_n)$ replaces the sample mean by the sample median in the formula for $T_a(X_1, X_2, \ldots, X_n)$, with an analogous change in the estimate of scale. We can argue as before that the distribution of $\widetilde{T}_a(Y_1, Y_2, \ldots, Y_n)$, when $Y_1, Y_2, \ldots, Y_n$ are independent Normal (a, σ^2), does not depend on σ^2, although it may still depend on n. Unfortunately, no useful representation is available that allows us to compute its exact tail probabilities. However, we can still use the simulation approach as before.

Continuing Example 10.4.3, we can obtain the value of $\widetilde{T}_a(X_1, X_2, \ldots, X_n)$, with $a = 0$, as follows.

```
a <- 0
x <- c(-5, -5, -4, -2, -2, 1, 2, 2, 3, 4, 4, 5, 7)
n <- length(x)
mu_x <- median(x)
sigma_x <- sum(abs(x - mu_x)) / n
T_x <- sqrt(n) * ((mu_x - a) / sigma_x)
T_x
```

```
[1] 2.232008
```

The observed value of $\widetilde{T}_a(X_1, X_2, \ldots, X_n)$ here is 2.232. To compute the corresponding p-value, we again simulate values of $\widetilde{T}_0(Y_1, Y_2, \ldots, Y_n)$ where $Y_1, Y_2, \ldots, Y_n$ are i.i.d. Normal $(0, 1)$.

```
T_sim_median <-
    replicate(1000000,
    {
        y <- rnorm(n, mean = 0, sd = 1)
        mu_y <- median(y)
        sigma_y <- sum(abs(y - mu_y)) / n
        T_y <- sqrt(n) * (mu_y / sigma_y)
        T_y
    })
uprob <- sum(T_sim_median >= abs(T_x)) / 1000000
uprob
```

```
[1] 0.094342
```

The desired approximation to the p-value in this case is again twice the upper tail probability, i.e. $2 \times 0.0943 = 0.1887$. As this value is larger than 0.05, we will not reject the null hypothesis that $\mu = 0$.

One may reasonably wonder which of the above tests is "better". This is an important question in general, but it is beyond the scope of this book. In this specific case, the first test can be shown to have more "power" to identify situations where the null hypothesis does not hold: This means that if the true value of μ differs from a, then the test based on the mean has higher probability of rejecting the null hypothesis (by producing a p-value less than 0.05) than the test based on the median. This holds for all values of $\mu \neq a$, although the actual probabilities of rejection would certainly depend on the value of μ. However, this assurance requires Normality of the underlying measurement. As in Chapter 9, simulation studies can usually provide helpful guidance regarding the performance of specific tests under various scenarios.

EXERCISES

10.5 THE GENERAL APPROACH: LIKELIHOOD RATIO TEST

The examples in the previous section illustrate the "intuitive" approach for developing a test statistic given a hypothesis of interest. While this approach is useful in many situations, one is often interested in a principle that may be applied in a general setup, much as the maximum likelihood principle served for estimation in Chapter 9. We will describe a similar principle for testing in this section. Although we will not delve into theoretical properties of this approach, we mention two important results about it: First, the resulting test can be easily shown to be optimal in the special case where both $\mathcal{P}_0$ and $\mathcal{P} \setminus \mathcal{P}_0$ are singleton sets (this is a fundamental result in statistics that is known as the Neyman-Pearson Lemma (see [CasBer90]) as well as more generally for certain families of distributions. Second, even though the distribution of the resulting test statistic under the null hypothesis may not always be computable, it can be determined *asymptotically* under fairly general conditions (see Wilks' Theorem, Theorem 10.7.2).

10.5.1 *The Likelihood Ratio Statistic*

Recall from Chapter 9 that the likelihood function given the sample $(X_1, X_2, \ldots, X_n)$ is defined as

$$L(p; X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} f(X_i \mid p).$$

and the maximum likelihood estimator (MLE) is given by

$$\hat{p} \equiv \hat{p}(X_1, X_2, \ldots, X_n) = \arg\max_{p \in \mathcal{P}} L(p; X_1, \ldots, X_n),$$

where we use the notation "arg max" to denote the value of the argument p for which the maximum is obtained. We can similarly define the MLE restricted to the null hypothesis being tested as

$$\hat{p}_0 \equiv \hat{p}_0(X_1, X_2, \ldots, X_n) = \arg\max_{p \in \mathcal{P}_0} L(p; X_1, \ldots, X_n).$$

We now define the *likelihood ratio* as

$$\lambda(X_1, \ldots, X_n) = \frac{L(\hat{p}_0; X_1, \ldots, X_n)}{L(\hat{p}; X_1, \ldots, X_n)}, \tag{10.5.1}$$

and the *likelihood ratio statistic*[1] as

$$\Lambda(X_1, \ldots, X_n) = -2 \log \lambda(X_1, \ldots, X_n) = 2 \log \frac{L(\hat{p}; X_1, \ldots, X_n)}{L(\hat{p}_0; X_1, \ldots, X_n)}. \tag{10.5.2}$$

[1] which should perhaps be called the log-likelihood ratio statistic, and often is.

The intuition behind these definitions is as follows. By definition of $\hat{p}$ and $\hat{p}_0$, we must have $L(\hat{p}; X_1, \ldots, X_n) \geq L(\hat{p}_0; X_1, \ldots, X_n) \geq 0$, hence $0 \leq \lambda(X_1, \ldots, X_n) \leq 1$ and thus $\Lambda(X_1, \ldots, X_n) \geq 0$. Equality is achieved if the unrestricted MLE $\hat{p} \in \mathcal{P}_0$. If $\hat{p} \notin \mathcal{P}_0$ then $\Lambda(X_1, \ldots, X_n) > 0$ gives a measure of how far $\hat{p}$ is from $\mathcal{P}_0$ (in terms of L). The further $\hat{p}$ is away from $\mathcal{P}_0$, the less likely it is that the null hypothesis $p \in \mathcal{P}_0$ is true. The general principle then is to believe the null hypothesis if the likelihood ratio λ is close to one, or equivalently, the likelihood ratio statistic Λ is small (close to zero), and suspect it when Λ is large. The reason for taking log and the including the factor of 2 is that doing so makes the distribution of Λ more convenient, as will become clear in due course.

10.5.2 *Type I and Type II Error*

It still remains for us to determine how large $\Lambda(X_1, \ldots, X_n)$ should be before we conclude that the balance of evidence suggests that the null hypothesis is false. We next present a simple example where the underlying distribution depends only on one parameter, but is nonetheless helpful in understanding this question.

Example 10.5.1. Suppose $X_1, X_2, \ldots, X_n$ is an i.i.d. sample from Bernoulli(p), where $p \in \mathcal{P} = (0, 1)$. We are interested in testing the null hypothesis that $p = p_0$, or equivalently, that $p \in \mathcal{P}_0 = \{p_0\}$. Intuitively, the further the sample proportion $\hat{p}$ is from p_0, the less confident we would be that the null hypothesis is true. It is easy to see that in this case $\Lambda(X_1, X_2, \ldots, X_n)$ is given by

$$\Lambda(X_1, X_2, \ldots, X_n) = 2T \log \frac{T/n}{p_0} + 2(n - T) \log \frac{1 - T/n}{1 - p_0},$$

where $T = \sum_{i=1}^{n} X_i$ (see (10.6.1) below). Therefore Λ is minimised (i.e. $\Lambda = 0$) when $T = np_0$ which matches our intuition. ■

Consider a specific instance in Example 10.5.1 with $n = 10$ and $p_0 = 0.5$. Then $T \sim$ Binomial$(10, p)$ distribution. If $T = 5$, we would obviously have no reason to suspect the null hypothesis, even though the probability of observing $T = 5$ when $p = 0.5$ is only around 0.25. Even if $T = 4$ or $T = 6$, we would not suspect the null hypothesis, as these are still quite plausible outcomes. However, if $T = 0$ or $T = 10$, we would possibly suspect that the null hypothesis is false, because $P(T = 0) = P(T = 10) = 0.0098$ is quite low. But where exactly do we draw the line?

To reasonably answer this question, we must distinguish between two types of mistakes we risk making. Suppose we decide that we will accept (i.e., not reject) the null hypothesis if $c_1 < T < c_2$ and reject it otherwise, where c_1 and c_2 need to be determined. Now imagine that the null hypothesis ($p = 0.5$) is indeed true. Suppose we obtain a random sample $X_1, X_2, \ldots, X_{10}$ and perform the test. Rejecting the null hypothesis would then be a mistake because we would conclude that the null hypothesis is false even though it is actually true. By convention, such a mistake is called a "Type I error" or "false positive". The probability of making a Type I error in this case is

$$P(T \leq c_1) + P(T \geq c_2) = 1 - \sum_{k=c_1+1}^{c_2-1} \binom{10}{k} \left(\frac{1}{2}\right)^{10}$$

For example, if $c_1 = 2$ and $c_2 = 8$, the probability of Type I error is 0.109. If instead $c_1 = 1$ and $c_2 = 9$, the probability of Type I error is 0.021.

On the other hand, if the null hypothesis is not true, we may still mistakenly accept the null hypothesis. This is called a "Type II error" or a "false negative". The probability of making a Type II error depends on the true value of the parameter p and is given by

$$P(T \leq c_1) + P(T \geq c_2) = 1 - \sum_{k=c_1+1}^{c_2-1} \binom{10}{k} p^k (1-p)^{10-k}.$$

For example, if $p = 0.75$, then with $c_1 = 2$ and $c_2 = 8$, probability of Type II error is 0.474, and with $c_1 = 1$ and $c_2 = 9$, probability of Type II error is 0.756. Similarly, for $p = 0.9$, with $c_1 = 2$ and $c_2 = 8$, probability of Type II error is 0.07, and with $c_1 = 1$ and $c_2 = 9$, probability of Type II error is 0.264.

These calculations illustrate a general principle, that trying to decrease the probability of Type I error (e.g., by controlling c_1 and c_2 in this example) generally results in an increase in the probability of Type II error. The standard approach to resolve this trade-off is a two-step approach: (a) Obtain a reasonable "test statistic", such as the likelihood ratio statistic $\Lambda(X_1, \ldots, X_n)$ defined above. (b) Choose a threshold for the test statistic,[2] beyond which the null hyothesis is declared to be rejected, in a manner that ensures that the probability of Type I error does not exceed a pre-determined limit.

An "optimal" choice of test statistic in (a) would ensure that any other test with equal or lower probability of Type I error will always have higher probability of Type II error. Using $\Lambda(X_1, \ldots, X_n)$ as the test statistic ensures such optimality in many situations, provided that it is possible to compute and control the probability of Type I error for the resulting test. Such a test is known as the *likelihood ratio test*.

With this background, we will now proceed to outline a general test procedure, before getting back to the specific examples cited above.

10.5.3 *The p-value for the Likelihood Ratio Test*

Let X be a random variable with probability mass function or probability density function $f(x)$, where $f(x) \equiv f(x \mid p)$ depends on one or more unknown parameters $p \in \mathcal{P}$. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. random sample with common distribution X. In our approach, the test statistic is the likelihood ratio statistic $\Lambda(X_1, \ldots, X_n)$, and we reject the null hypothesis if Λ is large. However, rather than trying to find a rejection cutoff that controls the probability of Type I error, we will take a more modern approach and introduce the closely related concept of p-value.

Note that $\Lambda(X_1, \ldots, X_n)$ is also a random variable having a corresponding sampling distribution of its own. When performing a test, we will work with a specific realisation of this sample. Henceforth, we will denote this realised value of $\Lambda(X_1, \ldots, X_n)$ by d, to emphasize that it is a constant for the purposes of the test.

[2] More generally, a set of possible values of the test statistic for which the null hypothesis is rejected.

Now, the sample $X_1, X_2, \ldots, X_n$ was generated with a particular value of p that may or may not have belonged to $\mathcal{P}_0$. The p-value is concerned with the probabilistic behaviour of the random variable $\Lambda(X_1, \ldots, X_n)$ when the underlying parameter *does* belong to $\mathcal{P}_0$. To distinguish a sample from such a "thought experiment" from the actual realized sample, imagine a second set of i.i.d. observations $Y_1, Y_2, \ldots, Y_n$ from the distribution of X for some $p \in \mathcal{P}$.

Definition 10.5.2. *The p-value for testing $p \in \mathcal{P}_0$ based on Λ is defined as*

$$\max_{p \in \mathcal{P}_0} P_p(\Lambda(Y_1, \ldots, Y_n) \geq d). \tag{10.5.3}$$

The notation P_p emphasizes that the probability calculations are done with parameter value p, and $d = \Lambda(X_1, X_2, \ldots, X_n)$ is the likelihood ratio statistic calculated from the observed sample.

Whether we can actually compute the p-value still remains to be seen, and will depend on the problem. Assuming that it can be, consider the following test procedure.

Definition 10.5.3. *(Level α test) Fix $0 < \alpha < 1$. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a population with distribution X. To test the null hypothesis $p \in \mathcal{P}_0$ at level α, compute the p-value as above, and reject the null hypothesis if the p-value is less than or equal to α. Otherwise, accept the null hypothesis.*

The following result establishes that the probability of Type I error for this test procedure does not exceed α. This property is conventionally taken to be defining characteristic of a level α test.

Theorem 10.5.4. *For a level α test obtained from the likelihood ratio statistic $\Lambda(X_1, \ldots, X_n)$ with p-value computed according to (10.5.3), the probability of Type I error does not exceed α.*

Proof. As we are interested in the probability of Type I error, suppose that $X_1, X_2, \ldots, X_n$ is a random sample from X with some parameter value $p_0 \in \mathcal{P}_0$. We need to show that the probability of rejection does not exceed α. Let $\Lambda(X_1, \ldots, X_n)$ be the likelihood ratio statistic. Consider another random sample $Y_1, Y_2, \ldots, Y_n$ from X with the same parameter p_0. If the test is rejected, the p-value must be α or less, and in particular, we must have (holding $\Lambda(X_1, \ldots, X_n)$ fixed)

$$P_Y(\Lambda(Y_1, \ldots, Y_n) \geq \Lambda(X_1, \ldots, X_n)) \leq \alpha$$

as this inequality must hold for all $p \in \mathcal{P}_0$ and hence in particular for p_0. Therefore, we must have

$$P_X(\text{null hypothesis rejected}) \leq P_X\left(P_Y(\Lambda(Y_1, \ldots, Y_n) \geq \Lambda(X_1, \ldots, X_n)) < \alpha\right) \leq \alpha.$$

This last result follows from Lemma10.5.5 below, with $Z = \Lambda(X_1, \ldots, X_n)$ and $W = \Lambda(Y_1, \ldots, Y_n)$. ∎

Lemma 10.5.5. *Let Z and W be i.i.d. random variables with distribution function F. Then*

$$P_Z\left(P_W(W \geq Z) < \alpha\right) \leq \alpha.$$

Proof. The proof is simple if we assume F is continuous and strictly increasing on its support.

$$\begin{aligned} P_Z\left(P_W(W \geq Z) < \alpha\right) &= P_Z(1 - F(Z) < \alpha) \\ &= P_Z(F(Z) > 1 - \alpha) \\ &= P_Z(Z > F^{-1}(1-\alpha)) \\ &= 1 - F(F^{-1}(1-\alpha)) \\ &= 1 - (1-\alpha) = \alpha \end{aligned}$$

The result also holds for more general F, but that case requires more careful manipulation, with the first two equalities in the proof above becoming inequalities. ∎

10.6 SPECIFIC EXAMPLES

10.6.1 *Binomial Test for Proportion*

Let us now return to Example 10.5.1, where $X_1, X_2, \ldots, X_n$ is an i.i.d. sample from Bernoulli(p), where $p \in \mathcal{P} = (0,1)$. Then $T = \sum_{i=1}^n X_i$ has a Binomial(n,p) distribution. We are interested in testing the null hypothesis that $p = p_0$.

The likelihood function for this example is

$$L(p; X_1, X_2, \ldots, X_n) = \binom{n}{T} p^T (1-p)^{n-T}, p \in \mathcal{P},$$

It is easy to see that $\hat{p} = T/n$ and $\hat{p}_0 = p_0$. Thus the likelihood ratio statistic is

$$\begin{aligned} \Lambda(X) &= 2\log\frac{L(\hat{p}; X)}{L(\hat{p}_0; X)} \\ &= 2X\log\frac{X/n}{p_0} + 2(n-X)\log\frac{1-X/n}{1-p_0}, \end{aligned} \tag{10.6.1}$$

where we interpret $0\log(0)$ as 0. To compute the p-value for this test for a specific realization of X, we need to first compute $d = \Lambda(X)$, and then compute

$$P_{p_0}(\Lambda(Y) \geq d),$$

where Y has the Binomial(n, p_0) distribution. Although we cannot express this probability in closed form, we can explicitly write it as the following sum involving Binomial(n, p_0) probabilities:

$$P(\Lambda(Y) \geq d) = \sum_{0 \leq k \leq n : \Lambda(k) \geq d} \binom{n}{k} p_0^k (1-p_0)^{n-k}, \tag{10.6.2}$$

which can be computed easily using R.

Example 10.6.1. Consider a specific instance of the above experiment where $n = 100$, and $X = 67$. Suppose we want to test the null hypothesis $p_0 = 0.5$. Then the observed value of d is

$$2X \log \frac{X/n}{p_0} + 2(n-X) \log \frac{1 - X/n}{1 - p_0} = 2 \times 67 \log \frac{0.67}{0.5} + 2 \times 33 \log \frac{0.33}{0.5} = 11.794$$

The required p-value then is

$$P(\Lambda(Y) \geq d)$$

where Y has a Binomial$(100, 0.5)$ distribution, and can be computed as follows.

```
d <- 2 * 67 * log(0.67 / 0.5) + 2 * 33 * log((1 - 0.67) / 0.5)
n <- 100
p0 <- 0.5
y <- 0:n
Py <- dbinom(y, size = n, prob = p0)
Ly <- ifelse(y %in% c(0, n),
             Inf,
             2 * y * log((y/n) / p0) + 2 * (n-y) * log((1-y/n) / (1-p0)))
sum(Py[ Ly >= d ])
```

```
[1] 0.0006412485
```

In the computation above, we need to treat $y = 0$ and $y = n$ specially, as direct computation of $\Lambda(y)$ fails for operations involving $\log 0$.

Not surprisingly, the p-value is quite small, suggesting that the observed value of $X = 67$ would be very unlikely if the null hypothesis $p = 0.5$ were true. In particular, as the p-value is less that 0.05, the null hypothesis would be rejected at level $\alpha = 0.05$.

It is instructive to investigate a little further and understand which values of k contribute to the sum in (10.6.2). Figure 10.1 plots values of $\Lambda(k)$ for a range of k centered around 50, along with the observed value $d = \Lambda(67)$. The traditional testing approach is to fix α and obtain a corresponding cutoff for d, usually by looking up the cutoff in a table. Suppose we fix $\alpha = 0.05$. It can be verified using R that $P(\Lambda(Y) \geq \Lambda(60)) = 0.057$ and $P(\Lambda(Y) \geq \Lambda(61)) = 0.035$. Thus, the highest probability of Type I error not exceeding $\alpha = 0.05$ that we can achieve in this example is 0.035, if we reject the null hypothesis when the observed $d \geq \Lambda(61) = 2.44$ (also shown in the figure).

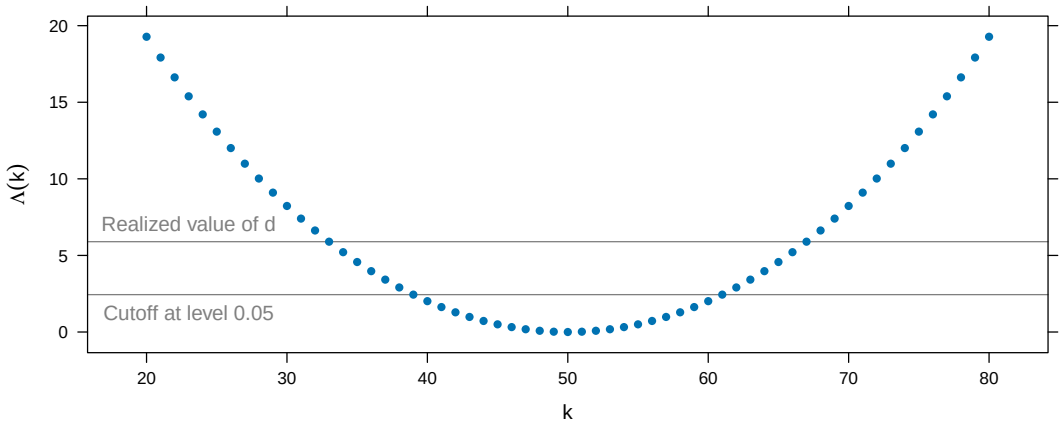


Figure 10.1: Values of $\Lambda(k)$ as a function of k for the Binomal test.

As can be seen by inspecting Figure 10.1, the null hypothesis will be rejected at level $\alpha = 0.05$ either if the realized X is 61 or higher, or if X is 39 or lower. In other words, this test is a *two-sided* test. ■

10.6.2 *Normal Test for Mean When Variance is Known*

Next we continue Example 10.4.2, where $X_1, X_2, \ldots, X_n$ is an i.i.d. sample from the Normal (μ, σ^2) distribution, where the variance σ^2 is known but the mean $\mu \in \mathcal{P} = \mathbb{R}$ is not. We are interested in testing the null hypothesis $\mu = \mu_0$, or equivalently, that $\mu \in \mathcal{P}_0 = \{\mu_0\}$, for a specific value μ_0. It is easily seen that

$$2 \log L(\mu; X_1, \ldots, X_n) = -\log 2\pi - \log \sigma^2 - \frac{1}{\sigma^2} \sum_i (X_i - \mu)^2$$

The unrestricted MLE of μ is $\hat{\mu} = \overline{X}$, and the restricted MLE is $\hat{\mu}_0 = \mu_0$. It follows that

$$\begin{aligned}
\Lambda(X_1, \ldots, X_n) &= 2 \log \frac{L(\hat{\mu}; X_1, \ldots, X_n)}{L(\hat{\mu}_0; X_1, \ldots, X_n)} \\
&= 2 \log L(\hat{\mu}; X_1, \ldots, X_n) - 2 \log L(\hat{\mu}_0; X_1, \ldots, X_n) \\
&= \frac{1}{\sigma^2} \left[\sum_i (X_i - \mu_0)^2 - \sum_i (X_i - \overline{X})^2 \right] \\
&= \frac{1}{\sigma^2} \left[n\mu_0^2 - 2n\mu_0 \overline{X} + n\overline{X}^2 \right] \\
&= \frac{n}{\sigma^2} (\overline{X} - \mu_0)^2 = \left(\frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} \right)^2
\end{aligned}$$

For a given realization of the sample $X_1, X_2, \ldots, X_n$, we need to first compute $d = \Lambda(X_1, X_2, \ldots, X_n)$ and then compute the p-value

$$P(\Lambda(Y_1, Y_2, \ldots, Y_n) \geq d)$$

where $Y_1, Y_2, \ldots, Y_n$ have a Normal (μ_0, σ^2) distribution. Now, we know that in that case $Z = \frac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}}$ has a Normal $(0, 1)$ distribution, and hence

$$\Lambda(Y_1, Y_2, \ldots, Y_n) = \left(\frac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}}\right)^2$$

has a χ^2_1 distribution. The p-value can thus be easily calculated using R. The following code computes the p-value for the data previously seen in Example 10.4.3.

```
x <- c(-5, -5, -4, -2, -2, 1, 2, 2, 3, 4, 4, 5, 7)
n <- length(x)
d <- (mean(x) / (3 / sqrt(n)))^2
pchisq(d, df = 1, lower.tail = FALSE)
```

```
[1] 0.3552259
```

As in the Binomial example, we can also derive a rejection cutoff for the observed d corresponding to a fixed level α. For example, with $\alpha = 0.05$, this cutoff would be the 0.95 quantile of the χ^2_1 distribution, which is

```
qchisq(0.95, df = 1)
```

```
[1] 3.841459
```

In other words, we reject the null hypothesis that $\mu = \mu_0$ if

$$d = \left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}\right)^2 \geq 3.84, \text{ or equivalently, if } \left|\sqrt{n}\,\frac{\overline{X} - \mu_0}{\sigma}\right| \geq \sqrt{3.84} = 1.96.$$

Thus our approach gives the same rule as the intuitive test derived earlier which which rejects the null hypothesis when the absolute value of $T_{\mu_0}(X_1, X_2, \ldots, X_n) = \sqrt{n}\left(\frac{\overline{X} - \mu_0}{\sigma}\right)$ is large. As with the test for Binomial proportion described earlier, this is a two-sided test. Recall that we have seen the cutoff of 1.96 before in Section 9.3 where we obtained a confidence interval for μ. This is not a coincidence, as the 0.95 quantile of χ^2_1 should indeed be the same as the 0.975 quantile (and the negative of the 0.025 quantile) of Normal $(0, 1)$.

More interestingly, it follows that given i.i.d. data $X_1, X_2, \ldots, X_n \sim$ Normal (μ, σ^2), a particular value of μ_0 will belong to the level $(1 - \alpha)$ confidence interval for μ obtained in Example 9.3.1 if and only if the null hypothesis that $\mu = \mu_0$ is accepted at level α. This observation applies more generally, and every hypothesis test can be used to generate a confidence region, and vice versa. This idea is often useful, especially in situations where a hypothesis test is available, but a confidence region may not be easy to obtain directly.

10.6.3 *One-sided Test for Normal Mean when Variance is Known*

An important variation of the above test is when we are interested in testing not a single point value of μ, but rather a range. For example, we may want to "reject" the null hypothesis only when the true mean is larger than the conjectured value, but not when it is lower. In this case, the null hypothsis is actually $\mu \leq \mu_0$, or $\mu \in \mathcal{P}_0 = (-\infty, \mu_0]$.

Although the unrestricted MLE $\hat{\mu} = \overline{X}$ remains unchanged, the restricted MLE $\hat{\mu}_0$ is now given by (See Exercise 10.6.1)

$$\hat{\mu}_0 = \begin{cases} \overline{X} & \text{if } \overline{X} \leq \mu_0 \\ \mu_0 & \text{otherwise.} \end{cases} = \min(\overline{X}, \mu_0)$$

The same calculations as above gives the likelihood ratio statistic as

$$\begin{aligned} \Lambda(X_1, X_2, \ldots, X_n) &= 2 \log \frac{L(\hat{\mu}; X_1, \ldots, X_n)}{L(\hat{\mu}_0; X_1, \ldots, X_n)} \\ &= \frac{1}{\sigma^2}\left[\sum_i (X_i - \hat{\mu}_0)^2 - \sum_i (X_i - \overline{X})^2\right] \\ &= \frac{n}{\sigma^2}(\overline{X} - \hat{\mu}_0)^2 = \begin{cases} 0 & \text{if } \overline{X} \leq \mu_0 \\ n(\overline{X} - \mu_0)^2/\sigma^2 & \text{otherwise.} \end{cases} \end{aligned}$$

We then need to compute the p-value

$$\max_{\mu \leq \mu_0} P_\mu(\Lambda(Y_1, Y_2, \ldots, Y_n) \geq d),$$

where $d = \Lambda(X_1, X_2, \ldots, X_n)$ is the observed likelihood ratio statistic.

If $\overline{X} \leq \mu_0$, $d = 0$ and hence the p-value is trivially 1. If $\overline{X} > \mu_0$, we must also have $d > 0$, and so $\Lambda(Y_1, \ldots, Y_n) \geq d$ if and only if $\overline{Y} > \mu_0$ and $n(\overline{Y} - \mu_0)^2/\sigma^2 \geq d$, or equivalently, $\sqrt{n}(\overline{Y} - \mu_0)/\sigma \geq \sqrt{d}$. Thus,

$$\begin{aligned} P_\mu(\Lambda(Y_1, \ldots, Y_n) \geq d) &= P_\mu(\sqrt{n}(\overline{Y} - \mu_0)/\sigma \geq \sqrt{d}) \\ &= P_\mu(\overline{Y} - \mu_0 \geq \sigma\sqrt{d/n}) \\ &= P_\mu(\overline{Y} - \mu \geq \sigma\sqrt{d/n} + \mu_0 - \mu) \\ &= P_\mu\left(\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \geq \sqrt{d} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z \geq \sqrt{d} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\sqrt{d} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right) \end{aligned}$$

where Z is a standard normal random variable with cumulative distribution function Φ. Now, as μ increases to μ_0, $\frac{\mu_0 - \mu}{\sigma/\sqrt{n}}$ decreases to 0, and $P_\mu(\Lambda(Y_1, \ldots, Y_n) \geq d)$ increases to $1 - \Phi(\sqrt{d})$. In other words, the maximum for calculating the p-value is achieved for $\mu = \mu_0$, giving p-value

$$1 - \Phi(\sqrt{d}) = P\left(Z \geq \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}\right).$$

As before, we can compute this p-value using R. To obtain a rejection cutoff, for example with $\alpha = 0.05$, we have

```
qnorm(0.95)
```

```
[1] 1.644854
```

We reject the null hypothesis if the p-value

$$P\left(Z \geq \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}\right) < 0.05,$$

or equivalently,

$$\sqrt{n}\,\frac{\overline{X} - \mu_0}{\sigma} > 1.645 \iff \overline{X} > \mu_0 + \frac{\sigma}{\sqrt{n}}\,1.645$$

Unlike the case where the null hypothesis allowed a single point value of μ, this is a one-sided test in terms of $\overline{X}$.

10.6.4 *Normal Test for Mean When Variance is Unknown*

As in the previous examples, suppose we have an i.i.d. sample $X_1, X_2, \ldots, X_n$ from the Normal(μ, σ^2) distribution, but now in addition to the mean μ, we assume more realistically that the variance σ^2 is also unknown. Formally, the parameters of the problem are $(\mu, \sigma^2) \in \mathcal{P} = \mathbb{R} \times (0, \infty)$. As in Section 10.6.2, we are interested in testing for a specific value of μ. To simplify notation, we will consider the null hypothesis $\mu = 0$ instead of the more general $\mu = \mu_0$. As noted earlier, this is not really a restriction; to test $\mu = \mu_0$, we can simply work with the transformed data $X_1 - \mu_0, X_2 - \mu_0, \ldots, X_n - \mu_0$. Under the null hypothesis, the restricted parameter set is $\mathcal{P}_0 = \{0\} \times (0, \infty)$.

Unlike the previous examples, we have two parameters in this problem, and we are interested in testing a null hypothesis that only puts restrictions on one of them. It is easy to see that the unrestricted MLEs of μ and σ^2 are given by $\hat{\mu} = \overline{X}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ (see Exercise 9.2.7).

The restricted MLE of μ is of course $\hat{\mu}_0 = 0$. It is easy to verify that the restricted MLE of σ^2 is $\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_0)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$. It follows that

$$\begin{aligned}
\Lambda(X_1, \ldots, X_n) &= 2\log\frac{L(\hat{\mu}, \hat{\sigma}^2; X_1, \ldots, X_n)}{L(\hat{\mu}_0, \hat{\sigma}_0^2; X_1, \ldots, X_n)} \\
&= 2\log L(\hat{\mu}, \hat{\sigma}^2; X_1, \ldots, X_n) - 2\log L(\hat{\mu}_0, \hat{\sigma}_0^2; X_1, \ldots, X_n) \\
&= n\log\hat{\sigma}_0^2 + \frac{1}{\hat{\sigma}_0^2}\sum_{i=1}^{n}(X_i - \hat{\mu}_0)^2 - n\log\hat{\sigma}^2 - \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 \\
&= n\log\frac{1}{n}\sum_{i=1}^{n} X_i^2 + n\left(\frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} X_i^2}\right) - n\log\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 - n\left(\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right) \\
&= n\log\frac{\sum X_i^2}{\sum(X_i - \overline{X})^2}
\end{aligned}$$

We are interested in the distribution of $\Lambda(Y_1, \ldots, Y_n)$ when $Y_1, Y_2, \ldots, Y_n$ has the Normal $(0, \sigma^2)$ distribution (i.e., $\mu = 0$). In general, this distribution could depend on the other parameters, namely, σ^2 in this example. As we saw in Section 10.4.3, this does not happen here; that is, the distribution of $\Lambda(Y_1, \ldots, Y_n)$ does *not* depend on the value of σ^2 when $\mu = 0$. To see this, notice that

$$\begin{aligned}
\Lambda(Y_1, \ldots, Y_n) &= n\log\frac{\sum Y_i^2}{\sum(Y_i - \overline{Y})^2} \\
&= n\log\frac{\sum(Y_i/\sigma)^2}{\sum(Y_i/\sigma - \overline{Y}/\sigma)^2} \\
&= n\log\frac{\sum Z_i^2}{\sum(Z_i - \overline{Z})^2} = \Lambda(Z_1, \ldots, Z_n)
\end{aligned}$$

where $Z_i = Y_i/\sigma \sim$ Normal $(0, 1)$ has distribution free of σ. In other words,

$$P_{\sigma^2}(\Lambda(Y_1, \ldots, Y_n) \geq d) = P_1(\Lambda(Y_1, \ldots, Y_n) \geq d)$$

for all σ^2, where P_{σ^2} denotes probability calculations when $Y_1, Y_2, \ldots, Y_n$ are i.i.d. from Normal$(0, \sigma^2)$. In particular, for an observed value $d = \Lambda(X_1, \ldots, X_n)$, the p-value is given by

$$\begin{aligned}
P(\Lambda(Z_1, \ldots, Z_n) \geq d) &= P\left(n\log\frac{\sum Z_i^2}{\sum(Z_i - \overline{Z})^2} \geq d\right) \\
&= P\left(\frac{\sum Z_i^2}{\sum(Z_i - \overline{Z})^2} \geq e^{d/n}\right) \\
&= P\left(\frac{\sum(Z_i - \overline{Z})^2}{\sum Z_i^2} \leq e^{-d/n}\right)
\end{aligned}$$

where $Z_1, Z_2, \ldots, Z_n$ are i.i.d. from Normal$(0,1)$. Now, simple algebraic manipulation shows that

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} (Z_i - \overline{Z} + \overline{Z})^2 = \sum_{i=1}^{n} (Z_i - \overline{Z})^2 + n\overline{Z}^2$$

Recall that $\overline{Z} \sim \text{Normal}(0, \frac{1}{n})$ and so $n\overline{Z}^2 = (\sqrt{n}\overline{Z})^2 \sim \chi^2_1$. An application of Theorem 8.1.10 further tells us that (i) $\sum(Z_i - \overline{Z})^2 \sim \chi^2_{n-1}$, and (ii) $\overline{Z}^2$ is independent of $\sum(Z_i - \overline{Z})^2$. Recall Example 8.1.5 to note that the χ^2_n distribution is the same as the Gamma$(\frac{n}{2}, \frac{1}{2})$ distribution. In other words,

$$\sum_{i=1}^{n} (Z_i - \overline{Z})^2 \sim \text{Gamma}\left(\frac{n-1}{2}, \frac{1}{2}\right) \text{ and } n\overline{Z}^2 \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right),$$

independently of each other. It then follows from Example 5.5.11 that

$$\frac{\sum_{i=1}^{n} (Z_i - \overline{Z})^2}{\sum_{i=1}^{n} Z_i^2} = \frac{\sum_{i=1}^{n} (Z_i - \overline{Z})^2}{\sum_{i=1}^{n} (Z_i - \overline{Z})^2 + n\overline{Z}^2}$$

has the Beta$\left(\frac{n-1}{2}, \frac{1}{2}\right)$ distribution. Thus, given an observed value for the likelihood ratio statistic

$$d = \Lambda(X_1, \ldots, X_n) = n \log \frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2},$$

the corresponding p-value can be computed as

$$P\left(U \leq e^{-d/n}\right)$$

where U has the Beta$\left(\frac{n-1}{2}, \frac{1}{2}\right)$ distribution.

For the data in Example 10.4.3, the p-value can be computed as follows.

```
x <- c(-5, -5, -4, -2, -2, 1, 2, 2, 3, 4, 4, 5, 7)
n <- length(x)
mu_x <- mean(x)
d <- n * log( sum(x^2) / sum((x - mu_x)^2) )
d
```

```
[1] 0.5151229
```

```
pbeta(exp(-d/n), 0.5 * (n-1), 0.5)
```

```
[1] 0.4994148
```

To obtain the rejection region at level α, we need to start with the lower α quantile, say b, which for $\alpha = 0.05$ can be computed as follows.

```
qbeta(0.05, 0.5 * (n-1), 0.5)
```

```
[1] 0.7165366
```

We reject at level α if $e^{-d/n} \leq b$, that is, if

$$\frac{\sum(X_i - \overline{X})^2}{\sum(X_i - \overline{X})^2 + n\overline{X}^2} \leq b,$$

or equivalently, if

$$\begin{aligned} \frac{\sum(X_i - \overline{X})^2 + n\overline{X}^2}{\sum(X_i - \overline{X})^2} &= 1 + n\frac{\overline{X}^2}{\sum(X_i - \overline{X})^2} \geq \frac{1}{b}, \\ \iff \frac{\overline{X}^2}{\frac{1}{n-1}\sum(X_i - \overline{X})^2} &\geq \frac{n-1}{n}\left(\frac{1}{b} - 1\right) \\ \iff \left|\frac{\overline{X}}{S}\right| &\geq \sqrt{\frac{n-1}{n}\left(\frac{1}{b} - 1\right)} \\ \iff \left|\sqrt{n}\,\frac{\overline{X}}{S}\right| &\geq \sqrt{(n-1)\left(\frac{1}{b} - 1\right)}, \end{aligned}$$

where S^2 is the sample variance (see Definition 7.1.5 in Chapter 8). Note the similarity of the rejection region with the one in Section 10.6.2 which addressed the analogous problem when σ^2 is known; as in that case, this is also a two-sided test. It follows from Corollary 8.1.11 that $\sqrt{n}\,\frac{\overline{X}}{S}$ has the t_{n-1} distribution when $\mu = 0$, and so p-values of cutoffs can be computed using the t distribution as well. It can be numerically verified that a two-sided t-test rejection cutoff is equivalent to the Beta test cutoff as follows.

```
b <- qbeta(0.05, 0.5 * (n-1), 0.5)
sqrt((n-1) * (1/b - 1))
```

```
[1] 2.178813
```

```
qt(0.025, df = n-1, lower.tail = FALSE)
```

```
[1] 2.178813
```

Before moving on, we make some observations and remarks about this important example. It should be easy to see that for the general null hypothesis $\mu = \mu_0$, the Beta statistic derived above generalizes to

$$e^{-\Lambda(X_1,\ldots,X_n)/n} = \frac{\sum(X_i - \overline{X})^2}{\sum(X_i - \mu_0)^2} = \frac{\sum(X_i - \overline{X})^2}{\sum(X_i - \overline{X})^2 + n(\overline{X} - \mu_0)^2}$$

and the corresponding statistic with the t_{n-1} distribution is

$$\sqrt{n}\,\frac{\overline{X} - \mu_0}{S}.$$

The latter form of the statistic is more conventional, not least due to its intuitively appealing form and similarity with the Normal statistic $\sqrt{n}\,\frac{\overline{X}-\mu_0}{\sigma}$ in the case when σ^2 is known, as it can be motivated as a modification where σ^2 is replaced by an estimator when it is unknown. As with the case where σ^2 is known, the test can be adjusted for a one-sided null hypothesis of the form $\mu \leq \mu_0$ or $\mu \geq \mu_0$, resulting in a one-sided rejection region in terms of the t statistic. Even though this test is derived under the assumption of Normality, it performs well for moderate departures from this assumption, and is among the most common statistical tests used in practice.

Another equivalent form of the test is also important. In its squared form, the statistic

$$\frac{n(\overline{X} - \mu_0)^2}{S^2}$$

has a $F_{1,n-1}$ distribution when $\mu = \mu_0$ (see Example 8.1.7). This form and its generalizations are also commonly used. A typical example is the problem of testing for equality of means of two or more populations, which we discuss next.

10.6.5 *The Two-sample Test for Equality of Population Means*

Connect this to a test of independence of treatment vs outcome, where outcome is now continuous rather than discrete. Or maybe keep as-is and use another section or subsection to make the connection and tie up other loose ends.

Hypothesis tests may be used to compare two samples to each other to see if the populations they were derived from are similar. This is of particular use in many applications. For instance: Are the political preferences of people of one region different from another? Are test scores at one school better than those at another school? These questions could be approached by taking random samples from each population and comparing them with each other.

Suppose $X_1, X_2, \ldots, X_m$ is an i.i.d. sample from a distribution $X \sim \text{Normal}(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_n$ is an i.i.d. sample from a distribution $Y \sim \text{Normal}(\mu_2, \sigma_2^2)$ independent of the X_j variables. Assume that μ_1 and μ_2 are unknown, as well as σ_1^2 and σ_2^2. How might we test the null hypothesis that $\mu_1 = \mu_2$ against the alternative hypothesis $\mu_1 \neq \mu_2$? It turns out that this problem is not easy to solve in its general form, and we will only consider the problem with the additional assumption that $\sigma_1^2 = \sigma_2^2$. We will use σ^2 to denote this common variance.

Based on the examples we have seen so far, we might guess that a good test would be based on $\overline{X} - \overline{Y}$, as this should be close to 0 if the null hypothesis that $\mu_1 = \mu_2$ were true. It is simple to check that $\overline{X} - \overline{Y}$ has a Normal distribution with mean $\mu_1 - \mu_2$ and variance $\left(\frac{1}{m} + \frac{1}{n}\right)\sigma^2$. If we can obtain an estimator S^2 of σ^2 that is independent of $\overline{X} - \overline{Y}$ and has a suitably scaled χ^2 distribution, we could plausibly expect that

$$\frac{\overline{X} - \overline{Y}}{S\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

would have a t distribution. In fact, it is not difficult to find such an estimator S^2. Consider the standard unbiased estimators of σ^2 from the two independent populations:

$$\begin{aligned} S_1^2 &= \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \overline{X})^2 \\ S_2^2 &= \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \overline{Y})^2 \end{aligned}$$

It follows from Theorem 8.1.10 that $(m-1)S_1^2/\sigma^2 \sim \chi^2_{m-1}$ independently of $\overline{X}$ and $(n-1)S_2^2/\sigma^2 \sim \chi^2_{n-1}$ independently of $\overline{Y}$. This suggests the natural estimator

$$S^2 = \frac{1}{m+n-2}\left[\sum_{i=1}^{m}(X_i - \overline{X})^2 + \sum_{j=1}^{n}(Y_j - \overline{Y})^2\right]$$

which is independent of both $\overline{X}$ and $\overline{Y}$ (and hence $\overline{X} - \overline{Y}$). By Example 5.5.6, $(m+n-2)S^2/\sigma^2 \sim \chi^2_{m+n-2}$, and so S^2 is an unbiased estimator of σ^2. It follows from Corollary 8.1.11 that

$$\frac{\overline{X} - \overline{Y}}{S\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

has a t_{m+n-2} distribution. This is the basis for the standard two-sided test for this problem, and this solution generalizes to the one-sided null hypotheses $\mu \leq \mu_0$ and $\mu \geq \mu_0$ in the usual way. As in the test for the mean of a single Normal population, the squared version of the statistic

$$\frac{\frac{mn}{m+n}(\overline{X} - \overline{Y})^2}{S^2}$$

has the $F_{1,m+n-2}$ distribution.

It turns out, not surprisingly, that this is equivalent to the test derived using the likelihood ratio statistic. The likelihood function for this model is (suppressing the dependence on $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ for brevity)

$$\begin{aligned} L(\mu_1, \mu_2, \sigma^2) &= \prod_{i=1}^{m} \frac{1}{\sqrt{n\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(X_i - \mu_1)^2\right\} \prod_{j=1}^{n} \frac{1}{\sqrt{n\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(Y_j - \mu_2)^2\right\} \\ &= \left(\frac{1}{\sqrt{n\pi}\sigma}\right)^{m+n} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{m}(X_i - \mu_1)^2 + \sum_{j=1}^{n}(Y_j - \mu_2)^2\right]\right\} \end{aligned}$$

It is easy to verify, again following the approach of Exercise 9.2.7, that the unrestricted MLEs are

$$\begin{aligned} \hat{\mu}_1 &= \overline{X} \\ \hat{\mu}_2 &= \overline{Y} \\ \hat{\sigma}^2 &= \frac{1}{m+n}\left[\sum_{i=1}^{m}(X_i - \hat{\mu}_1)^2 + \sum_{j=1}^{n}(Y_j - \hat{\mu}_2)^2\right], \end{aligned}$$

and therefore

$$\begin{aligned} L(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2) &= \left(\frac{1}{\sqrt{n\pi}\hat{\sigma}}\right)^{m+n} \exp\left\{-\frac{m+n}{2}\left[\frac{\sum_{i=1}^{m}(X_i - \hat{\mu}_1)^2 + \sum_{j=1}^{n}(Y_j - \hat{\mu}_2)^2}{\sum_{i=1}^{m}(X_i - \hat{\mu}_1)^2 + \sum_{j=1}^{n}(Y_j - \hat{\mu}_2)^2}\right]\right\} \\ &= \left(n\pi e\hat{\sigma}^2\right)^{-\frac{m+n}{2}}. \end{aligned}$$

Similarly, under the null hypothesis $\mu_1 = \mu_2$, the restricted MLEs are

$$\begin{aligned} \hat{\mu}_0 = \hat{\mu}_{10} = \hat{\mu}_{20} &= \frac{1}{m+n}(m\overline{X} + n\overline{Y}) \\ \hat{\sigma}_0^2 &= \frac{1}{m+n}\left[\sum_{i=1}^{m}(X_i - \hat{\mu}_0)^2 + \sum_{j=1}^{n}(Y_j - \hat{\mu}_0)^2\right] \end{aligned}$$

and therefore

$$L(\hat{\mu}_{10}, \hat{\mu}_{20}, \hat{\sigma}_0^2) = \left(n\pi e\hat{\sigma}_0^2\right)^{-\frac{m+n}{2}}.$$

The likelihood ratio statistic is thus

$$\Lambda(X_1, \ldots, X_m, Y_1, \ldots, Y_n) = 2\log\frac{L(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2)}{L(\hat{\mu}_{10}, \hat{\mu}_{20}, \hat{\sigma}_0^2)} = (m+n)\log\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right),$$

which is essentially equivalent to

$$\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} = \frac{\sum_{i=1}^{m}(X_i - \hat{\mu}_0)^2 + \sum_{j=1}^{n}(Y_j - \hat{\mu}_0)^2}{\sum_{i=1}^{m}(X_i - \overline{X})^2 + \sum_{j=1}^{n}(Y_j - \overline{Y})^2}.$$

The test rejects the null hypothesis when this ratio is large, which is in line with our intuition because if the null hypothesis is not true, then we would expect the deviations in the numerator (from a common mean) to be larger than the deviations in the denominator (from group-specific means). The denominator clearly has a χ^2_{m+n-2} distribution (by Theorem 8.1.10 followed by Example 5.5.6), but the distribution of the numerator is not immediately obvious. To simplify the numerator, note that

$$\sum_{i=1}^{m}(X_i - \hat{\mu}_0)^2 = \sum_{i=1}^{m}(X_i - \overline{X} + \overline{X} - \hat{\mu}_0)^2 = \sum_{i=1}^{m}(X_i - \overline{X})^2 + m(\overline{X} - \hat{\mu}_0)^2 + 0.$$

Recall that $\hat{\mu}_0 = \frac{1}{m+n}(m\overline{X} + n\overline{Y})$, and so

$$\overline{X} - \hat{\mu}_0 = \frac{(m+n)\overline{X} - m\overline{X} - n\overline{Y}}{m+n} = \frac{1}{m+n}(m\overline{X} + n\overline{X} - m\overline{X} - n\overline{Y}) = \frac{n}{m+n}(\overline{X} - \overline{Y}),$$

and consequently,

$$\sum_{i=1}^{m}(X_i - \hat{\mu}_0)^2 = \sum_{i=1}^{m}(X_i - \overline{X})^2 + \frac{mn^2}{(m+n)^2}(\overline{X} - \overline{Y})^2.$$

Analogously,

$$\sum_{j=1}^{n}(Y_j - \hat{\mu}_0)^2 = \sum_{j=1}^{n}(Y_j - \overline{Y})^2 + \frac{m^2 n}{(m+n)^2}(\overline{X} - \overline{Y})^2$$

and therefore the numerator term simplifies to

$$\sum_{i=1}^{m}(X_i - \hat{\mu}_0)^2 + \sum_{j=1}^{n}(Y_j - \hat{\mu}_0)^2 = \sum_{i=1}^{m}(X_i - \overline{X})^2 + \sum_{j=1}^{n}(Y_j - \overline{Y})^2 + \frac{mn}{m+n}(\overline{X} - \overline{Y})^2.$$

As before, it follows from Theorem 8.1.10 and the discussion above that $\overline{X} - \overline{Y}$ is independent of the denominator

$$\sum_{i=1}^{m}(X_i - \overline{X})^2 + \sum_{j=1}^{n}(Y_j - \overline{Y})^2$$

and that $\frac{mn}{m+n}(\overline{X} - \overline{Y})^2$ has a χ^2_1 distribution. Arguing similarly as we did in Section 10.6.4, it follows that

$$\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} = 1/\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} = \frac{\sum_{i=1}^{m}(X_i - \overline{X})^2 + \sum_{j=1}^{n}(Y_j - \overline{Y})^2}{\sum_{i=1}^{m}(X_i - \overline{X})^2 + \sum_{j=1}^{n}(Y_j - \overline{Y})^2 + \frac{mn}{m+n}(\overline{X} - \overline{Y})^2}$$

has the Beta($\frac{m+n-2}{2}, \frac{1}{2}$) distribution (test rejects null when small), and

$$\frac{\frac{mn}{m+n}(\overline{X}-\overline{Y})^2}{\frac{1}{m+n-2}\left[\sum(X_i-\overline{X})^2+\sum(Y_j-\overline{Y})^2\right]} = \frac{\frac{mn}{m+n}(\overline{X}-\overline{Y})^2}{S^2}$$

has the $F_{1,m+n-2}$ distribution (test rejects null when large), and both are equivalent to the "intuitive" two-sided t test derived above.

10.6.6 *Equality of Population Means with Different Variances*

The examples above may give the impression that the likelihood ratio approach always leads to a useful test statistic. The next example, which is a simple and natural extension of the previous problem, shows that this is not so.

Recall the setup of the previous problem, where we suppose that $X_1, X_2, \ldots, X_m$ is an i.i.d. sample from $X \sim \text{Normal}(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_n$ is an independent i.i.d. sample from $Y \sim \text{Normal}(\mu_2, \sigma_2^2)$ independent of the X_j variables. We are still interested in testing the null hypothesis that $\mu_1 = \mu_2$ against the alternative hypothesis $\mu_1 \neq \mu_2$, but this time we do not wish to assume that $\sigma_1^2 = \sigma_2^2$.

The unrestricted MLEs of the parameters are straightforward, as the X and Y observations do not share any common parameters. The detailed calculations for the restricted case are more involved. The details are left as an exercise, but it can be shown that the MLEs satisfy the following equations:

$$\begin{aligned}
\hat{\mu}_0 = \hat{\mu}_{10} = \hat{\mu}_{20} &= \frac{m\hat{\sigma}_{20}^2\overline{X} + n\hat{\sigma}_{10}^2\overline{Y}}{m\hat{\sigma}_{20}^2 + n\hat{\sigma}_{10}^2} \\
\hat{\sigma}_{10}^2 &= \frac{1}{m}\sum_{i=1}^{m}(X_i - \hat{\mu}_0)^2 \\
\hat{\sigma}_{20}^2 &= \frac{1}{n}\sum_{j=1}^{n}(Y_j - \hat{\mu}_0)^2
\end{aligned}$$

An exact solution can be obtained, although in practice an iterative approach works well. It can be further shown that the likelihood ratio statistic can be expressed as follows in terms of these estimates as follows.

$$\Lambda(X_1, \ldots, X_m, Y_1, \ldots, Y_n) = 2\log\frac{L(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, , \hat{\sigma}_2^2)}{L(\hat{\mu}_{10}, \hat{\mu}_{20}, \hat{\sigma}_{10}^2, \hat{\sigma}_{20}^2)} = m\log\left(\frac{\hat{\sigma}_{10}^2}{\hat{\sigma}_1^2}\right) + n\log\left(\frac{\hat{\sigma}_{20}^2}{\hat{\sigma}_2^2}\right).$$

This is where the usual procedure breaks down. Unlike in the previous examples, the distribution of this quantity is *not* completely determined when the null hypothesis holds, because it depends on the unknown ratio σ_1^2/σ_2^2. Thus, even using simulation to estimate the p-value is not feasible.

As alluded to earlier, one benefit of the likelihood ratio test is that even when the distribution of the statistic is difficult to study, a powerful result gives its *asymptotic* distribution under fairly general conditions. Applied to this problem, this result says that as $m, n \to \infty$, the distribution of

$\Lambda(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ converges to a χ^2_1 distribution under the null hypothesis. This is indeed true; however, for small sample sizes m and n, calculating p-values using this null distribution leads to substantially larger probability of Type I error than nominally specified. A modification works quite well in practice, but the details are beyond our scope. For us, this example serves to illustrate the limitations of the likelihood ratio test approach.

EXERCISES

Ex. 10.6.1. Consider an i.i.d. sample $X_1, X_2, \ldots, X_n \sim \text{Normal}\left(\mu, \sigma^2\right)$. Show that the restricted MLE $\hat{\mu}_0$ for $\mu \in (-\infty, \mu_0]$ is now given by

$$\hat{\mu}_0 = \begin{cases} \overline{X} & \text{if } \overline{X} \leq \mu_0 \\ \mu_0 & \text{otherwise.} \end{cases} = \min(\overline{X}, \mu_0)$$

Hint: Notice that the likelihood function $L(\mu; X_1, \ldots, X_n)$, as a function of μ, is proportional to a normal density with mean $\overline{X}$, and is therefore maximum at $\overline{X}$ and symmetrically decreasing on either side of it.

10.7 TESTING FOR GOODNESS OF FIT

We now return to the problems discussed earlier in this chapter, in Sections 10.2 and 10.3, which we left largely unresolved. We start with the "goodness of fit" problem, and specifically Example 10.2.1 which we can generalize as follows.

Consider a random experiment (such as rolling a die) which has k possible outcomes. As in Example 3.2.12, let p_j represent the probability that any individual trial results in the j-th outcome, and let X_j represent the number of the n trials that result in the j-th outcome. The joint distribution of the random variables $(X_1, X_2, \ldots, X_k)$ is then a multinomial distribution with parameters n and $p = (p_1, p_2, \ldots, p_k)$. Formally, the parameter space for the problem is

$$\mathcal{P} = \left\{ p = (p_1, p_2, \ldots, p_k) : 0 \leq p_j \leq 1 \text{ for all } j, \sum_{j=1}^{k} p_j = 1 \right\}.$$

We wish to test the null hypothesis that $p = p_0$ for some $p_0 = (p_{01}, p_{02}, \ldots, p_{0k})$. In Example 10.2.1, where the experiment was rolling a die, we had $k = 6$ and $p_0 = (1/6, 1/6, \ldots, 1/6)$.

As $\mathcal{P}_0 = \{p_0\}$ is a singleton set, the MLE $\hat{p}_0$ under the null hypothesis is simply p_0. As we saw in Exercise 9.2.11, the unrestricted MLE $\hat{p}$ of p is given by the coordinatewise sample proportions, that is, $\hat{p}_j = X_j/n$.

From the multinomial distribution obtained in Example 3.2.12, it follows easily that the likelihood function, given observed data $X_1, X_2, \ldots, X_k$, has the form

$$L(p; X_1, X_2, \ldots, X_k) = \frac{n!}{X_1! X_2! \ldots X_k!} \prod_{j=1}^{k} p_j^{X_j}, \text{ for } p \in \mathcal{P}$$

and so, suppressing the dependence of $L(p)$ on $X_1, X_2, \ldots, X_k$, we have

$$\log L(p) = \log n! - \sum_{j=1}^{k} \log X_j! + \sum_{j=1}^{k} X_j \log p_j, \text{ for } p \in \mathcal{P}.$$

Substituting $\hat{p}$ and $\hat{p}_0$, we have

$$\Lambda(X_1, \ldots, X_k) = 2 \log \frac{L(\hat{p})}{L(\hat{p}_0)} = 2 \sum_{j=1}^{k} X_j \log \frac{\hat{p}_j}{\hat{p}_{0j}} = 2 \sum_{j=1}^{k} X_j \log \frac{X_j}{n p_{0j}}.$$

It is conventional in this context to define $E_j = n p_{0j}$, representing the expected value of X_j. This leads to the test statistic

$$\Lambda(X_1, \ldots, X_k) = 2 \sum_{j=1}^{k} X_j \log \frac{X_j}{E_j}. \tag{10.7.1}$$

Example 10.7.1. In a survey, a class of $n = 71$ students were asked what their birth month was, and the following summary results were obtained:

January	7	February	4	March	5	April	8
May	2	June	6	July	6	August	7
September	3	October	8	November	9	December	6

Using this data, we wish to test the null hypothesis that all days in the year are equally likely as birthdays. We will ignore the possibility of leap years for simplicity and assume that there are 365 days in a year. We also assume that the class represents an i.i.d. sample from some larger population. We can set up the problem and calculate the test statistic in R as follows.

```
X <- c(7, 4, 5, 8, 2, 6, 6, 7, 3, 8, 9, 6)
n <- sum(X)
p0 <- c(31, 28, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31) / 365
E <- n * p0
Lambda_x <- 2 * sum(X * log(X / E))
Lambda_x
```

```
[1] 9.008024
```

To compute the p-value for this test, we need the distribution of $\Lambda(Y_1, \ldots, Y_k)$ when $(Y_1, Y_2, \ldots, Y_k)$ follow the multinomial distribution with parameters n and p_0. Unfortunately, this distribution is not easy to compute explicitly. However, as the distribution of $(Y_1, Y_2, \ldots, Y_k)$ is completely known

(because $\mathcal{P}_0$ has only one element, p_0), we can easily simulate values of $\Lambda(Y_1, \ldots, Y_k)$ and estimate the p-value. We can do this, using 1000000 replications, as follows.

```
Lambda_sim <-
    replicate(1000000,
    {
        Y <- rmultinom(1, size = n, prob = p0)
        2 * sum(Y * log(Y / E), na.rm = TRUE) # Use na.rm to allow for Y = 0
    })
uprob <- sum(Lambda_sim >= Lambda_x) / 1000000
uprob
```

```
[1] 0.648674
```

A remarkable result, a version of which is stated without proof below in Section 10.7.1, tells us that even if we had been unable to estimate the exact p-value, the asymptotic distribution of $\Lambda(Y_1, \ldots, Y_k)$ (as the sample size $n \to \infty$) *is* known, and is in fact χ^2_{k-1}. This asymptotic result allows us to obtain an approximate p-value for this example as follows.

```
pchisq(Lambda_x, df = 11, lower.tail = FALSE)
```

```
[1] 0.6211516
```

While not exactly the same, this is reasonably close to the exact p-value estimated using simulation. ■

10.7.1 *Asymptotic Distribution of the Likelihood Ratio Test Statistic*

Much of the appeal of the likelihood-based approach we have explored in this chapter comes from a very powerful result that states that the null distribution of the likelihood ratio test statistic can be obtained asymptotically, even if it cannot be computed exactly for any finite sample size. In simplified form, the result may be stated as follows.

Theorem 10.7.2. *(Wilks' Theorem) Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a distribution X with a probability mass function or probability density function $f(x)$, where $f(x) = f(x \mid p_1, p_2, \ldots, p_d) = f(x \mid p)$ depends on one or more unknown parameters $p = (p_1, p_2, \ldots, p_d) \in \mathcal{P} \subset \mathbb{R}^d$ for some $d \geq 1$. We are interested in testing the null hypothesis that $p \in \mathcal{P}_0$, where $\mathcal{P}_0$ is a proper subset of $\mathcal{P}$. Then, under certain conditions, the distribution of the likelihood ratio statistic $\Lambda(X_1, X_2, \ldots, X_n)$ defined in (10.5.2) converges in distribution to the χ^2_k distribution as $n \to \infty$, where k is the number of independent constraints that $\mathcal{P}_0$ puts on $\mathcal{P}$, or equivalently, the difference in the number of independent parameters in $\mathcal{P}$ and $\mathcal{P}_0$.*

The proof of this result is beyond the scope of this book. In fact, even a proper statement would be unnecessarily complicated. For our purposes, it is sufficient to know that the result is applicable

in most situations, as long as two important conditions are met: The number of independent parameters in $\mathcal{P}$ and $\mathcal{P}_0$ are fixed finite numbers, and no $p \in \mathcal{P}_0$ is on the "boundary" of $\mathcal{P}$. We will not get into the precise details of what it means for p to be in the boundary of $\mathcal{P}$, but in the goodness of fit example, an obvious way for this to happen would be if one of the components p_{0j} equals 0. If that were the case, the result would not be applicable.

To appreciate the power of this result, note that it applies even when the null parameter space $\mathcal{P}_0$ is not a singleton set. Recall that to solve a testing problem we need two things: to find a suitable test statistic, and then to find its null distribution. The likelihood ratio approach often gives us a test statistic in situations where no natural candidate is available. In Example 10.7.1 above, this approach gave a test statistic that we would probably not consider to be natural. However, once we determine the test statistic, finding its null distribution is simple; even though this null distribution is "unkown" in the sense that we cannot identify it as a standard distribution, it is still possible to simulate from it because $\mathcal{P}_0$ is a singleton set, so any probabilities related to that distribution can be computed as precisely as we wish.

The situation is fundamentally different when $\mathcal{P}_0$ contains multiple (usually infinitely many) values. The likelihood ratio test statistic need not have a single "null distribution", but rather a different one for every $p_0 \in \mathcal{P}_0$, making the simulation approach impractical. Theorem 10.7.2 comes to our rescue in such sitations, giving us a single null distribution that is at least asymptotically valid. We will see an example of such a test in Section 10.8.

10.7.2 *The Standard χ^2 Test for Goodness of Fit*

We conclude this section with a discussion of a much more well known test for the goodness of fit problem, given by the following test statistic.

$$T(X_1, \ldots, X_k) = \sum_{j=1}^{k} \frac{(X_j - E_j)^2}{E_j}$$

This is an intutively appealing test statistic. As with the likelihood ratio test described above, the exact distribution of $T(X_1, \ldots, X_k)$ cannot be computed explicitly, but can be studied using simulation for any specific problem. This distribution also converges to the same χ^2_{k-1} distribution as $n \to \infty$, which as we will see soon, is not a coincidence. This asymptotic result can be proved without appealing to the more general Theorem 10.7.2, although the proof is still beyond the scope of this book. For Example 10.7.1, the resulting test statistic and the corresponding p-value can be obtained as follows.

```
T_x <- sum((X-E)^2 / E)
T_x
```

```
[1] 8.110577
```

```
pchisq(T_x, df = 11, lower.tail = FALSE)
```

```
[1] 0.7033655
```

One may wonder, given the same asymptotic null distribution of $\Lambda(X_1, \ldots, X_k)$ and $T(X_1, \ldots, X_k)$, whether they are related to each other. The answer is that they are indeed related. To see this, define $\varepsilon_j = X_j - E_j$, so that we can write (10.7.1) as

$$\begin{aligned}\Lambda(X_1, \ldots, X_k) &= 2\sum_{j=1}^{k} X_j \log \frac{X_j}{E_j} \\ &= 2\sum_{j=1}^{k} (E_j + \varepsilon_j) \log\left(1 + \frac{\varepsilon_j}{E_j}\right)\end{aligned}$$

Now, for x close to 0, we can write $\log(1+x) \approx x - \frac{1}{2}x^2$, so we can write, provided $\frac{\varepsilon_j}{E_j} \approx 0$,

$$\Lambda(X_1, \ldots, X_k) \approx 2\sum_{j=1}^{k} \left[(E_j + \varepsilon_j)\left(\frac{\varepsilon_j}{E_j} - \frac{1}{2}\frac{\varepsilon_j^2}{E_j^2}\right)\right] \tag{10.7.2}$$

It is easy to see that

$$(E_j + \varepsilon_j)\left(\frac{\varepsilon_j}{E_j} - \frac{1}{2}\frac{\varepsilon_j^2}{E_j^2}\right) = (X_j - E_j) + \frac{1}{2}\frac{(X_j - E_j)^2}{E_j}\left(1 - \frac{X_j - E_j}{E_j}\right)$$

Substituting in (10.7.2) and noting that $\sum X_j = \sum E_j = n$ and hence $\sum(X_j - E_j) = 0$, we have

$$\Lambda(X_1, \ldots, X_k) \approx 0 + 2\,\frac{1}{2}\sum_{j=1}^{k}\left[\frac{(X_j - E_j)^2}{E_j}\left(1 - \frac{X_j - E_j}{E_j}\right)\right]$$

This approximation is of course only valid when $\frac{\varepsilon_j}{E_j} = \frac{X_j - E_j}{E_j} \approx 0$. In that case, we can further approximate $\Lambda(X_1, \ldots, X_k)$ by assuming that $1 - \frac{X_j - E_j}{E_j} \approx 1$, to get

$$\Lambda(X_1, \ldots, X_k) \approx \sum_{j=1}^{k} \frac{(X_j - E_j)^2}{E_j} = T(X_1, \ldots, X_k).$$

It is not particularly difficult to show that $\frac{\varepsilon_j}{E_j} \xrightarrow{p} 0$ as $n \to 0$, and thus establish using Slutsky's theorem that $\Lambda(X_1, \ldots, X_k)$ and $T(X_1, \ldots, X_k)$ have the same asymptotic distribution.

A natural next question is to ask which of these tests is better in terms of their power to identify situations where the null hypothesis does not hold. The results we have cited so far do not give a clear answer, but simulation studies can be used to get an indication.

10.8 TESTING FOR INDEPENDENCE OF CATEGORICAL ATTRIBUTES

Next we consider the problem of testing whether two categorical attributes are independent. There are several formulations of this problem that give potentially different answers. We start with the multinomial formulation discussed earlier in Section 10.3.

10.8.1 *The Multinomial Model*

Recall that in Section 10.3, we described a multinomial model for this problem parameterized by the parameter vector $p = (p_{11}, p_{21}, p_{12}, p_{22})$. This model is natural when the units studied in the problem can be viewed as an i.i.d. sample from some population, as would be appropriate in an observational study. An alternative formulation that is more appropriate for randomized controlled trials is parameterized by $p = (\pi_1, \pi_2, q_{11}, q_{12}, q_{21}, q_{22})$, where π_1 and π_2 are the probabilities of a specific unit being assigned to treatment 1 and treatment 2 respectively, and $q_{1\ell}$ and $q_{2\ell}$ are corresponding conditional probabilities of outcome ℓ. As the two formulations are equivalent (as long as units are allocated treatment independently), we will only consider the first formulation.

We have already obtained (see Section 10.3.2) maximum likelihood estimates of p under the unconstrained model, as well as under the null hypothesis of independence which restricts the parameter values to

$$\mathcal{P}_0 = \{p : p_{k\ell} = p_{k\circ} p_{\circ\ell} \text{ for } k, \ell = 1, 2\}.$$

To obtain the likelihood ratio statistic, we can follow the same calculations as in the goodness of fit problem, to write the test statistic as

$$\Lambda(X_{11}, X_{12}, X_{21}, X_{22}) = 2 \log \frac{L(\hat{p})}{L(\hat{p}_0)} = 2 \sum_{k=1}^{2} \sum_{\ell=1}^{2} X_{k\ell} \log \frac{\hat{p}_{k\ell}}{\hat{p}_{0,k\ell}}$$

Substituting the maximum likelihood estimates of $\hat{p}$ and $\hat{p}_0$, we have

$$\Lambda(X_{11}, X_{12}, X_{21}, X_{22}) = 2 \sum_{k=1}^{2} \sum_{\ell=1}^{2} X_{k\ell} \log \frac{X_{k\ell}/n}{(X_{k1} + X_{k2})(X_{1\ell} + X_{2\ell})/n^2}.$$

As in the case of the goodness of fit test, it is conventional to write this statistic as

$$\Lambda(X_{11}, X_{12}, X_{21}, X_{22}) = 2 \sum_{k=1}^{2} \sum_{\ell=1}^{2} X_{k\ell} \log \frac{X_{k\ell}}{E_{k\ell}}, \tag{10.8.1}$$

where $E_{k\ell} = \frac{1}{n}(X_{k1} + X_{k2})(X_{1\ell} + X_{2\ell}) = n\hat{p}_{k\circ}\hat{p}_{\circ\ell}$ is interpreted as the "expected" value of $X_{k\ell}$ if the null hypothesis is true. The more popular Pearson's χ^2 test of independence, with test statistic

$$T(X_{11}, X_{12}, X_{21}, X_{22}) = 2 \sum_{k=1}^{2} \sum_{\ell=1}^{2} \frac{(X_{k\ell} - E_{k\ell})^2}{E_{k\ell}},$$

can be similarly viewed as an approximation of $\Lambda(X_{11}, X_{12}, X_{21}, X_{22})$. Both these tests can be easily generalized to situations where there more than two treatments or more than two outcomes.

Unlike the goodness of fit problem, the null parameter space $\mathcal{P}_0$ is no longer a singleton set. Also, unlike in the examples involving the Normal distribution, the distribution of the test statistic does not become independent of the choice of $p_0 \in \mathcal{P}_0$. This effectively rules out the simulation approach to obtain the null distribution, as there are infinitely many choices of p_0 which need to be considered if we wish to compute the p-value using (10.5.3).

Fortunately, Theorem 10.7.2 is still applicable in this case. As the number of independent parameters is 3 in $\mathcal{P}$ and 2 in $\mathcal{P}_0$, both the likelihood ratio statistic and Pearson's test statistic have the χ^2_1 distribution asymptotically. In general, the degrees of freedom of the asymptotic null distribution will depend on the number of treatments and outcomes.

10.8.2 *Binomial Model with Fixed Row Margins*

A common strategy when designing clinical trials such as the one in Example 10.3.1 is to fix the number of individuals in each treatment group in advance. In other words, the numbers n_1 and n_2 to be given treatment 1 and treatment 2 respectively are fixed in advance.

This still leaves open the question of *how* to choose the n_1 individuals to be given treatment 1. A natural choice is to choose them randomly, uniformly from the $\binom{n}{n_1}$ possibilities. Such an allocation scheme forms the basis of the approach described in Section 10.8.3.

Here, we consider the model where the treatment attribute is not random at all, but rather fixed in advance. One may think of this as comparing two different populations, based on samples of size n_1 and n_2. In this setup, the null hypothesis of independence of treatment and outcome can be reinterpreted to mean that the *distribution* of the outcome attribute does not depend on the population from which the individual comes. It is similar in that sense to the two-sample test for equality of population means discussed in Section 10.6.5. Alternatively, this model can be thought to have been derived from the multinomial model by conditioning on the treatment attribute of *all* individuals. In terms of Lemma 10.3.2, we have

$$\begin{aligned} Y_i = 1 \quad | \quad T_i = 1 &\sim \text{Bernoulli}(q_{11}), \\ Y_i = 1 \quad | \quad T_i = 2 &\sim \text{Bernoulli}(q_{21}). \end{aligned}$$

By exchangeability of individuals within each treatment group, this is equivalent to conditioning on the *totals* n_1 and n_2 in each treatment group. This interpretation allows us to use this (conditional) model wherever the multinomial model is appropriate. In this conditional model, the parameter π_1 is not required, so the parameter space reduces to $\mathcal{P} = \{p = (q_{11}, q_{21}) : 0 \leq q_{11}, q_{21} \leq 1\}$, and and the null hypothesis is represented by the subset $\mathcal{P}_0 = \{p \in \mathcal{P} : q_{11} = q_{21}\}$.

The likelihood function for this model is simply the product of two Binomial likelihoods, and the maximum likelihood estimators of p are suitable sample proportions. Specifically, the unconstrained estimators are given by $\hat{q}_{k1} = X_{k1}/n_k$ for $k = 1, 2$, and the estimators under the null hypothesis are $\hat{q}_{0,11} = \hat{q}_{0,21} = (X_{11} + X_{21})/n$. It is left as an exercise to verify that the likelihood ratio statistic

is exactly the same as (10.8.1), the statistic obtained in the multinomial model. Although both $\mathcal{P}$ and $\mathcal{P}_0$ have one fewer independent parameter in this case, $\mathcal{P}_0$ is still not a singleton set, so the null distribution will depend on the unknown common value of $q_{11} = q_{21}$. Again, the asymptotic distribution of the test statistic is χ^2_1.

10.8.3 *Fisher's Exact Test of Independence*

The tests of independence derived above are approximate tests, valid asymptotically. An interesting and somewhat natural extension of this setup *does* allow us to obtain an exact test, although its formulation is such that it does not fit nicely into our usual parametric setup. Nonetheless, we end this section with a discussion of this formulation because it provides a useful perspective on the testing problem in general. The resulting test is known as Fisher's exact test of independence.

Technically, the setup of Fisher's test can be obtained from the multinomial model by conditioning on a certain event. In this sense, it is an extension of the previous two models. Specifically, the multinomial model in Section 10.3.1 defines a probability distribution on the set of all 2×2 tables of the form

	Outcome 1	Outcome 2	Total
Treatment 1	X_{11}	X_{12}	$N_{1\circ}$
Treatment 2	X_{21}	X_{22}	$N_{2\circ}$
Total	$N_{\circ 1}$	$N_{\circ 2}$	n

where $X_{k\ell}, k, \ell = 1, 2$ are non-negative integer counts, $N_{k\circ}$ and $N_{\circ\ell}$ are row and column totals, and $\sum_{k,\ell} X_{k\ell} = \sum_k N_{k\circ} = \sum_\ell N_{\circ\ell} = n$, the total number of participants. The row and column totals $N_{k\circ}$ and $N_{\circ\ell}$ are random, although the total sum n is fixed as the total number of units (the sample size) is fixed in advance. The model considered in Section 10.8.2, as we have noted above, can be derived from the multinomial model by conditioning on the treatment attribute of each participant, or equivalently by conditioning on $N_{1\circ}$ and $N_{2\circ}$. The interpretation of this conditioning from the perspective of the random experiment is that the number of individuals in each treatment group (for instance, the number of individuals given the placebo and the vaccine in Example 10.3.1) is fixed in advance. To derive Fisher's exact test, we further condition on the column totals $N_{\circ 1}$ and $N_{\circ 2}$.

If both row and column totals are fixed in advance, then it is immediate that any one element of the table determines the others. This solves one of our problems, namely, finding a test statistic: without loss of generality, we can take X_{11} to be our test statistic, as it completely defines the entire table. In the multinomial setup, the conditional distribution of X_{11} given the row and column totals turns out to be the familiar hypergeometric distribution (Example 2.3.1).

It is not, however, immediately obvious how the conditioning on column totals can be interpreted from the perspective of the underlying random experiment. Fixing the total number of individuals to be assigned treatment 1 and treatment 2 in advance is reasonable. However, it is completely unreasonable to expect that the number of individuals who have outcome 1 and outcome 2 would also be fixed in advance. To link the conditional model above to a reasonable experimental setup, we have to view the experiment from a different, nonparametric perspective.

Recall that there are four possible treatment-outcome combinations for each individual or unit, namely $(1,1), (1,2), (2,1)$, and $(2,2)$, identified respectively with the four matrices

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \text{ and } \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

As in Lemma 10.3.2, let (T_i, Y_i) denote the treatment and outcome pair for the i-th individual. The treatment attribute T_i for each unit is usually random, but the nature of the randomness depends on the type of experiment. For observational studies, individuals are sampled from a population and both "treatment" and "outcome" attributes are observed; neither are under the control of the experimenter. However, for controlled trials such as the vaccine trial described in Example 10.3.1, the treatment is assigned randomly as part of the experiment, typically by ensuring that all participants are equally likely to get a particular treatment. The outcome Y_i is also random, presumably in a way that depends on the individual involved. In the vaccine example, the outcome may depend on age, gender, and other attributes of a participant that are not available to us. The outcome is also possibly affected by the treatment assigned; in fact, in the vaccine trial, we hope that being vaccinated makes an individual less likely to become affected with the disease.

However, we are interested in testing the hypothesis of *independence* of the two attributes, which posits that the outcome is *not affected* by the treatment, even if it does depend on the individual. How can we formulate this idea as a probability model? The multinomial model with fixed row totals described in Section 10.8.2 assumes a parametric model where the distribution of $Y_i \mid T_i = t$ depends only on the parameter q_{11} or q_{21}, depending on whether $t = 1$ or $t = 2$. Fisher's exact test uses a different probability formulation which in principle allows the distribution of $Y_i \mid T_i = t$ to vary from individual to individual, without explicitly providing a parametric model. We describe this formulation next.

Normally, probability statements are interpreted in terms of the outcome in repeated performances of an experiment. Here, repeating the experiment may mean selecting a new sample of units (participants) via some random sampling mechanism, randomly assigning them treatments, and observing the outcomes. However, let us suppose that we simplify the process of repeating the experiment by skipping the first step: instead of selecting a new set of units on which to perform the experiment, suppose we use the *same* set of n participants. However, we do still randomize the allocation of treatments, by randomly selecting a new subset of size n_1 uniformly from the $\binom{n}{n_1}$ possibilities, to receive treatment 1. Here n_1, the number of participants who get treatment 1, is assumed to be fixed in advance, and thus n_1 can be viewed as the first row total $N_{1\circ}$ in the multinomial formulation. Of course, we cannot actually perform this experiment and observe the outcomes on the same units again (after all, someone who has been vaccinated cannot be un-vaccinated), but we *can* conjecture about what could have happened if the experiment had been performed with this new treatment assignment.

In general, we cannot say what the outcome would have been, as they could have changed if different treatments had been received. However, suppose we restrict ourselves to the case when the outcome is *independent* of the treatment, which is what the position of the skeptic would be. Interpreting the notion of independence literally rather than probabilistically, we can then say

that for each unit, the outcome depends only on the individual and not the treatment, so would have remained the same as the outcome observed with the original assignment. In other words, for the i-th unit if we had $Y_i = 1$ in the original experiment, we would again have $Y_i = 1$ in the new hypothetical experiment, regardless of whether the value of T_i had changed. For each such hypothetical experiment then, we can recreate the summary 2×2 table without requiring any new information other than the new treatment assignments. An important point, which is easy to see, is that the row and column totals of these tables remain unchanged from the original by construction. The row totals are the number of units assigned treatments 1 and 2, which are always n_1 and $n - n_1$. The column totals are the total number of units with outcome 1 and 2, which also remain unchanged, *provided* that outcomes are not affected by treatment assignment.

The argument above gives us a very concrete (if still somewhat abstract) procedure to test the null hypothesis (the skeptic's conjecture) that treatment does not affect outcome: If this conjecture were true, then we can randomly choose a hypothetical treatment assignment to get a random summary table, forming a probability distribution whose sample space consists of all 2×2 summary tables with the given marginal row and column totals, or equivalently, just its first entry X_{11}. Even if we could not say anything more about this distribution, we could always simulate as many such tables as we wanted to get an empirical distribution of X_{11}. As it happens, we *can* actually say more, because the distribution of X_{11} again turns out to be the same hypergeometric distribution.

Finally, to decide whether or not to reject the null hypothesis, we use the same approach as earlier. Depending on the problem at hand, departure from the null may be indicated either by high values of X_{11}, or low values of X_{11}, or both. Accordingly, the test is one-sided or two-sided. For one-sided tests, the p-value is given by the corresponding tail probability of the null distribution starting from the observed value of X_{11}. For two-sided tests, the computation is less obvious as the null distribution is not symmetric. Following the principle described in Example 10.6.1, the p-value in this case is obtained by adding up the individual probabilities of all outcomes in the null distribution that are *at most as likely* as the observed value of X_{11}.

A

SOME MATHEMATICAL DETAILS

A.1 TRANSFORMATION OF CONTINUOUS RANDOM VARIABLES- JACOBIAN

Suppose the random variable $X : S \to \mathbb{R}$ is a continuous random variable, with probability density function $f_X : \mathbb{R} \to \mathbb{R}$. Let $g : \mathbb{R} \to \mathbb{R}$ and $Y = g(X)$. In general it may be hard to find the distribution of Y. For some specific class of g the random variable Y will also be a continous random variable and we can calculate its probability density function. We recall the method discussed in Section 5.3. One immediately observes that the distribution function of Y is given by

$$P(Y \leq y) = P(g(X) \in (-\infty, y]) = P(X \in g^{-1}(-\infty, y]).$$

Thus the above formula provides a theoretical expression for the distribution function of Y provided for all y the function g is such that $g^{-1}(-\infty, y]$ is an event. Now, let us assume that g is strictly increasing and differentiable function with g' being continuous and $g'(x) > 0$ for all $x \in \mathbb{R}$. This implies that $g^{-1} : \mathbb{R} \to \mathbb{R}$ exists and is differentiable. The distribution function of Y is given by

$$P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = \int_{-\infty}^{g^{-1}(y)} f_X(x)dx.$$

From the above, using the fundamental theorem of calculus, we see that Y has a probability density function $f_Y : \mathbb{R} \to$ given by

$$f_Y(y) = \frac{dg^{-1}}{dy}(y) f_X(g^{-1}(y)), \tag{A.1.1}$$

for all $y \in \mathbb{R}$. Now, let us assume that g is strictly decreasing and differentiable function with g' being continuous and $g'(x) < 0$ for all $x \in \mathbb{R}$. This implies that $g^{-1} : \mathbb{R} \to \mathbb{R}$ exists and is differentiable. The distribution function of Y is given by

$$P(Y \leq y) = P(g(X) \leq y) = P(X \geq g^{-1}(y)) = \int_{g^{-1}(y)}^{\infty} f_X(x)dx.$$

From the above, using the fundamental theorem of calculus, we see that Y has a probability density function $f_Y : \mathbb{R} \to$ given by

$$f_Y(y) = -\frac{dg^{-1}}{dy}(y) f_X(g^{-1}(y)), \tag{A.1.2}$$

for all $y \in \mathbb{R}$. We now present the above deductions as a theorem below.

Theorem A.1.1. *Let $X : S \to \mathbb{R}$ be a continuous random variable with probability density function $f_X(\cdot)$ with $I = \{x \in \mathbb{R} : f_X(x) > 0\}$. Let $g : I \to \mathbb{R}$ be a differentiable function on I such that $g' : I \to \mathbb{R}$ is continuous with $g'(\cdot) \neq 0$ on I. Then the random variable $Y = g(X)$[a] has a density $f_Y : \mathbb{R} \to \mathbb{R}$ given by*

$$f_Y(y) = \begin{cases} |\, \frac{dg^{-1}}{dy}(y) \,|\, f_X(g^{-1}(y)), & y \in Range(g) \\ \\ 0 & otherwise \end{cases} \tag{A.1.3}$$

or equivalently

$$\begin{cases} f_Y(y) = \frac{1}{|g'(x)|} f_X(x), & with\ y = g(x), x \in I \\ \\ 0 & otherwise. \end{cases} \tag{A.1.4}$$

[a] We can assume without loss of generality that $X(s) \in I$ for all $s \in S$ so that $Y(s) = g(X(s))$ is well defined for all $s \in S$.

Example A.1.2. Let $X \sim$ Uniform $(-1, 1)$. Let $g : (-1, 1) \to \mathbb{R}$ given by $g(x) = x^2$ and $Y = g(X)$. Observe that g is differentiable on $(-1, 1)$ with $g' : (-1, 1) \to \mathbb{R}$ given $g'(x) = 2x$. Again $g'(0) = 0$ so Theorem A.1.3 is not applicable. As we have seen before we can calculate the probability density function of Y. We first calculate the distribution function of Y.

$$P(Y \leq y) = \begin{cases} 0 & y \leq 0 \\ 1 & y \geq 1 \end{cases}$$

For $0 < y < 1$, using the probability density function of X we have

$$P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{2} dz = y^{\frac{1}{2}}.$$

We note that the distribution function of Y is piecewise differentiable and hence Y has a probability density function given by

$$f_Y(y) = \begin{cases} \frac{1}{2} y^{-\frac{1}{2}} & y \in (0, 1), \\ 0 & \text{otherwise} \end{cases}$$

■

In the above example the transformation g was not one-one and hence was not invertible. We note that the function g has a well defined inverse in the interval $(-1, 0)$ and $(0, 1)$. Intuitively one should be able to apply Theorexm A.1.3 on each of these intervals. The next theorem formalises this.

Theorem A.1.3. *Let $X : S \to \mathbb{R}$ be a continuous random variable with probability density function $f_X(\cdot)$ with $I = \{x \in \mathbb{R} : f_X(x) > 0\}$. Let $g : I \to \mathbb{R}$ be a differentiable function on I such that*

(a) $Z = \{x \in I : g'(x) = 0\}$ is finite.

(b) $g' : I \setminus Z \to \mathbb{R}$ is continuous with $g'(\cdot) \neq 0$ on $I \setminus Z$.

Let $B = \{g(x) : x \in I \setminus Z\}$. Then $Z_y = \{x \in I \setminus Z : g(x) = y\}$ is necessarily a finite non-empty set for $y \in B$ and the random variable $Y = g(X)$ has a density $f_Y : \mathbb{R} \to \mathbb{R}$ given by

$$\begin{cases} f_Y(y) = \sum_{x_i \in Z_y} \frac{f_X(x_i)}{|g'(x_i)|}, & \text{with } y \in B \\ \\ 0 & \text{otherwise.} \end{cases} \tag{A.1.5}$$

A.1.1 *Multiple Continuous Random Variables and the Jacobian*

An alternate way of viewing the change of distribution formula in the previous subsection is as the familiar u-substitution from calculus. If $g(x)$ is a differentiable function with differentiable inverse and $Y = g(X)$ then using the subsitution $y = g(x)$ (so $x = g^{-1}(y)$) we have

$$\begin{aligned} \int_A f_Y(y)\,dy &= P(Y \in A) \\ &= P(g(X) \in A) \\ &= P(X \in g^{-1}(A)) \\ &= \int_{g^{-1}(A)} f_X(x)\,dx \\ &= \int_A f_X(g^{-1}(y)) \frac{dg^{-1}}{dy}(y)\,dy \end{aligned}$$

But if $\int_A f_Y(y)\,dy = \int_A f_X(g^{-1}(y)) \frac{dg^{-1}}{dy}\,dy$ for every event A, then the integrands must be the same and $f_Y(y) = f_X(g^{-1}(y)) \frac{dg^{-1}}{dy}(y)$.

When multiple random variables are involved a similar formula may be derived using the multivariate change of variables formula invovling the Jacobian. Recall the following result from multivariate calculus:

Theorem A.1.4. *Let $S, T \subset \mathbb{R}^n$ and let $h : S \to T$ be an invertible function*

$$h(y_1, y_2, \ldots, y_n) = (x_1, x_2, \ldots, x_n)$$

Suppose the Jacobian $J = \frac{\partial(x_1, x_2, \ldots, x_n)}{\partial(y_1, y_2, \ldots, y_n)}$ exists and is never zero. Then

$$\begin{aligned} &\int \int \cdots \int_{h(A)} f(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \ldots dx_n \\ &= \int \int \cdots \int_{A} f(x_1, x_2, \ldots, x_n) \mid J(y_1, y_2, \ldots, y_n) \mid \, dy_1 dy_2 \ldots dy_n \end{aligned}$$

where in the final line it is understood that the x_j variables have been written in terms of $y_1, y_2, \ldots, y_n$ via the h^{-1} function.

Now let $n \geq 1$ and suppose $X_1, X_2, \ldots, X_n$ are random variables with a joint density $f_{\mathbf{X}}(x_1, x_2, \ldots, x_n)$. Let $h : S \to T$ be as in the theorem above and define an $\mathbb{R}^n$-valued random vector

$$(Y_1, Y_2, \ldots, Y_n) = h(X_1, X_2, \ldots, X_n).$$

Let A be an event in $\mathbb{R}^n$. Then,

$$\begin{aligned} P((Y_1, Y_2, \ldots, Y_n) \in A) &= P(h(X_1, X_2, \ldots, X_n) \in A) \\ &= P((X_1, X_2, \ldots, X_n) \in h^{-1}(A)) \\ &= \int \int \cdots \int_{h^{-1}(A)} f_{\mathbf{X}}(x_1, x_2, \ldots, x_n)\, dx_1 dx_2, \ldots dx_n \\ &= \int \int \cdots \int_{A} f_{\mathbf{X}}(x_1, x_2, \ldots, x_n) \mid J(y_1, y_2, \ldots, y_n)\, dy_1 dy_2, \ldots dy_n \end{aligned}$$

where, as in the prior theorem, the x_j variables are understood to have been written in terms of $y_1, y_2, \ldots, y_n$.

Let $f_{\mathbf{Y}}(y_1, y_2, \ldots, y_n)$ represent the joint density for the $(Y_1, Y_2, \ldots, Y_n)$. Since that density is defined by the equation

$$P((Y_1, Y_2, \ldots, Y_n) \in A) = \int \int \cdots \int_{A} f_{\mathbf{Y}}(y_1, y_2, \ldots, y_n)\, dy_1 dy_2, \ldots dy_n$$

it must be that $f_{\mathbf{Y}}(y_1, y_2, \ldots, y_n) = f_{\mathbf{X}}(x_1, x_2, \ldots, x_n) \mid J(y_1, y_2, \ldots, y_n)$ which provides an equation relating the joint density of the $(X_1, X_2, \ldots, X_n)$ with the joint density of $(Y_1, Y_2, \ldots, Y_n)$.

Example A.1.5. Let $X_1, X_2 \sim \text{Uniform}((0, 1))$ be independent random variables. Let $Y_1 = X_1 + X_2$ and let $Y_2 = X_1 - X_2$. This is a globally invertable linear transformation for which $X_1 = \frac{1}{2}Y_1 + \frac{1}{2}Y_2$ and $X_2 = \frac{1}{2}Y_1 - \frac{1}{2}Y_2$. The Jacobian is the determinant

$$J = \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)} = det \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{bmatrix} = -\frac{1}{2}.$$

Since the joint density for the X variables is

$$f_{\mathbf{X}}(x_1, x_2) = \begin{cases} 1 & \text{if } 0 < x_1 < 1 \text{ and } 0 < x_2 < 1 \\ 0 & \text{otherwise} \end{cases}$$

then the joint density for the Y variables is

$$f_{\mathbf{Y}}(y_1, y_2) = f_{\mathbf{X}}(x_1, x_2)|J(y_1, y_2)| = \begin{cases} 1/2 & \text{if } 0 < y_1 + y_2 < 2 \text{ and } 0 < y_1 - y_2 < 2 \\ 0 & \text{otherwise} \end{cases}$$

where the region in which the density is non-zero is the square with corners $(0,0), (1,1), (2,0)$, and $(1,-1)$ in the (y_1, y_2)-plane. In particular we could use this joint density to provide another derivation of the density for the sum of two independent Uniform$(0,1)$ random variables. That is simply the marginal distribution of Y_1 alone which we can now calculatate.

When $0 < y_1 < 1$ we have

$$\begin{aligned} f_{Y_1}(y_1) &= \int_{-\infty}^{\infty} f_{\mathbf{Y}}(y_1, y_2)\, dy_2 \\ &= \int_{-y_1}^{y_1} \frac{1}{2}\, dy_2 \\ &= y_1 \end{aligned}$$

while when $1 < y_1 < 2$ we have

$$\begin{aligned} f_{Y_1}(y_1) &= \int_{-\infty}^{\infty} f_{\mathbf{Y}}(y_1, y_2)\, dy_2 \\ &= \int_{y_1-2}^{2-y_1} \frac{1}{2}\, dy_2 \\ &= 2 - y_1 \end{aligned}$$

In other words, we have the familiar result

$$f_{Y_1}(y_1) = \begin{cases} y_1 & \text{if } 0 < y_1 < 1 \\ 2 - y_1 & \text{if } 1 < y_1 < 2 \\ 0 & \text{otherwise} \end{cases}$$

which was proven in Example 5.5.2 using a different technique. ■

The change of variables formula requries that the number of X-variables be equal to the number of Y-variables in order for the Jacobian determinant to exist. For some problems this may

not originally be the case, but this problem can often be alleviated by inserting extra variables. For instance, consider Exercise 5.5.4 from Chapter 5. In that problem (X_1, X_2, X_3) are given as independent and uniformly distributed on $(0, 1)$. The problem asks for the value of $P(X_1X_3 < X_2^2)$. While that problem can be solved using the techniques from that chapter, it can also be solved using the Jacobian method.

Example A.1.6. Let $Y_1 = X_1X_3$ and let $Y_2 = X_2^2$. Note that these are the quantities of interest in the probability we are asked to compute. To use the Jacobian technique we will also introduce $Y_3 = X_1$ simply to maintain an equal number of variables. On the region $X_1, X_2, X_3 \in (0, 1)$ where the density is non-zero, this transformation is invertible. Solving for the X-variables gives: $X_1 = Y_3$, $X_2 = \sqrt{Y_2}$, and $X_3 = \frac{Y_1}{Y_3}$. Therefore,

$$J = \frac{\partial(x_1, x_2, x_3)}{\partial(y_1, y_2, y_3)} = det \begin{bmatrix} 0 & 0 & 1 \\ 0 & \frac{1}{2\sqrt{y_2}} & 0 \\ \frac{1}{y_3} & 0 & -\frac{y_1}{y_3^2} \end{bmatrix} = -\frac{1}{2\sqrt{y_2}y_3}.$$

Since the joint density for the X-variables is $f_{\mathbf{X}}(x_1, x_2, x_3) = 1$ whenever $0 < x_1, x_2, x_3 < 1$ (and 0 otherwise) we have

$$f_{\mathbf{Y}}(y_1, y_2, y_3) = f_{\mathbf{X}}(x_1, x_2, x_3)|J(y_1, y_2, y_3)| = \begin{cases} \frac{1}{2\sqrt{y_2}y_3} & \text{if } 0 < y_1 < 1 \text{ and } 0 < \sqrt{y_2} < 1 \text{ and } 0 < \frac{y_1}{y_3} < 1 \\ 0 & \text{otherwise} \end{cases}$$

which may be written more compactly as

$$f_{\mathbf{Y}}(y_1, y_2, y_3) = \begin{cases} \frac{1}{2\sqrt{y_2}y_3} & \text{if } 0 < y_1 < y_3 < 1 \text{ and } 0 < y_2 < 1 \\ 0 & \text{otherwise} \end{cases}$$

At that point we may calculate the desired probability using the new joint density.

$$\begin{aligned} P(X_1X_3 < X_2^2) &= P(Y_1 < Y_2) \\ &= \int\int\int_{\{y_1<y_2\}} f_{\mathbf{Y}}(y_1, y_2, y_3)\, dy_1\, dy_2\, dy_3 \\ &= \int_0^1 \int_0^{y_3} \int_{y_1}^1 \frac{1}{2\sqrt{y_2}y_3}\, dy_2\, dy_1\, dy_3 \\ &= \int_0^1 \int_0^{y_3} \frac{1}{y_3} - \frac{\sqrt{y_1}}{y_3}\, dy_1\, dy_3 \\ &= \int_0^1 1 - \frac{2}{3}\sqrt{y_3}\, dy_3 \\ &= \frac{5}{9} \end{aligned}$$

■

A.2 STRONG LAW OF LARGE NUMBERS

In this section we shall state and prove the strong law of large numbers.

Theorem A.2.1. (Strong Law of Large Numbers) *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables. Assume that X_1 has finite mean μ and finite variance σ^2. Let $A = \{\lim_{n\to\infty} \overline{X}_n = \mu\}$. Then*

$$P(A) = 1. \tag{A.2.1}$$

As remarked in Chapter 8, the above results states that the convergence of sample mean to μ actually happens with Probability one. This mode of convergence of the sample mean to the true mean is called "convergence with probability 1." We define it precisely below.

Definition A.2.2. *A sequence $X_1, X_2, \ldots$ is said to converge with probability one to a random variable X if $A = \{\lim_{n\to\infty} \overline{X}_n = X\}$.*

$$P(A) = 1. \tag{A.2.2}$$

The following notation

$$X_n \xrightarrow{w.p.1} X$$

is typically used to convey that the sequence $X_1, X_2, \ldots$ converges with probability one to X.

As alluded earlier that this is a stronger mode of convergence. We prove it in the next proposition.

Proposition A.2.3. *Let $X_1, X_2, \ldots$ be a sequence of random variables on a sample space S. Suppose X_n converges to a random variable X with probability 1 then X_n converges to a random variable X in probability.*

Proof- Let $\epsilon > 0$ and $\delta > 0$ be given. We need to show $\exists N$ such that

$$P(|X_m - X| > \epsilon) < \delta, \forall m \geq N. \tag{A.2.3}$$

Let $A = \{\omega \in S : \lim_{n\to\infty} X_n(\omega) = X\}$. We are given that

$$P(A) = 1. \tag{A.2.4}$$

Suppose we denote, for $\eta > 0$ and $n \geq 1$,

$$A_n^\eta = \{\omega \in S : |X_n(\omega) - X(\omega)| \leq \epsilon\}.$$

then

$$A = \cap_{\eta>0} \cup_{k=1}^{\infty} \cap_{n=k}^{\infty} A_n^\eta.$$

This can be verified using the fact that $\omega \in A$ if and only if for all $\eta > 0$, there is a $k \equiv k(\omega)$ such that

$$|X_n(\omega) - X(\omega)| \leq \epsilon, \ \forall\, n \geq k.$$

For $m \geq 1$, define $B_m^\epsilon = \cap_{n=m}^\infty A_n^\epsilon$. Note

$$B_m^\epsilon \subset B_{m+1}^\epsilon, \tag{A.2.5}$$

for all $m \geq 1$. So by Exercise 1.1.13, we have

$$\lim_{m\to\infty} P(B_m^\epsilon) \uparrow P(\cup_{m=1}^\infty B_m^\epsilon). \tag{A.2.6}$$

As $A \subset \cup_{m=1}^\infty B_m^\epsilon$, using (A.2.4) we have $1 = P(A) \leq P(\cup_{m=1}^\infty B_m^\epsilon) \leq 1$. So

$$P(\cup_{m=1}^\infty B_m^\epsilon) = 1. \tag{A.2.7}$$

By (A.2.6) and (A.2.7) $\exists N$ such that

$$P(B_m^{\epsilon_0}) > 1 - \delta, \forall m \geq N.$$

As $B_m^\epsilon \subset A_m^\epsilon$,

$$P(A_m^\epsilon) > 1 - \delta, \forall m \geq N.$$

Therefore by considering the complement of A_m^ϵ we obtain (A.2.3). ■

We will need a technical Lemma regarding convergence in probabilty which we state and prove below.

Lemma A.2.4. *Suppose a sequence of random variables X_n is such that*

$$X_n \xrightarrow{p} X \text{ and } X_n \xrightarrow{p} Y$$

for some random variables X, Y then $P(X = Y) = 1$.

Proof- Let $k \geq 1$. Let $A_k = \{|\, X - Y \,| \geq \frac{1}{k}\}$. Notice that $A_k \subset A_{k+1}$ and $\cup_{k=1}^\infty A_k = \{X \neq Y\}$. Let $k \geq 1, \delta > 0$ be given. As $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{p} Y$, (applying Definition 8.2.6 with $\epsilon = \frac{1}{2k}$), there exists N such that for all $n \geq N$

$$0 \ \leq \ P\left(|\, X_n - X \,| > \frac{1}{2k}\right) \ < \ \frac{\delta}{2} \qquad \text{and} \qquad 0 \ \leq \ P\left(|\, X_n - Y \,| > \frac{1}{2k}\right) \ < \ \frac{\delta}{2}. \tag{A.2.8}$$

Using the triangle inequality we observe that $|\, X - Y \,| \leq |\, X - X_n \,| + |\, X_n - Y \,|$ for all $n \geq 1$. So,

$$A_k \subset \{|\, X_n - X \,| > \frac{1}{2k}\} \cup \{|\, X_n - X \,| > \frac{1}{2k}\} \tag{A.2.9}$$

for all $n \geq 1$. Combining (A.2.8) and (A.2.9) we have (using any $n \geq N$)

$$0 \ \leq \ P(A_k) \ \leq \ P\left(|\, X_n - X \,| > \frac{1}{2k}\right) + P\left(|\, X_n - Y \,| > \frac{1}{2k}\right) \ \leq \ \frac{\delta}{2} + \frac{\delta}{2} \ = \ \delta.$$

As $\delta > 0$ was arbitrary we have $P(A_k) = 0$. Further by Exercise 1.1.13,

$$P(X \neq Y) = \lim_{k\to\infty} P(A_k) = 0.$$

Hence $P(X = Y) = 1$. ■

Proof of Theorem A.2.1**(Special Case)**- We provide a complete proof of Theorem A.2.1 in the special case when the random variables are i.i.d Bernoulli (p) random variables. We will proceed in two steps.

Step 1: $\overline{X}_n$ converges with probability one to a random variable X.

Let $\overline{S} = \limsup_{n\to\infty} \overline{X}_n$ and $\underline{S} = \liminf_{n\to\infty} \overline{X}_n$. Clearly,

$$0 \leq \underline{S} \leq \overline{S} \leq 1.$$

Fix $\epsilon > 0$, then for every k define

$$N_k = \inf\{n \in \mathbb{N} : \frac{X_k + X_{k+1} \ldots + X_{k+n-1}}{n} \geq \overline{S} - \epsilon\}.$$

The random variable N_k, in some sense, measures how close we are to $\overline{S}$ and our main effort will be to control the size N_k. It is easy to see that N_k is finite a.e. and are all identically distributed (because of independence of X_i). Hence we can choose an m such that $P(N_k > m) < \epsilon$ for all k. Define random variables Y_k and N_k^Y by the following mechanism:

$$Y_k = \begin{cases} X_k & \text{if} \quad N_k \leq m \\ 1 & \text{if} \quad N_k > m \end{cases} \tag{A.2.10}$$

$$N_k^Y = \inf\{n \in \mathbb{N} : \frac{Y_k + Y_{k+1} \ldots + Y_{k+n-1}}{n} \geq \overline{S} - \epsilon\}. \tag{A.2.11}$$

Clearly $N_k^Y \leq N_k$ and if k is such that $N_k \geq m$ then $N_k^Y = 1$ (since setting $Y_k = 1$ ensures that we are above $\overline{S} - \epsilon$ immediately). So we have

$$N_k^Y \leq m. \quad a.e.$$

So for large enough $n \in \mathbb{N}$ we can break up $\sum_{k=1}^n Y_k$ into pieces of lengths atmost M such that the average over each piece is atleast $\overline{S} - \epsilon$. Then finally stop at the n-th term. Then it is clear that,

$$\sum_{k=1}^{n} Y_k \geq (n - m)(\overline{S} - \epsilon). \tag{A.2.12}$$

By our choice of m

$$E(Y_k) = E(X_k 1(N_k \leq m)) + P(N_k > m) < E(X_k) + \epsilon = E(X) + \epsilon,$$

for any k. Take expectations in (A.2.12) and use the above inequality to obtain

$$n(E(X)+\epsilon) \geq (n-m)(E(\overline{S})-\epsilon).$$

Divide by n and first let $n \to \infty$ followed by $\epsilon \to 0$, to get

$$E(\overline{S}) \leq E(X). \tag{A.2.13}$$

Let $\widetilde{X_k} = 1 - X_k$. Applying the above argument to $\widetilde{X}$ (verify this) we have

$$E(\overline{\widetilde{S}}) \leq E(\widetilde{X}).$$

Since $\underline{S} = -\overline{\widetilde{S}}$ this implies

$$E(\underline{S}) \geq E(X). \tag{A.2.14}$$

Now, $\underline{S} \leq \overline{S}$ a.e. So only way (A.2.14) and (A.2.13) can hold only if $\underline{S} = \overline{S} a.e.$ Therefore $\lim_{n\to\infty} \overline{X}_n$ exists almost everywhere and let us call it X. This completes step 1.

Step 2: We shall now use the Weak Law of Large numbers (Theorem 8.2.1), along with Proposition A.2.3, and Lemma A.2.4 to complete the proof. The weak law implies that

$$\overline{X}_n \xrightarrow{p} \mu \text{ as } n \to \infty.$$

From Step 1, we know that

$$\overline{X}_n \xrightarrow{w.p.1} X \text{ as } n \to \infty.$$

Proposition A.2.3 then implies that

$$\overline{X}_n \xrightarrow{p} X \text{ as } n \to \infty.$$

Finally Lemma A.2.4 implies $P(X = \mu) = 1$. Therefore

$$\overline{X}_n \xrightarrow{w.p.1} \mu \text{ as } n \to \infty.$$

■

Proof of Theorem A.2.1**(General Case)** The essence of the proof is contained in the special case proven above. We provide a sketch of the proof.

Case 1:($0 \leq X \leq 1$) An imitation of Step 1 of the proof for Bernoulli p random variables will show that there is a limit. Step 2 of the above proof follows readily.

Case 2: Bounded Case When the random variable X is bounded, i.e. $\mid X \mid \leq M$ for some $M > 0$. One can consider $Y = \frac{X-M}{2M}$ and $Y_i = \frac{X_i - M}{2M}$. As $0 \leq Y \leq 1$ then one can use Case 1 for Y_i to establish that there is a limit. Step 2 of the above proof follows readily.

Case 3: (General Case by Truncation) One fixes $\alpha, \beta > 0$ and defines

$$\overline{S}_{(\alpha)} = \min\{\overline{S}, \alpha\}, X^{(\beta)} = \max\{X, -\beta\} \text{ and } X_k^{(\beta)} = \max\{X_k, -\beta\}\ \forall k \in \mathbb{N}.$$

The above quantities are all bounded. One imitates Step 1 of the above proof and this will result in inequalities depending on α, β. One then allows α, β approach infinity to establish that $\overline{X}_n \xrightarrow{w.p.1} X$ for a random variable X. Step 2 of the above proof follows readily. We refer the reader to [AS09] for the complete proof. ■

B

TABLES

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Table B.1: Normal tables evaluating : $\frac{1}{2\pi}\int_{-\infty}^{z} e^{-\frac{x^2}{2}}dx$

BIBLIOGRAPHY

[AS09] Siva, Athreya, Sunder, V.S. *Measure and Probability* CRC Press (Outside India) 2009 ISBN 14 3980 126 6.

[CasBer90] Casella, George; Berger, Roger L.*Statistical inference.* The Wadsworth & Brooks/Cole Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1990. xviii+650 pp. ISBN: 0-534-11958-1

[Fel68] W. Feller, *An introduction to probability theory and its applications* Vol. I. Third edition John Wiley & Sons, Inc., New York-London-Sydney 1968.

[Fel71] W. Feller, *An introduction to probability theory and its applications* Vol. II. Second edition John Wiley & Sons, Inc., New York-London-Sydney 1971.

[FPP98] T.Freedman, R.Pisani, R.Purves, *Statistics* Third edition W.W.Norton and Company, Inc., New York-London, 1998.

[FG97] B. Fristedt and L. Gray, *A Modern Approach to Probability Thoery* Birkhauser, Boston 1997.

[Ghah00] S. Ghahramani *Fundamentals of probability.* Second edition Prentice Hall, New Jersey, 2000.

[HMC05] R.Hogg, J. McKean, A. Craig, *Introduction to Mathematical Statistics*, Sixth Edition, Pearson Education, Inc., New Jersey, 2005.

[HPS72] P.G. Hoel, S.C. Port, and C.J. Stone, *Introduction to Probability Theory*, Houghton Mifflin Company, 1972.

[Keane] M. Keane, *The Essence of the Law of Large Numbers*, , Pages 125–129, Algorithms, Fractals, and Dynamics, Springer USA, 1995.

[Pit89] Pitman, *Probability,* Springer-verlag, New York, 1989.

[Ram97] Ramasubramanian, *The Normal Distribution 1. From Binomial to Normal* Resonance, Vol 2, No.6, (June 1997), pp 15-14.

[Rao73] C.R.Rao, *Linear Statistical Inference and its Application* , Second Edition, John Wiley, New York, 1973.

[Ross84] S. Ross, *A first course in probability*, Second edition Macmillan Publishing Company, New York, 1984.

[Ser09] Serfling RJ., *Approximation theorems of mathematical statistics*, John Wiley & Sons; 2009 Sep 25.

[Stig84] Stephen M. Stigler, *Kruskal's Proof of the Joint Distribution of* $\overline{X}_n$ *and* s_n^2, The American Statistician, Vol 38, No. 2, (May 1984), pp 134-135.

[Wil27] Wilson E.B., *Probable inference, the law of succession, and statistical inference*, Journal of the American Statistical Association. 1927 Jun 1;22(158):209-12.

INDEX