Introduction to Probability and Statistics Using R

Second Edition

G. Jay Kerns

December 28, 2012

IPSUR: Introduction to Probability and Statistics Using R
Copyright © 2011 G. Jay Kerns ISBN: 978-0-557-24979-4

Date: December 28, 2012

Contents

Preface to the Second Edition

What is new in the Second Edition? Almost everything. I have addressed two of the goals from the first edition. I have now converted most of the plots to `ggplot2` format.

The Second Edition marks a departure from LyX to Emacs Org-Mode. I went with Org-Mode for many reasons. I liked LyX, and LyX is definitely a more user-friendly approach to writing a free book. My workflow, however, has radically changed over the last two years, and I've converted to using Org-Mode for (almost) everything. It truly is "Your Life in Plain Text".

An advantage of the Org-Mode approach is that I can generate an HTML version (that even looks good, still) with a few keystrokes. That means I can post an HTML version of IPSUR, which I've done.

The HTML version is **very** important, for **more** than the following reasons: 1) a person can read IPSUR without need to do anything else, period, 2) automatic full-text indexing by Google, Bing, YaCY, etc., and, most importantly to me, 3) **automatic translation to over 40 languages at the click of a button** (with Google Translate, which comes for free with Google Chrome/Chromium).

Acknowledgements

The success of the Second Edition (if any) would be due in no small part to the successes of the First Edition, so it would be apropos to copy-paste the acknowledgements from the earlier Preface here.

I think, though, that the *failures* of the First Edition have played an important role as well. I would like to extend gracious thanks to Mr. P.J.C. Dwarshuis (Hans), Statistician, from The Hague, Netherlands, and Jésus Juan, who, armed with a sharp eye, have pointed out mistakes, misstatements, and places where better discussion is warranted. It is the selfless contributions of people just like these gentlemen which make the hours spent polishing a FREE book all the more worthwhile.

Preface to the First Edition

This book was expanded from lecture materials I use in a one semester upper-division undergraduate course entitled *Probability and Statistics* at Youngstown State University. Those lecture materials, in turn, were based on notes that I transcribed as a graduate student at Bowling Green State University. The course for which the materials were written is 50-50 Probability and Statistics, and the attendees include mathematics, engineering, and computer science majors (among others). The catalog prerequisites for the course are a full year of calculus.

The book can be subdivided into three basic parts. The first part includes the introductions and elementary *descriptive statistics*; I want the students to be knee-deep in data right out of the gate. The second part is the study of *probability*, which begins at the basics of sets and the equally likely model, journeys past discrete/continuous random variables, and continues through to multivariate distributions. The chapter on sampling distributions paves the way to the third part, which is *inferential statistics*. This last part includes point and interval estimation, hypothesis testing, and finishes with introductions to selected topics in applied statistics.

I usually only have time in one semester to cover a small subset of this book. I cover the material in Chapter 2 in a class period that is supplemented by a take-home assignment for the students. I spend a lot of time on Data Description, Probability, Discrete, and Continuous Distributions. I mention selected facts from Multivariate Distributions in passing, and discuss the meaty parts of Sampling Distributions before moving right along to Estimation (which is another chapter I dwell on considerably). Hypothesis Testing goes faster after all of the previous work, and by that time the end of the semester is in sight. I normally choose one or two final chapters (sometimes three) from the remaining to survey, and regret at the end that I did not have the chance to cover more.

In an attempt to be correct I have included material in this book which I would normally not mention during the course of a standard lecture. For instance, I normally do not highlight the intricacies of measure theory or integrability conditions when speaking to the class. Moreover, I often stray from the matrix approach to multiple linear regression because many of my students have not yet been formally trained in linear algebra. That being said, it is important to me for the students to hold something in their hands which acknowledges the world of mathematics and statistics beyond the classroom, and which may be useful to them for many semesters to come. It also mirrors my own experience as a student.

The vision for this document is a more or less self contained, essentially complete, correct, introductory textbook. There should be plenty of exercises for the student, with full solutions for some, and no solutions for others (so that the instructor may assign them for grading). By Sweave's dynamic nature it is possible to write randomly generated exercises and I had planned to implement this idea already throughout the book. Alas, there are only 24 hours in a day. Look for more in future editions.

Seasoned readers will be able to detect my origins: *Probability and Statistical Inference* by Hogg and Tanis [**?**], *Statistical Inference* by Casella and Berger [**?**], and *Theory of Point Estimation* and *Testing Statistical Hypotheses* by Lehmann [**?**, **?**]. I highly recommend each of those books to every reader of this one. Some R books with "introductory" in the title that I recommend are *Introductory Statistics with R* by Dalgaard [**?**] and *Using R for Introductory Statistics* by Verzani [**?**]. Surely there are many, many other good introductory books about R, but frankly, I have tried to steer clear of them for the past year or so to avoid any undue influence on my own writing.

I would like to make special mention of two other books: *Introduction to Statistical Thought* by Michael Lavine [**?**] and *Introduction to Probability* by Grinstead and Snell [**?**]. Both of these books are *free* and are what ultimately convinced me to release IPSURunder a free license, too.

Please bear in mind that the title of this book is "Introduction to Probability and Statistics Using R", and not "Introduction to R Using Probability and Statistics", nor even "Introduction to Probability and Statistics and R Using Words". The people at the party are Probability and Statistics; the handshake is R. There are several important topics about R which some individuals will feel are underdeveloped, glossed over, or wantonly omitted. Some will feel the same way about the probabilistic and/or statistical content. Still others will just want to learn R and skip all of the mathematics.

Despite any misgivings: here it is, warts and all. I humbly invite said individuals to take this book, with the GNU Free Documentation License (GNU-FDL) in hand, and make it better. In that spirit there are at least a few ways in my view in which this book could be improved.

Better data. The data analyzed in this book are almost entirely from the datasets package in base R, and here is why:

- I made a conscious effort to minimize dependence on contributed packages,
- The data are instantly available, already in the correct format, so we need not take time to manage them, and
- The data are *real*.

I made no attempt to choose data sets that would be interesting to the students; rather, data were chosen for their potential to convey a statistical point. Many of the data sets are decades old or more (for instance, the data used to introduce simple linear regression are the speeds and stopping distances of cars in the 1920's).

In a perfect world with infinite time I would research and contribute recent, *real* data in a context crafted to engage the students in *every* example. One day I hope to stumble over said time. In the meantime, I will add new data sets incrementally as time permits.

More proofs. I would like to include more proofs for the sake of completeness (I understand that some people would not consider more proofs to be improvement). Many proofs have been skipped entirely, and I am not aware of any rhyme or reason to the current omissions. I will add more when I get a chance.

More and better˜graphics. I have not used the `ggplot2` package [**?**] because I do not know how to use it yet. It is on my to-do list.

More and better exercises. There are only a few exercises in the first edition simply because I have not had time to write more. I have toyed with the `exams` package [**?**] and I believe that it is a right way to move forward. As I learn more about what the package can do I would like to incorporate it into later editions of this book.

About This Document

IPSURcontains many interrelated parts: the *Document*, the *Program*, the *Package*, and the *Ancillaries*. In short, the *Document* is what you are reading right now. The *Program* provides an efficient means to modify the Document. The *Package* is an R package that houses the Program and the Document. Finally, the *Ancillaries* are extra materials that reside in the Package and were produced by the Program to supplement use of the Document. We briefly describe each of them in turn.

The Document

The *Document* is that which you are reading right now – IPSUR's *raison d'être*. There are transparent copies (nonproprietary text files) and opaque copies (everything else). See the GNU-FDL in Appendix **??** for more precise language and details.

IPSUR.tex is a transparent copy of the Document to be typeset with a LaTeX distribution such as MikTeX or TeX Live. Any reader is free to modify the Document and release the modified version in accordance with the provisions of the GNU-FDL. Note that this file cannot be used to generate a randomized copy of the Document. Indeed, in its released form it is only capable of typesetting the exact version of IPSURwhich you are currently reading. Furthermore, the `.tex` file is unable to generate any of the ancillary materials.

IPSUR-xxx.eps, IPSUR-xxx.pdf are the image files for every graph in the Document. These are needed when typesetting with LaTeX.

IPSUR.pdf is an opaque copy of the Document. This is the file that instructors would likely want to distribute to students.

IPSUR.dvi is another opaque copy of the Document in a different file format.

The Program

The *Program* includes `IPSUR.lyx` and its nephew `IPSUR.Rnw`; the purpose of each is to give individuals a way to quickly customize the Document for their particular purpose(s).

IPSUR.lyx is the source LyX file for the Program, released under the GNU General Public License (GNU GPL) Version 3. This file is opened, modified, and compiled with LyX, a sophisticated open-source document processor, and may be used (together with `Sweave`) to generate a randomized, modified copy of the Document with brand new data sets for some of the exercises and the solution manuals (in the Second Edition). Additionally, LyX can easily activate/deactivate entire blocks of the document, *e.g.* the *proofs* of the theorems, the student *solutions* to the exercises, or the instructor *answers* to the problems, so that the new author may choose which sections (s)he would like to include in the final Document (again, Second Edition). The `IPSUR.lyx` file is all that a person needs (in addition to a properly configured system – see Appendix **??**) to generate/compile/export to all of the other formats described above and below, which includes the ancillary materials `IPSUR.Rdata` and `IPSUR.R`.

IPSUR.Rnw is another form of the source code for the Program, also released under the GNU GPL Version 3. It was produced by exporting `IPSUR.lyx` into R/Sweave format (`.Rnw`). This file may be processed with Sweave to generate a randomized copy of `IPSUR.tex` – a transparent copy of the Document – together with the ancillary materials `IPSUR.Rdata` and `IPSUR.R`. Please note, however, that `IPSUR.Rnw` is just a simple text file which does not support many of the extra features that LyX offers such as WYSIWYM editing, instantly (de)activating branches of the manuscript, and more.

The Package

There is a contributed package on `CRAN`, called `IPSUR`. The package affords many advantages, one being that it houses the Document in an easy-to-access medium. Indeed, a student can have the Document at his/her fingertips with only three commands:

Another advantage goes hand in hand with the Program's license; since IPSURis free, the source code must be freely available to anyone that wants it. A package hosted on `CRAN` allows the author to obey the license by default.

A much more important advantage is that the excellent facilities at R-Forge are building and checking the package daily against patched and development versions of the absolute latest pre-release of R. If any problems surface then I will know about it within 24 hours.

And finally, suppose there is some sort of problem. The package structure makes it *incredibly* easy for me to distribute bug-fixes and corrected typographical errors. As an author I can make my corrections, upload them to the repository at R-Forge, and they will be reflected *worldwide* within hours. We aren't in Kansas anymore, Toto.

Ancillary Materials

These are extra materials that accompany IPSUR. They reside in the `/etc` subdirectory of the package source.

IPSUR.RData is a saved image of the R workspace at the completion of the Sweave processing of IPSUR. It can be loaded into memory with `File ▷ Load Workspace` or with the command `load("/path/to/IPSUR.Rdata")`. Either method will make every single object in the file immediately available and in memory. In particular, the data BLANK from Exercise BLANK in Chapter BLANK on page BLANK will be loaded. Type BLANK at the command line (after loading `IPSUR.RData`) to see for yourself.

IPSUR.R is the exported R code from `IPSUR.Rnw`. With this script, literally every R command from the entirety of IPSURcan be resubmitted at the command line.

Notation

We use the notation `x` or `stem.leaf` notation to denote objects, functions, *etc.*. The sequence `Statistics ▷ Summaries ▷ Active Dataset` means to click the `Statistics` menu item, next click the `Summaries` submenu item, and finally click `Active Dataset`.

Acknowledgements

This book would not have been possible without the firm mathematical and statistical foundation provided by the professors at Bowling Green State University, including Drs. Gábor Székely, Craig Zirbel, Arjun K. Gupta, Hanfeng Chen, Truc Nguyen, and James Albert. I would also like to thank Drs. Neal Carothers and Kit Chan.

I would also like to thank my colleagues at Youngstown State University for their support. In particular, I would like to thank Dr. G. Andy Chang for showing me what it means to be a statistician.

I would like to thank Richard Heiberger for his insightful comments and improvements to several points and displays in the manuscript.

Finally, and most importantly, I would like to thank my wife for her patience and understanding while I worked hours, days, months, and years on a *free book*. Looking back, I can't believe I ever got away with it.

List of Figures

List of Tables

1 Discrete Distributions

In this chapter we introduce discrete random variables, those who take values in a finite or countably infinite support set. We discuss probability mass functions and some special expectations, namely, the mean, variance and standard deviation. Some of the more important discrete distributions are explored in detail, and the more general concept of expectation is defined, which paves the way for moment generating functions.

We give special attention to the empirical distribution since it plays such a fundamental role with respect to resampling and Chapter **??**; it will also be needed in Section **??** where we discuss the Kolmogorov-Smirnov test. Following this is a section in which we introduce a catalogue of discrete random variables that can be used to model experiments.

There are some comments on simulation, and we mention transformations of random variables in the discrete case. The interested reader who would like to learn more about any of the assorted discrete distributions mentioned here should take a look at *Univariate Discrete Distributions* by Johnson *et al*[**?**].

What do I want them to know?

- how to choose a reasonable discrete model under a variety of physical circumstances
- item the notion of mathematical expectation, how to calculate it, and basic properties-moment generating functions (yes, I want them to hear about those)
- the general tools of the trade for manipulation of continuous random variables, integration, *etc*.
- some details on a couple of discrete models, and exposure to a bunch of other ones
- how to make new discrete random variables from old ones

1.1 Discrete Random Variables

1.1.1 Probability Mass Functions

Discrete random variables are characterized by their supports which take the form

$$S_X = \{u_1, u_2, \ldots, u_k\} \text{ or } S_X = \{u_1, u_2, u_3 \ldots\}. \tag{1.1}$$

Every discrete random variable X has associated with it a probability mass function (PMF) $f_X : S_X \to [0,1]$ defined by

$$f_X(x) = \mathbb{P}(X = x), \quad x \in S_X. \tag{1.2}$$

Since values of the PMF represent probabilities, we know from Chapter **??** that PMFs enjoy certain properties. In particular, all PMFs satisfy

1. $f_X(x) > 0$ for $x \in S$,
2. $\sum_{x\in S} f_X(x) = 1$, and
3. $\mathbb{P}(X \in A) = \sum_{x\in A} f_X(x)$, for any event $A \subset S$.

Example 1.1. Toss a coin 3 times. The sample space would be

$$S = \{HHH,\ HTH,\ THH,\ TTH,\ HHT,\ HTT,\ THT,\ TTT\}.$$

Now let X be the number of Heads observed. Then X has support $S_X = \{0,1,2,3\}$. Assuming that the coin is fair and was tossed in exactly the same way each time, it is not unreasonable to suppose that the outcomes in the sample space are all equally likely.

What is the PMF of X? Notice that X is zero exactly when the outcome TTT occurs, and this event has probability $1/8$. Therefore, $f_X(0) = 1/8$, and the same reasoning shows that $f_X(3) = 1/8$. Exactly three outcomes result in $X = 1$, thus, $f_X(1) = 3/8$ and $f_X(3)$ holds the remaining $3/8$ probability (the total is 1). We can represent the PMF with a table:

Table 1.1: Flipping a coin three times: the PMF.

$x \in S_X$	0	1	2	3	Total
$f_X(x) = \mathbb{P}(X = x)$	1/8	3/8	3/8	1/8	1

1.1.2 Mean, Variance, and Standard Deviation

There are numbers associated with PMFs. One important example is the mean μ, also known as $\mathbb{E}X$ (which we will discuss later):

$$\mu = \mathbb{E}X = \sum_{x\in S} x f_X(x), \tag{1.3}$$

provided the (potentially infinite) series $\sum |x| f_X(x)$ is convergent. Another important number is the variance:

$$\sigma^2 = \sum_{x\in S} (x-\mu)^2 f_X(x), \tag{1.4}$$

which can be computed (see Exercise 1.3) with the alternate formula $\sigma^2 = \sum x^2 f_X(x) - \mu^2$. Directly defined from the variance is the standard deviation $\sigma = \sqrt{\sigma^2}$.

Example 1.2. We will calculate the mean of X in Example 1.1.

$$\mu = \sum_{x=0}^{3} x f_X(x) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = 1.5.$$

We interpret $\mu = 1.5$ by reasoning that if we were to repeat the random experiment many times, independently each time, observe many corresponding outcomes of the random variable X, and take the sample mean of the observations, then the calculated value would fall close to 1.5. The approximation would get better as we observe more and more values of X (another form of the Law of Large Numbers; see Section **??**). Another way it is commonly stated is that X is 1.5 "on the average" or "in the long run".

Remark 1.3. Note that although we say X is 3.5 on the average, we must keep in mind that our X never actually equals 3.5 (in fact, it is impossible for X to equal 3.5).

Related to the probability mass function $f_X(x) = \mathbb{P}(X = x)$ is another important function called the *cumulative distribution function* (CDF), F_X. It is defined by the formula

$$F_X(t) = \mathbb{P}(X \leq t), \quad -\infty < t < \infty. \tag{1.5}$$

We know that all PMFs satisfy certain properties, and a similar statement may be made for CDFs. In particular, any CDF F_X satisfies

- F_X is nondecreasing ($t_1 \leq t_2$ implies $F_X(t_1) \leq F_X(t_2)$).
- F_X is right-continuous ($\lim_{t \to a^+} F_X(t) = F_X(a)$ for all $a \in \mathbb{R}$).
- $\lim_{t \to -\infty} F_X(t) = 0$ and $\lim_{t \to \infty} F_X(t) = 1$.

We say that X has the distribution F_X and we write $X \sim F_X$. In an abuse of notation we will also write $X \sim f_X$ and for the named distributions the PMF or CDF will be identified by the family name instead of the defining formula.

How to do it with R

The mean and variance of a discrete random variable is easy to compute at the console. Let's return to Example 1.2. We will start by defining a vector `x` containing the support of X, and a vector `f` to contain the values of f_X at the respective outcomes in `x`:

```
x <- c(0,1,2,3)
f <- c(1/8, 3/8, 3/8, 1/8)
```

To calculate the mean μ, we need to multiply the corresponding values of `x` and `f` and add them. This is easily accomplished in R since operations on vectors are performed *element-wise* (see Section **??**):

```
mu <- sum(x * f)
mu
```

```
 [1] 1.5
```

To compute the variance σ^2, we subtract the value of mu from each entry in x, square the answers, multiply by f,and sum. The standard deviation σ is simply the square root of σ^2.

```
sigma2 <- sum((x-mu)^2 * f)
sigma2
```

```
 [1] 0.75
```

```
sigma <- sqrt(sigma2)
sigma
```

```
 [1] 0.8660254
```

Finally, we may find the values of the CDF F_X on the support by accumulating the probabilities in f_X with the cumsum function.

```
F <- cumsum(f)
F
```

```
 [1] 0.125 0.500 0.875 1.000
```

As easy as this is, it is even easier to do with the distrEx package [**?**]. We define a random variable X as an object, then compute things from the object such as mean, variance, and standard deviation with the functions E, var, and sd:

```
X <- DiscreteDistribution(supp = 0:3, prob = c(1,3,3,1)/8)
E(X); var(X); sd(X)
```

```
 [1] 1.5
 [1] 0.75
 [1] 0.8660254
```

1.2 The Discrete Uniform Distribution

We have seen the basic building blocks of discrete distributions and we now study particular models that statisticians often encounter in the field. Perhaps the most fundamental of all is the *discrete uniform* distribution.

A random variable X with the discrete uniform distribution on the integers $1, 2, \ldots, m$ has PMF

$$f_X(x) = \frac{1}{m}, \quad x = 1, 2, \ldots, m. \tag{1.6}$$

We write $X \sim \mathsf{disunif}(m)$. A random experiment where this distribution occurs is the choice of an integer at random between 1 and 100, inclusive. Let X be the number chosen. Then $X \sim \mathsf{disunif}(m = 100)$ and

$$\mathbb{P}(X = x) = \frac{1}{100}, \quad x = 1, \ldots, 100.$$

We find a direct formula for the mean of $X \sim \mathsf{disunif}(m)$:

$$\mu = \sum_{x=1}^{m} x f_X(x) = \sum_{x=1}^{m} x \cdot \frac{1}{m} = \frac{1}{m}(1 + 2 + \cdots + m) = \frac{m+1}{2}, \tag{1.7}$$

where we have used the famous identity $1 + 2 + \cdots + m = m(m + 1)/2$. That is, if we repeatedly choose integers at random from 1 to m then, on the average, we expect to get $(m + 1)/2$. To get the variance we first calculate

$$\sum_{x=1}^{m} x^2 f_X(x) = \frac{1}{m} \sum_{x=1}^{m} x^2 = \frac{1}{m} \frac{m(m+1)(2m+1)}{6} = \frac{(m+1)(2m+1)}{6},$$

and finally,

$$\sigma^2 = \sum_{x=1}^{m} x^2 f_X(x) - \mu^2 = \frac{(m+1)(2m+1)}{6} - \left(\frac{m+1}{2}\right)^2 = \cdots = \frac{m^2-1}{12}. \tag{1.8}$$

Example 1.4. Roll a die and let X be the upward face showing. Then $m = 6$, $\mu = 7/2 = 3.5$, and $\sigma^2 = (6^2 - 1)/12 = 35/12$.

1.2.1 How to do it with R

1.2.2 From the console:

One can choose an integer at random with the `sample` function. The general syntax to simulate a discrete uniform random variable is `sample(x, size, replace = TRUE)`.

The argument `x` identifies the numbers from which to randomly sample. If `x` is a number, then sampling is done from 1 to `x`. The argument `size` tells how big the sample size should be, and `replace` tells whether or not numbers should be replaced in the urn after having been sampled. The default option is `replace = FALSE` but for discrete uniforms the sampled values should be replaced. Some examples follow.

1.2.3 Examples

- To roll a fair die 3000 times, do `sample(6, size = 3000, replace = TRUE)`.
- To choose 27 random numbers from 30 to 70, do `sample(30:70, size = 27, replace = TRUE)`.
- To flip a fair coin 1000 times, do `sample(c("H","T"), size = 1000, replace = TRUE)`.

1.2.4 With the R Commander:

Follow the sequence `Probability` ▹ `Discrete Distributions` ▹ `Discrete Uniform distribution` ▹ `Simulate Discrete uniform variates...`.

Suppose we would like to roll a fair die 3000 times. In the `Number of samples` field we enter 1. Next, we describe what interval of integers to be sampled. Since there are six faces numbered 1 through 6, we set `from = 1`, we set `to = 6`, and set `by = 1` (to indicate that we travel from 1 to 6 in increments of 1 unit). We will generate a list of 3000 numbers selected from among 1, 2, . . . , 6, and we store the results of the simulation. For the time being, we select `New Data set`. Click `OK`.

Since we are defining a new data set, the R Commander requests a name for the data set. The default name is `Simset1`, although in principle you could name it whatever you like (according to R's rules for object names). We wish to have a list that is 3000 long, so we set `Sample Size = 3000` and click `OK`.

In the R Console window, the R Commander should tell you that `Simset1` has been initialized, and it should also alert you that `There was 1 discrete uniform variate sample stored in Simset 1.`. To take a look at the rolls of the die, we click `View data set` and a window opens.

The default name for the variable is `disunif.sim1`.

1.3 The Binomial Distribution

The binomial distribution is based on a *Bernoulli trial*, which is a random experiment in which there are only two possible outcomes: success (S) and failure (F). We conduct the Bernoulli trial and let

$$X = \begin{cases} 1 & \text{if the outcome is } S, \\ 0 & \text{if the outcome is } F. \end{cases} \tag{1.9}$$

If the probability of success is p then the probability of failure must be $1 - p = q$ and the PMF of X is

$$f_X(x) = p^x (1-p)^{1-x}, \quad x = 0, 1. \tag{1.10}$$

It is easy to calculate $\mu = \mathbb{E}X = p$ and $\mathbb{E}X^2 = p$ so that $\sigma^2 = p - p^2 = p(1-p)$.

1.3.1 The Binomial Model

The Binomial model has three defining properties:

- Bernoulli trials are conducted n times,
- the trials are independent,
- the probability of success p does not change between trials.

If X counts the number of successes in the n independent trials, then the PMF of X is

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n. \tag{1.11}$$

We say that X has a *binomial distribution* and we write $X \sim$ `binom(size` $= n$, `prob` $= p$). It is clear that $f_X(x) \geq 0$ for all x in the support because the value is the product of nonnegative numbers. We next check that $\sum f(x) = 1$:

$$\sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = [p + (1-p)]^n = 1^n = 1.$$

We next find the mean:

$$\begin{aligned}
\mu =& \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x}, \\
=& \sum_{x=1}^{n} x \, \frac{n!}{x!(n-x)!} p^x q^{n-x}, \\
=& n \cdot p \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x}, \\
=& np \sum_{x-1=0}^{n-1} \binom{n-1}{x-1} p^{(x-1)} (1-p)^{(n-1)-(x-1)}, \\
=& np.
\end{aligned}$$

A similar argument shows that $\mathbb{E}X(X-1) = n(n-1)p^2$ (see Exercise 1.4). Therefore

$$\begin{aligned}
\sigma^2 =& \mathbb{E}X(X-1) + \mathbb{E}X - [\mathbb{E}X]^2, \\
=& n(n-1)p^2 + np - (np)^2, \\
=& n^2p^2 - np^2 + np - n^2p^2, \\
=& np - np^2 = np(1-p).
\end{aligned}$$

Example 1.5. A four-child family. Each child may be either a boy (B) or a girl (G). For simplicity we suppose that $\mathbb{P}(B) = \mathbb{P}(G) = 1/2$ and that the genders of the children are determined independently. If we let X count the number of B's, then $X \sim \mathsf{binom}(\mathtt{size} = 4, \mathtt{prob} = 1/2)$. Further, $\mathbb{P}(X = 2)$ is

$$f_X(2) = \binom{4}{2}(1/2)^2(1/2)^2 = \frac{6}{2^4}.$$

The mean number of boys is $4(1/2) = 2$ and the variance of X is $4(1/2)(1/2) = 1$.

How to do it with R

The corresponding R function for the PMF and CDF are `dbinom` and `pbinom`, respectively. We demonstrate their use in the following examples.

Example 1.6. We can calculate it in R Commander under the `Binomial Distribution` menu with the `Binomial probabilities` menu item.

```
      Pr
0 0.0625
1 0.2500
2 0.3750
3 0.2500
4 0.0625
```

We know that the $\mathsf{binom}(\mathtt{size} = 4, \mathtt{prob} = 1/2)$ distribution is supported on the integers 0, 1, 2, 3, and 4; thus the table is complete. We can read off the answer to be $\mathbb{P}(X = 2) = 0.3750$.

Example 1.7. Roll 12 dice simultaneously, and let X denote the number of 6's that appear. We wish to find the probability of getting seven, eight, or nine 6's. If we let $S = \{\text{get a 6 on one roll}\}$, then $\mathbb{P}(S) = 1/6$ and the rolls constitute Bernoulli trials; thus $X \sim \mathsf{binom}(\mathtt{size} = 12, \mathtt{prob} = 1/6)$ and our task is to find $\mathbb{P}(7 \leq X \leq 9)$. This is just

$$\mathbb{P}(7 \leq X \leq 9) = \sum_{x=7}^{9} \binom{12}{x}(1/6)^x(5/6)^{12-x}.$$

Again, one method to solve this problem would be to generate a probability mass table and add up the relevant rows. However, an alternative method is to notice that $\mathbb{P}(7 \leq X \leq 9) = \mathbb{P}(X \leq 9) - \mathbb{P}(X \leq 6) = F_X(9) - F_X(6)$, so we could get the same answer by using the `Binomial tail probabilities...` menu in the R Commander or the following from the command line:

```
pbinom(9, size=12, prob=1/6) - pbinom(6, size=12, prob=1/6)
diff(pbinom(c(6,9), size = 12, prob = 1/6))  # same thing
```

```
 [1] 0.001291758
 [1] 0.001291758
```

Example 1.8. Toss a coin three times and let X be the number of Heads observed. We know from before that $X \sim$ binom(`size` = 3, `prob` = 1/2) which implies the following PMF:

Table 1.2: Flipping a coin three times: the PMF.

x = num. of Heads	0	1	2	3	Total
$f(x) = \mathbb{P}(X = x)$	1/8	3/8	3/8	1/8	1

Our next goal is to write down the CDF of X explicitly. The first case is easy: it is impossible for X to be negative, so if $x < 0$ then we should have $\mathbb{P}(X \leq x) = 0$. Now choose a value x satisfying $0 \leq x < 1$, say, $x = 0.3$. The only way that $X \leq x$ could happen would be if $X = 0$, therefore, $\mathbb{P}(X \leq x)$ should equal $\mathbb{P}(X = 0)$, and the same is true for any $0 \leq x < 1$. Similarly, for any $1 \leq x < 2$, say, $x = 1.73$, the event $\{X \leq x\}$ is exactly the event $\{X = 0 \text{ or } X = 1\}$. Consequently, $\mathbb{P}(X \leq x)$ should equal $\mathbb{P}(X = 0 \text{ or } X = 1) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1)$. Continuing in this fashion, we may figure out the values of $F_X(x)$ for all possible inputs $-\infty < x < \infty$, and we may summarize our observations with the following piecewise defined function:

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0, & x < 0, \\ \frac{1}{8}, & 0 \leq x < 1, \\ \frac{1}{8} + \frac{3}{8} = \frac{4}{8}, & 1 \leq x < 2, \\ \frac{4}{8} + \frac{3}{8} = \frac{7}{8}, & 2 \leq x < 3, \\ 1, & x \geq 3. \end{cases}$$

In particular, the CDF of X is defined for the entire real line, $\mathbb{R}$. The CDF is right continuous and nondecreasing. A graph of the binom(`size` = 3, `prob` = 1/2) CDF is shown in Figure 1.1.

Figure 1.1: A graph of the binom(`size` = 3, `prob` = 1/2) CDF.

Example 1.9. Another way to do Example 1.8 is with the `distr` family of packages [**?**]. They use an object oriented approach to random variables, that is, a random variable is stored in an object `X`, and then questions about the random variable translate to functions on and involving `X`. Random variables with distributions from the `base` package[**?**] are specified by capitalizing the name of the distribution.

```
X <- Binom(size = 3, prob = 1/2)
X
```

```
 X11cairo
        2
 X11cairo
        2
 Distribution Object of Class: Binom
  size: 3
  prob: 0.5
```

The analogue of the `dbinom` function for `X` is the `d(X)` function, and the analogue of the `pbinom` function is the `p(X)` function. Compare the following:

```
d(X)(1)   # pmf of X evaluated at x = 1
p(X)(2)   # cdf of X evaluated at x = 2
```

```
 [1] 0.375
 [1] 0.875
```

Random variables defined via the `distr` package [**?**] may be *plotted*, which will return graphs of the PMF, CDF, and quantile function (introduced in Section **??**). See Figure 1.2 for an example.

```
plot(X, cex = 0.2)
```

Figure 1.2: The binom(`size` = 3, `prob` = 0.5) distribution from the `distr` package.

1.4 Expectation and Moment Generating Functions

1.4.1 The Expectation Operator

We next generalize some of the concepts from Section 1.1.2. There we saw that every [1] PMF has two important numbers associated with it:

$$\mu = \sum_{x\in S} x f_X(x), \quad \sigma^2 = \sum_{x\in S} (x-\mu)^2 f_X(x). \tag{1.12}$$

[1] Not every, only those PMFs for which the (potentially infinite) series converges.

Table 1.3: Correspondence between `stats` and `distr`. We are given $X \sim$ dbinom(`size` $= n$, `prob` $= p$). For the `distr` package we must first set `X <- Binom(size =`n`, prob =`p`)`.

How to do:	with `stats` (default)	with `distr`
PMF: $\mathbb{P}(X = x)$	`dbinom(x, size = n, prob = p)`	`d(X)(x)`
CDF: $\mathbb{P}(X \leq x)$	`pbinom(x, size = n, prob = p)`	`p(X)(x)`
Simulate k variates	`rbinom(k, size = n, prob = p)`	`r(X)(k)`

Intuitively, for repeated observations of X we would expect the sample mean to closely approximate μ as the sample size increases without bound. For this reason we call μ the *expected value* of X and we write $\mu = \mathbb{E}X$, where $\mathbb{E}$ is an *expectation operator*.

Definition 1.10. More generally, given a function g we define the *expected value of* $g(X)$ by

$$\mathbb{E}\, g(X) = \sum_{x \in S} g(x) f_X(x), \tag{1.13}$$

provided the (potentially infinite) series $\sum_x |g(x)| f(x)$ is convergent. We say that $\mathbb{E}g(X)$ *exists*.

In this notation the variance is $\sigma^2 = \mathbb{E}(X - \mu)^2$ and we prove the identity

$$\mathbb{E}(X - \mu)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2 \tag{1.14}$$

in Exercise 1.3. Intuitively, for repeated observations of X we would expect the sample mean of the $g(X)$ values to closely approximate $\mathbb{E}\, g(X)$ as the sample size increases without bound.

Let us take the analogy further. If we expect $g(X)$ to be close to $\mathbb{E}g(X)$ on the average, where would we expect $3g(X)$ to be on the average? It could only be $3\mathbb{E}g(X)$. The following theorem makes this idea precise.

Proposition 1.11. *For any functions g and h, any random variable X, and any constant c:*

1. $\mathbb{E}\, c = c$,
2. $\mathbb{E}[c \cdot g(X)] = c\mathbb{E}g(X)$
3. $\mathbb{E}[g(X) + h(X)] = \mathbb{E}g(X) + \mathbb{E}h(X)$,

provided $\mathbb{E}g(X)$ and $\mathbb{E}h(X)$ exist.

Proof. Go directly from the definition. For example,

$$\mathbb{E}[c \cdot g(X)] = \sum_{x \in S} c \cdot g(x) f_X(x) = c \cdot \sum_{x \in S} g(x) f_X(x) = c\mathbb{E}g(X).$$

□

1.4.2 Moment Generating Functions

Definition 1.12. Given a random variable X, its *moment generating function* (abbreviated MGF) is defined by the formula

$$M_X(t) = \mathbb{E}\mathrm{e}^{tX} = \sum_{x \in S} \mathrm{e}^{tx} f_X(x), \tag{1.15}$$

provided the (potentially infinite) series is convergent for all t in a neighborhood of zero (that is, for all $-\epsilon < t < \epsilon$, for some $\epsilon > 0$).

Note that for any MGF M_X,

$$M_X(0) = \mathbb{E}\mathrm{e}^{0 \cdot X} = \mathbb{E}1 = 1. \tag{1.16}$$

We will calculate the MGF for the two distributions introduced above.

Example 1.13. Find the MGF for $X \sim \mathsf{disunif}(m)$. Since $f(x) = 1/m$, the MGF takes the form

$$M(t) = \sum_{x=1}^{m} \mathrm{e}^{tx} \frac{1}{m} = \frac{1}{m}(\mathrm{e}^{t} + \mathrm{e}^{2t} + \cdots + \mathrm{e}^{mt}), \quad \text{for any } t.$$

Example 1.14. Find the MGF for $X \sim \mathsf{binom}(\mathtt{size} = n, \mathtt{prob} = p)$.

$$\begin{aligned} M_X(t) &= \sum_{x=0}^{n} \mathrm{e}^{tx} \binom{n}{x} p^x (1-p)^{n-x}, \\ &= \sum_{x=0}^{n} \binom{n}{x} (p\mathrm{e}^t)^x q^{n-x}, \\ &= (p\mathrm{e}^t + q)^n, \quad \text{for any } t. \end{aligned}$$

Applications

We will discuss three applications of moment generating functions in this book. The first is the fact that an MGF may be used to accurately identify the probability distribution that generated it, which rests on the following:

Theorem 1.15. *The moment generating function, if it exists in a neighborhood of zero, determines a probability distribution* uniquely.

Proof. Unfortunately, the proof of such a theorem is beyond the scope of a text like this one. Interested readers could consult Billingsley [**?**]. □

We will see an example of Theorem 1.15 in action.

Example 1.16. Suppose we encounter a random variable which has MGF

$$M_X(t) = (0.3 + 0.7\mathrm{e}^t)^{13}.$$

Then $X \sim$ `binom(size` $= 13$, `prob` $= 0.7$).

An MGF is also known as a "Laplace Transform" and is manipulated in that context in many branches of science and engineering.

Why is it called a Moment Generating Function?

This brings us to the second powerful application of MGFs. Many of the models we study have a simple MGF, indeed, which permits us to determine the mean, variance, and even higher moments very quickly. Let us see why. We already know that

$$M(t) = \sum_{x \in S} \mathrm{e}^{tx} f(x).$$

Take the derivative with respect to t to get

$$M'(t) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\sum_{x \in S} \mathrm{e}^{tx} f(x)\right) = \sum_{x \in S} \frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{e}^{tx} f(x)\right) = \sum_{x \in S} x\mathrm{e}^{tx} f(x), \tag{1.17}$$

and so if we plug in zero for t we see

$$M'(0) = \sum_{x \in S} x\mathrm{e}^{0} f(x) = \sum_{x \in S} x f(x) = \mu = \mathbb{E}X. \tag{1.18}$$

Similarly, $M''(t) = \sum x^2 \mathrm{e}^{tx} f(x)$ so that $M''(0) = \mathbb{E}X^2$. And in general, we can see [2] that

$$M_X^{(r)}(0) = \mathbb{E}X^r = r^{\text{th}} \text{ moment of } X \text{ about the origin.} \tag{1.19}$$

These are also known as *raw moments* and are sometimes denoted μ_r'. In addition to these are the so called *central moments* μ_r defined by

$$\mu_r = \mathbb{E}(X - \mu)^r, \quad r = 1, 2, \ldots \tag{1.20}$$

Example 1.17. Let $X \sim$ `binom(size` $= n$, `prob` $= p$) with $M(t) = (q + p\mathrm{e}^t)^n$.

We calculated the mean and variance of a binomial random variable in Section 1.3 by means of the binomial series. But look how quickly we find the mean and variance with the moment generating function.

$$\begin{aligned} M'(t) =& n(q + p\mathrm{e}^t)^{n-1} p\mathrm{e}^t \,|_{t=0}\,, \\ =& n \cdot 1^{n-1} p, \\ =& np. \end{aligned}$$

[2] We are glossing over some significant mathematical details in our derivation. Suffice it to say that when the MGF exists in a neighborhood of $t = 0$, the exchange of differentiation and summation is valid in that neighborhood, and our remarks hold true.

And

$$\begin{aligned}M''(0) =& n(n-1)[q+pe^t]^{n-2}(pe^t)^2 + n[q+pe^t]^{n-1}pe^t\,|_{t=0}\ ,\\ \mathbb{E}X^2 =& n(n-1)p^2 + np.\end{aligned}$$

Therefore

$$\begin{aligned}\sigma^2 =& \mathbb{E}X^2 - (\mathbb{E}X)^2,\\ =& n(n-1)p^2 + np - n^2p^2,\\ =& np - np^2 = npq.\end{aligned}$$

See how much easier that was?

Remark 1.18. We learned in this section that $M^{(r)}(0) = \mathbb{E}X^r$. We remember from Calculus II that certain functions f can be represented by a Taylor series expansion about a point a, which takes the form

$$f(x) = \sum_{r=0}^{\infty} \frac{f^{(r)}(a)}{r!}(x-a)^r, \quad \text{for all } |x-a| < R, \tag{1.21}$$

where R is called the *radius of convergence* of the series (see Appendix **??**). We combine the two to say that if an MGF exists for all t in the interval $(-\epsilon, \epsilon)$, then we can write

$$M_X(t) = \sum_{r=0}^{\infty} \frac{\mathbb{E}X^r}{r!} t^r, \quad \text{for all } |t| < \epsilon. \tag{1.22}$$

How to do it with R

The `distrEx` package [**?**] provides an expectation operator `E` which can be used on random variables that have been defined in the ordinary `distr` sense:

```
X <- Binom(size = 3, prob = 0.45)
E(X)
E(3*X + 4)
```

```
 X11cairo
        2
 X11cairo
        2
 [1] 1.35
 [1] 8.05
```

For discrete random variables with finite support, the expectation is simply computed with direct summation. In the case that the random variable has infinite support and the function is crazy, then the expectation is not computed directly, rather, it is estimated by first generating a random sample from the underlying model and next computing a sample mean of the function of interest.

There are methods for other population parameters:

```
var(X)
sd(X)
```

```
 [1] 0.7425
 [1] 0.8616844
```

There are even methods for `IQR`, `mad`, `skewness`, and `kurtosis`.

1.5 The Empirical Distribution

Do an experiment n times and observe n values $x_1, x_2, \ldots, x_n$ of a random variable X. For simplicity in most of the discussion that follows it will be convenient to imagine that the observed values are distinct, but the remarks are valid even when the observed values are repeated.

Definition 1.19. The *empirical cumulative distribution function F_n* (written ECDF) is the probability distribution that places probability mass $1/n$ on each of the values $x_1, x_2, \ldots, x_n$. The empirical PMF takes the form

$$f_X(x) = \frac{1}{n}, \quad x \in \{x_1, x_2, ..., x_n\}. \tag{1.23}$$

If the value x_i is repeated k times, the mass at x_i is accumulated to k/n.

The mean of the empirical distribution is

$$\mu = \sum_{x \in S} x f_X(x) = \sum_{i=1}^{n} x_i \cdot \frac{1}{n} \tag{1.24}$$

and we recognize this last quantity to be the sample mean, $\overline{x}$. The variance of the empirical distribution is

$$\sigma^2 = \sum_{x \in S} (x - \mu)^2 f_X(x) = \sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot \frac{1}{n} \tag{1.25}$$

and this last quantity looks very close to what we already know to be the sample variance.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2. \tag{1.26}$$

The *empirical quantile function* is the inverse of the ECDF. See Section **??**.

1.5.1 How to do it with R

The empirical distribution is not directly available as a distribution in the same way that the other base probability distributions are, but there are plenty of resources available for the determined investigator. Given a data vector of observed values `x`, we can see the empirical CDF with the `ecdf` function:

```
x <- c(4, 7, 9, 11, 12)
ecdf(x)
```

```
 Empirical CDF
 Call: ecdf(x)
  x[1:5] =      4,      7,      9,     11,     12
```

The above shows that the returned value of `ecdf(x)` is not a *number* but rather a *function*. The ECDF is not usually used by itself in this form. More commonly it is used as an intermediate step in a more complicated calculation, for instance, in hypothesis testing (see Chapter **??**) or resampling (see Chapter **??**). It is nevertheless instructive to see what the `ecdf` looks like, and there is a special plot method for `ecdf` objects.

```
plot(ecdf(x))
```

Figure 1.3: The empirical CDF.

See Figure 1.3. The graph is of a right-continuous function with jumps exactly at the locations stored in `x`. There are no repeated values in `x` so all of the jumps are equal to 1/5 = 0.2.

The empirical PDF is not usually of particular interest in itself, but if we really wanted we could define a function to serve as the empirical PDF:

```
epdf <- function(x) function(t){sum(x %in% t)/length(x)}
x <- c(0,0,1)
epdf(x)(0)       # should be 2/3
```

```
 X11cairo
        2
 X11cairo
        2
 [1] 0.6666667
```

To simulate from the empirical distribution supported on the vector `x`, we use the `sample` function.

```
x <- c(0,0,1)
sample(x, size = 7, replace = TRUE)
```

```
 [1] 1 1 0 1 0 0 1
```

We can get the empirical quantile function in R with `quantile(x, probs = p, type = 1)`; see Section **??**.

As we hinted above, the empirical distribution is significant more because of how and where it appears in more sophisticated applications. We will explore some of these in later chapters – see, for instance, Chapter **??**.

1.6 Other Discrete Distributions

The binomial and discrete uniform distributions are popular, and rightly so; they are simple and form the foundation for many other more complicated distributions. But the particular uniform and binomial models only apply to a limited range of problems. In this section we introduce situations for which we need more than what the uniform and binomial offer.

1.6.1 Dependent Bernoulli Trials

The Hypergeometric Distribution

Consider an urn with 7 white balls and 5 black balls. Let our random experiment be to randomly select 4 balls, without replacement, from the urn. Then the probability of observing 3 white balls (and thus 1 black ball) would be

$$\mathbb{P}(3W, 1B) = \frac{\binom{7}{3}\binom{5}{1}}{\binom{12}{4}}. \tag{1.27}$$

More generally, we sample without replacement K times from an urn with M white balls and N black balls. Let X be the number of white balls in the sample. The PMF of X is

$$f_X(x) = \frac{\binom{M}{x}\binom{N}{K-x}}{\binom{M+N}{K}}. \tag{1.28}$$

We say that X has a *hypergeometric distribution* and write $X \sim$ hyper(`m` $= M$, `n` $= N$, `k` $= K$).

The support set for the hypergeometric distribution is a little bit tricky. It is tempting to say that x should go from 0 (no white balls in the sample) to K (no black balls in the sample), but that does not work if $K > M$, because it is impossible to have more white balls in the sample than there were white balls originally in the urn. We have the same trouble if $K > N$. The good news is that the majority of examples we study have $K \leq M$ and $K \leq N$ and we will happily take the support to be $x = 0,\ 1,\ \ldots,\ K$.

It is shown in Exercise 1.5 that

$$\mu = K\frac{M}{M+N},\qquad \sigma^2 = K\frac{MN}{(M+N)^2}\,\frac{M+N-K}{M+N-1}. \tag{1.29}$$

The associated R functions for the PMF and CDF are `dhyper(x, m, n, k)` and `phyper`, respectively. There are two more functions: `qhyper`, which we will discuss in Section **??**, and `rhyper`, discussed below.

Example 1.20. Suppose in a certain shipment of 250 Pentium processors there are 17 defective processors. A quality control consultant randomly collects 5 processors for inspection to determine whether or not they are defective. Let X denote the number of defectives in the sample.

Find the probability of exactly 3 defectives in the sample, that is, find $\mathbb{P}(X = 3)$. *Solution:* We know that $X \sim$ hyper(`m` = 17, `n` = 233, `k` = 5). So the required probability is just

$$f_X(3) = \frac{\binom{17}{3}\binom{233}{2}}{\binom{250}{5}}.$$

To calculate it in R we just type

```
dhyper(3, m = 17, n = 233, k = 5)

 [1] 0.002351153
```

To find it with the R Commander we go `Probability` ▹ `Discrete Distributions` ▹ `Hypergeometric distribution` ▹ `Hypergeometric probabilities...`. We fill in the parameters $m = 17$, $n = 233$, and $k = 5$. Click `OK`, and the following table is shown in the window.

```
A <- data.frame(Pr=dhyper(0:4, m = 17, n = 233, k = 5))
rownames(A) <- 0:4
A

             Pr
 0 7.011261e-01
 1 2.602433e-01
 2 3.620776e-02
 3 2.351153e-03
 4 7.093997e-05
```

We wanted $\mathbb{P}(X = 3)$, and this is found from the table to be approximately 0.0024. The value is rounded to the fourth decimal place. We know from our above discussion that the sample space should be $x = 0, 1, 2, 3, 4, 5$, yet, in the table the probabilities are only displayed for $x = 1, 2, 3$, and 4. What is happening? As it turns out, the R Commander will only display probabilities that are 0.00005 or greater. Since $x = 5$ is not shown, it suggests that the outcome has a tiny probability. To find its exact value we use the `dhyper` function:

```
dhyper(5, m = 17, n = 233, k = 5)

  [1] 7.916049e-07
```

In other words, $\mathbb{P}(X = 5) \approx 0.0000007916049$, a small number indeed. Find the probability that there are at most 2 defectives in the sample, that is, compute $\mathbb{P}(X \leq 2)$. *Solution:* Since $\mathbb{P}(X \leq 2) = \mathbb{P}(X = 0, 1, 2)$, one way to do this would be to add the 0, 1, and 2 entries in the above table. this gives $0.7011 + 0.2602 + 0.0362 = 0.9975$. Our answer should be correct up to the accuracy of 4 decimal places. However, a more precise method is provided by the R Commander. Under the `Hypergeometric distribution` menu we select `Hypergeometric tail probabilities...`. We fill in the parameters m, n, and k as before, but in the `Variable value(s)` dialog box we enter the value 2. We notice that the `Lower tail` option is checked, and we leave that alone. Click `OK`.

```
phyper(2, m = 17, n = 233, k = 5)

  [1] 0.9975771
```

And thus $\mathbb{P}(X \leq 2) \approx 0.9975771$. We have confirmed that the above answer was correct up to four decimal places. Find $\mathbb{P}(X > 1)$. The table did not give us the explicit probability $\mathbb{P}(X = 5)$, so we can not use the table to give us this probability. We need to use another method. Since $\mathbb{P}(X > 1) = 1 - \mathbb{P}(X \leq 1) = 1 - F_X(1)$, we can find the probability with `Hypergeometric tail probabilities...`. We enter 1 for `Variable Value(s)`, we enter the parameters as before, and in this case we choose the `Upper tail` option. This results in the following output.

```
phyper(1, m = 17, n = 233, k = 5, lower.tail = FALSE)

  [1] 0.03863065
```

In general, the `Upper tail` option of a tail probabilities dialog computes $\mathbb{P}(X > x)$ for all given `Variable Value(s)` x. Generate 100,000 observations of the random variable X. We can randomly simulate as many observations of X as we want in R Commander. Simply choose `Simulate hypergeometric variates...` in the `Hypergeometric` `distribution` dialog. In the `Number of samples` dialog, type 1. Enter the parameters as above. Under the `Store Values` section, make sure `New Data set` is selected. Click `OK`. A new dialog should open, with the default name `Simset1`. We could change this if we like, according to the rules for `R` object names. In the sample size box, enter 100000. Click `OK`. In the Console Window, `R` Commander should issue an alert that `Simset1` has been initialized, and in a few seconds, it should also state that 100,000 hypergeometric variates were stored in `hyper.sim1`. We can view the sample by clicking the `View Data Set` button on the `R` Commander interface. We know from our formulas that $\mu = K \cdot M/(M+N) = 5 * 17/250 = 0.34$. We can check our formulas using the fact that with repeated observations of X we would expect about 0.34 defectives on the average. To see how our sample reflects the true mean, we can compute the sample mean

```
    Rcmdr> mean(Simset2$hyper.sim1, na.rm=TRUE)
    [1] 0.340344

    Rcmdr> sd(Simset2$hyper.sim1, na.rm=TRUE)
    [1] 0.5584982
    ...
```

We see that when given many independent observations of X, the sample mean is very close to the true mean μ. We can repeat the same idea and use the sample standard deviation to estimate the true standard deviation of X. From the output above our estimate is 0.5584982, and from our formulas we get

$$\sigma^2 = K \frac{MN}{(M+N)^2} \frac{M+N-K}{M+N-1} \approx 0.3117896,$$

with $\sigma = \sqrt{\sigma^2} \approx 0.5583811944$. Our estimate was pretty close. From the console we can generate random hypergeometric variates with the `rhyper` function, as demonstrated below.

```
rhyper(10, m = 17, n = 233, k = 5)

    [1] 0 0 1 0 1 1 0 0 1 2
```

Sampling With and Without Replacement

Suppose that we have a large urn with, say, M white balls and N black balls. We take a sample of size n from the urn, and let X count the number of white balls in the sample. If we sample

without replacement, then $X \sim$ hyper(`m` $=M$, `n` $= N$, `k` $= n$) and has mean and variance

$$\begin{aligned}\mu =& n\frac{M}{M+N},\\ \sigma^2 =& n\frac{MN}{(M+N)^2}\frac{M+N-n}{M+N-1},\\ =& n\frac{M}{M+N}\left(1-\frac{M}{M+N}\right)\frac{M+N-n}{M+N-1}.\end{aligned}$$

On the other hand, if we sample

with replacement, then $X \sim$ binom(`size` $= n$, `prob` $= M/(M+N)$) with mean and variance

$$\begin{aligned}\mu =& n\frac{M}{M+N},\\ \sigma^2 =& n\frac{M}{M+N}\left(1-\frac{M}{M+N}\right).\end{aligned}$$

We see that both sampling procedures have the same mean, and the method with the larger variance is the “with replacement” scheme. The factor by which the variances differ,

$$\frac{M+N-n}{M+N-1}, \tag{1.30}$$

is called a *finite population correction*. For a fixed sample size n, as $M, N \to \infty$ it is clear that the correction goes to 1, that is, for infinite populations the sampling schemes are essentially the same with respect to mean and variance.

1.6.2 Waiting Time Distributions

Another important class of problems is associated with the amount of time it takes for a specified event of interest to occur. For example, we could flip a coin repeatedly until we observe Heads. We could toss a piece of paper repeatedly until we make it in the trash can.

The Geometric Distribution

Suppose that we conduct Bernoulli trials repeatedly, noting the successes and failures. Let X be the number of failures before a success. If $\mathbb{P}(S) = p$ then X has PMF

$$f_X(x) = p(1-p)^x, \quad x = 0, 1, 2, \ldots \tag{1.31}$$

(Why?) We say that X has a *Geometric distribution* and we write $X \sim$ geom(`prob` $= p$). The associated R functions are `dgeom(x, prob)`, `pgeom`, `qgeom`, and `rhyper`, which give the PMF, CDF, quantile function, and simulate random variates, respectively.

Again it is clear that $f(x) \geq 0$ and we check that $\sum f(x) = 1$ (see Equation **??** in Appendix **??**):

$$\sum_{x=0}^{\infty} p(1-p)^x = p \sum_{x=0}^{\infty} q^x = p \, \frac{1}{1-q} = 1.$$

We will find in the next section that the mean and variance are

$$\mu = \frac{1-p}{p} = \frac{q}{p} \text{ and } \sigma^2 = \frac{q}{p^2}. \tag{1.32}$$

Example 1.21. The Pittsburgh Steelers place kicker, Jeff Reed, made 81.2% of his attempted field goals in his career up to 2006. Assuming that his successive field goal attempts are approximately Bernoulli trials, find the probability that Jeff misses at least 5 field goals before his first successful goal.

Solution: If X = the number of missed goals until Jeff's first success, then $X \sim$ geom(`prob` $= 0.812$) and we want $\mathbb{P}(X \geq 5) = \mathbb{P}(X > 4)$. We can find this in R with

```
pgeom(4, prob = 0.812, lower.tail = FALSE)
```

```
 [1] 0.0002348493
```

Note 1.22. Some books use a slightly different definition of the geometric distribution. They consider Bernoulli trials and let Y count instead the number of trials until a success, so that Y has PMF

$$f_Y(y) = p(1-p)^{y-1}, \quad y = 1, 2, 3, \ldots \tag{1.33}$$

When they say "geometric distribution", this is what they mean. It is not hard to see that the two definitions are related. In fact, if X denotes our geometric and Y theirs, then $Y = X + 1$. Consequently, they have $\mu_Y = \mu_X + 1$ and $\sigma_Y^2 = \sigma_X^2$.

The Negative Binomial Distribution

We may generalize the problem and consider the case where we wait for *more* than one success. Suppose that we conduct Bernoulli trials repeatedly, noting the respective successes and failures. Let X count the number of failures before r successes. If $\mathbb{P}(S) = p$ then X has PMF

$$f_X(x) = \binom{r+x-1}{r-1} p^r (1-p)^x, \quad x = 0, 1, 2, \ldots \tag{1.34}$$

We say that X has a *Negative Binomial distribution* and write $X \sim$ nbinom(`size` $= r$, `prob` $= p$). The associated R functions are `dnbinom(x, size, prob)`, `pnbinom`, `qnbinom`, and `rnbinom`, which give the PMF, CDF, quantile function, and simulate random variates, respectively.

As usual it should be clear that $f_X(x) \geq 0$ and the fact that $\sum f_X(x) = 1$ follows from a generalization of the geometric series by means of a Maclaurin's series expansion:

$$\frac{1}{1-t} = \sum_{k=0}^{\infty} t^k, \quad \text{for } -1 < t < 1, \text{ and} \tag{1.35}$$

$$\frac{1}{(1-t)^r} = \sum_{k=0}^{\infty} \binom{r+k-1}{r-1} t^k, \quad \text{for } -1 < t < 1. \tag{1.36}$$

Therefore

$$\sum_{x=0}^{\infty} f_X(x) = p^r \sum_{x=0}^{\infty} \binom{r+x-1}{r-1} q^x = p^r (1-q)^{-r} = 1, \tag{1.37}$$

since $|q| = |1-p| < 1$.

Example 1.23. We flip a coin repeatedly and let X count the number of Tails until we get seven Heads. What is $\mathbb{P}(X = 5)$? *Solution*: We know that $X \sim$ nbinom(`size` = 7, `prob` = 1/2).

$$\mathbb{P}(X = 5) = f_X(5) = \binom{7+5-1}{7-1} (1/2)^7 (1/2)^5 = \binom{11}{6} 2^{-12}$$

and we can get this in R with

```
dnbinom(5, size = 7, prob = 0.5)
```

```
 [1] 0.112793
```

Let us next compute the MGF of $X \sim$ nbinom(`size` = r, `prob` = p).

$$\begin{aligned} M_X(t) =& \sum_{x=0}^{\infty} \mathrm{e}^{tx} \binom{r+x-1}{r-1} p^r q^x \\ =& p^r \sum_{x=0}^{\infty} \binom{r+x-1}{r-1} [q\mathrm{e}^t]^x \\ =& p^r (1 - q e^t)^{-r}, \quad \text{provided } |q\mathrm{e}^t| < 1, \end{aligned}$$

and so

$$M_X(t) = \left(\frac{p}{1 - q\mathrm{e}^t} \right)^r, \quad \text{for } q\mathrm{e}^t < 1. \tag{1.38}$$

We see that $q\mathrm{e}^t < 1$ when $t < -\ln(1-p)$.

Let $X \sim$ nbinom(`size` = r, `prob` = p) with $M(t) = p^r (1 - q\mathrm{e}^t)^{-r}$. We proclaimed above the values of the mean and variance. Now we are equipped with the tools to find these directly.

$$\begin{aligned}M'(t) =&p^r(-r)(1-qe^t)^{-r-1}(-qe^t),\\ =&rqe^t p^r(1-qe^t)^{-r-1},\\ =&\frac{rqe^t}{1-qe^t}M(t), \text{ and so}\\ M'(0) =&\frac{rq}{1-q}\cdot 1 = \frac{rq}{p}.\end{aligned}$$

Thus $\mu = rq/p$. We next find $\mathbb{E}X^2$.

$$\begin{aligned}M''(0) =& \left.\frac{rqe^t(1-qe^t)-rqe^t(-qe^t)}{(1-qe^t)^2}M(t)+\frac{rqe^t}{1-qe^t}M'(t)\right|_{t=0},\\ =&\frac{rqp+rq^2}{p^2}\cdot 1+\frac{rq}{p}\left(\frac{rq}{p}\right),\\ =&\frac{rq}{p^2}+\left(\frac{rq}{p}\right)^2.\end{aligned}$$

Finally we may say $\sigma^2 = M''(0) - [M'(0)]^2 = rq/p^2$.

Example 1.24. A random variable has MGF

$$M_X(t) = \left(\frac{0.19}{1-0.81e^t}\right)^{31}.$$

Then $X \sim$ nbinom(`size` = 31, `prob` = 0.19).

Note 1.25. As with the Geometric distribution, some books use a slightly different definition of the Negative Binomial distribution. They consider Bernoulli trials and let Y be the number of trials until r successes, so that Y has PMF

$$f_Y(y) = \binom{y-1}{r-1}p^r(1-p)^{y-r}, \quad y = r, r+1, r+2, \ldots \tag{1.39}$$

It is again not hard to see that if X denotes our Negative Binomial and Y theirs, then $Y = X + r$. Consequently, they have $\mu_Y = \mu_X + r$ and $\sigma_Y^2 = \sigma_X^2$.

1.6.3 Arrival Processes

The Poisson Distribution

This is a distribution associated with "rare events", for reasons which will become clear in a moment. The events might be:

- traffic accidents,
- typing errors, or
- customers arriving in a bank.

Let λ be the average number of events in the time interval $[0,1]$. Let the random variable X count the number of events occurring in the interval. Then under certain reasonable conditions it can be shown that

$$f_X(x) = \mathbb{P}(X = x) = \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots \tag{1.40}$$

We use the notation $X \sim$ pois(`lambda` $= \lambda$). The associated R functions are `dpois(x, lambda)`, `ppois`, `qpois`, and `rpois`, which give the PMF, CDF, quantile function, and simulate random variates, respectively.

What are the reasonable conditions?

Divide $[0,1]$ into subintervals of length $1/n$. A *Poisson process* satisfies the following conditions:

- the probability of an event occurring in a particular subinterval is $\approx \lambda/n$.
- the probability of two or more events occurring in any subinterval is ≈ 0.
- occurrences in disjoint subintervals are independent.

Remark 1.26. If X counts the number of events in the interval $[0,t]$ and λ is the average number that occur in unit time, then $X \sim$ pois(`lambda` $= \lambda t$), that is,

$$\mathbb{P}(X = x) = \mathrm{e}^{-\lambda t}\frac{(\lambda t)^x}{x!}, \quad x = 0, 1, 2, 3 \ldots \tag{1.41}$$

Example 1.27. On the average, five cars arrive at a particular car wash every hour. Let X count the number of cars that arrive from 10AM to 11AM. Then $X \sim$ pois(`lambda` $= 5$). Also, $\mu = \sigma^2 = 5$. What is the probability that no car arrives during this period? *Solution*: The probability that no car arrives is

$$\mathbb{P}(X = 0) = \mathrm{e}^{-5}\frac{5^0}{0!} = \mathrm{e}^{-5} \approx 0.0067.$$

Example 1.28. Suppose the car wash above is in operation from 8AM to 6PM, and we let Y be the number of customers that appear in this period. Since this period covers a total of 10 hours, from Remark 1.26 we get that $Y \sim$ pois(`lambda` $= 5 * 10 = 50$). What is the probability that there are between 48 and 50 customers, inclusive? *Solution*: We want $\mathbb{P}(48 \leq Y \leq 50) = \mathbb{P}(X \leq 50) - \mathbb{P}(X \leq 47)$.

```
diff(ppois(c(47, 50), lambda = 50))
```

```
 [1] 0.1678485
```

1.7 Functions of Discrete Random Variables

We have built a large catalogue of discrete distributions, but the tools of this section will give us the ability to consider infinitely many more. Given a random variable X and a given function h, we may consider $Y = h(X)$. Since the values of X are determined by chance, so are the values of Y. The question is, what is the PMF of the random variable Y? The answer, of course, depends on h. In the case that h is one-to-one (see Appendix **??**), the solution can be found by simple substitution.

Example 1.29. Let $X \sim \mathsf{nbinom}(\mathtt{size} = r, \mathtt{prob} = p)$. We saw in 1.6 that X represents the number of failures until r successes in a sequence of Bernoulli trials. Suppose now that instead we were interested in counting the number of trials (successes and failures) until the r^{th} success occurs, which we will denote by Y. In a given performance of the experiment, the number of failures (X) and the number of successes (r) together will comprise the total number of trials (Y), or in other words, $X + r = Y$. We may let h be defined by $h(x) = x + r$ so that $Y = h(X)$, and we notice that h is linear and hence one-to-one. Finally, X takes values $0,\ 1,\ 2,\ldots$ implying that the support of Y would be $\{r,\ r+1,\ r+2,\ldots\}$. Solving for X we get $X = Y - r$. Examining the PMF of X

$$f_X(x) = \binom{r+x-1}{r-1} p^r (1-p)^x, \tag{1.42}$$

we can substitute $x = y - r$ to get

$$\begin{aligned} f_Y(y) &= f_X(y-r), \\ &= \binom{r+(y-r)-1}{r-1} p^r (1-p)^{y-r}, \\ &= \binom{y-1}{r-1} p^r (1-p)^{y-r}, \quad y = r,\ r+1,\ldots \end{aligned}$$

Even when the function h is not one-to-one, we may still find the PMF of Y simply by accumulating, for each y, the probability of all the x's that are mapped to that y.

Proposition 1.30. *Let X be a discrete random variable with PMF f_X supported on the set S_X. Let $Y = h(X)$ for some function h. Then Y has PMF f_Y defined by*

$$f_Y(y) = \sum_{\{x \in S_X \mid h(x) = y\}} f_X(x) \tag{1.43}$$

Example 1.31. Let $X \sim \mathsf{binom}(\mathtt{size} = 4, \mathtt{prob} = 1/2)$, and let $Y = (X-1)^2$. Consider the following table:

From this we see that Y has support $S_Y = \{0, 1, 4, 9\}$. We also see that $h(x) = (x-1)^2$ is not one-to-one on the support of X, because both $x = 0$ and $x = 2$ are mapped by h

Table 1.4: Transforming a discrete random variable.

x	0	1	2	3	4
$f_X(x)$	1/16	1/4	6/16	1/4	1/16
$y = (x-1)^2$	1	0	1	4	9

to $y = 1$. Nevertheless, we see that $Y = 0$ only when $X = 1$, which has probability $1/4$; therefore, $f_Y(0)$ should equal $1/4$. A similar approach works for $y = 4$ and $y = 9$. And $Y = 1$ exactly when $X = 0$ or $X = 2$, which has total probability $7/16$. In summary, the PMF of Y may be written:

Table 1.5: Transforming a discrete random variable, its PMF.

y	0	1	4	9
$f_Y(y)$	1/4	7/16	1/4	1/16

There is not a special name for the distribution of Y, it is just an example of what to do when the transformation of a random variable is not one-to-one. The method is the same for more complicated problems.

Proposition 1.32. *If X is a random variable with $\mathbb{E}X = \mu$ and $Var(X) = \sigma^2$, then the mean and variance of $Y = mX + b$ is*

$$\mu_Y = m\mu + b, \quad \sigma_Y^2 = m^2\sigma^2, \quad \sigma_Y = |m|\sigma. \tag{1.44}$$

1.8 Exercises

Exercise 1.1. A recent national study showed that approximately 44.7% of college students have used Wikipedia as a source in at least one of their term papers. Let X equal the number of students in a random sample of size $n = 31$ who have used Wikipedia as a source.

- How is X distributed?
- Sketch the probability mass function (roughly).
- Sketch the cumulative distribution function (roughly).
- Find the probability that X is equal to 17.
- Find the probability that X is at most 13.
- Find the probability that X is bigger than 11.
- Find the probability that X is at least 15.
- Find the probability that X is between 16 and 19, inclusive.
- Give the mean of X, denoted $\mathbb{E}X$.
- Give the variance of X.
- Give the standard deviation of X.
- Find $\mathbb{E}(4X + 51.324)$.

Exercise 1.2. For the following situations, decide what the distribution of X should be. In nearly every case, there are additional assumptions that should be made for the distribution to apply; identify those assumptions (which may or may not hold in practice.)

- We shoot basketballs at a basketball hoop, and count the number of shots until we make a goal. Let X denote the number of missed shots. On a normal day we would typically make about 37% of the shots.
- In a local lottery in which a three digit number is selected randomly, let X be the number selected.
- We drop a Styrofoam cup to the floor twenty times, each time recording whether the cup comes to rest perfectly right side up, or not. Let X be the number of times the cup lands perfectly right side up.

- We toss a piece of trash at the garbage can from across the room. If we miss the trash can, we retrieve the trash and try again, continuing to toss until we make the shot. Let X denote the number of missed shots.

- Working for the border patrol, we inspect shipping cargo as when it enters the harbor looking for contraband. A certain ship comes to port with 557 cargo containers. Standard practice is to select 10 containers randomly and inspect each one very carefully, classifying it as either having contraband or not. Let X count the number of containers that illegally contain contraband.

- At the same time every year, some migratory birds land in a bush outside for a short rest. On a certain day, we look outside and let X denote the number of birds in the bush.

- We count the number of rain drops that fall in a circular area on a sidewalk during a ten minute period of a thunder storm.

- We count the number of moth eggs on our window screen.

- We count the number of blades of grass in a one square foot patch of land.

- We count the number of pats on a baby's back until (s)he burps.

Exercise 1.3. Show that $\mathbb{E}(X-\mu)^2 = \mathbb{E}X^2 - \mu^2$. *Hint*: expand the quantity $(X-\mu)^2$ and distribute the expectation over the resulting terms.

Exercise 1.4. If $X \sim \mathtt{binom}(\mathtt{size} = n,\ \mathtt{prob} = p)$ show that $\mathbb{E}X(X-1) = n(n-1)p^2$.

Exercise 1.5. Calculate the mean and variance of the hypergeometric distribution. Show that

$$\mu = K\frac{M}{M+N}, \quad \sigma^2 = K\frac{MN}{(M+N)^2}\frac{M+N-K}{M+N-1}. \tag{1.45}$$