PROBABILITY AND STATISTICS

FOR SCIENCE & ENGINEERING WITH EXAMPLES IN R

2nd EDITION

BY HONGSHIK AHN

SECOND EDITION

PROBABILITY AND STATISTICS

FOR SCIENCE AND ENGINEERING

WITH EXAMPLES IN R

BY HONGSHIK AHN

cognella® | ACADEMIC PUBLISHING

Bassim Hamadeh, CEO and Publisher
Carrie Montoya, Manager, Revisions and Author Care
Kaela Martin, Project Editor
Christian Berk, Associate Production Editor
Miguel Macias, Senior Graphic Designer
Alexa Lucido, Licensing Associate
Natalie Piccotti, Senior Marketing Manager
Kassie Graves, Vice President of Editorial
Jamie Giganti, Director of Academic Publishing

Printed in the United States of America.

ISBN: 978-1-5165-3110-3 (pbk) / 978-1-5165-3111-0 (br)

Contents

Preface

This book is designed for a one-semester course in probability and statistics, specifically for students in the natural science or engineering. It is also suitable for business and economics students with a calculus background. The text is based on my past teachings over the course of many years at Stony Brook University. Most existing textbooks contain topics to fulfill at least one whole year of instruction, and therefore, they are excessive for a one semester course in probability and statistics. The purpose of this book is to cover just the necessary topics for a one-semester course, thus reducing the typical volume of a textbook and lowering the financial burden on students.

This book provides examples of how to use the R software to obtain summary statistics, calculate probabilities and quantiles, find confidence intervals, and conduct statistical testing. In addition to using distribution tables, students can calculate probabilities of various distributions using their smartphones or computers. Since R is available under an open-source license, everyone can download it and use it free of charge.

In the second edition, the new Section 1.4 on relationship between two variables has been added to the first addition. In Chapter 8 and Chapter 9, flowcharts have been added for identifying the appropriate test methods for population means. Also, 27% more exercise problems have been added, and some problems in the first edition have been modified.

This book is organized as follows: Chapter 1 covers descriptive statistics, Chapter 2 through Chapter 5 cover probability and distributions, Chapter 6 provides concepts about sampling, and Chapter 7 through 9 cover estimations and hypothesis testing. Hypothesis testing can be overwhelming for students, due to the inundating formulas needed for numerous cases. To help students understand and identify the correct formula to use, a comprehensive table for each type of test is provided. As mentioned earlier, flowcharts have been added in the second edition. Students can easily follow these tables to choose the appropriate confidence intervals and statistical tests. A summary is given at the end of each chapter. Distribution tables for various distributions are provided in the appendix.

I wish to thank Hyojeong Son for carefully reviewing the first edition, checking answers to the exercise problems, and creating the online tools; David Saltz for contributing some examples and exercise problems; Chelsea Kennedy for proofreading the preliminary edition and checking the solutions to the exercise problems; Yan Yu for proofreading the preliminary edition; Jerson Cochancela for proofreading the manuscript, contributing an exercise problem, and providing helpful suggestions; and Mingshen Chen for checking the solutions to the exercise problems for the preliminary edition. I also thank Cognella for inviting me to write this book.

Hongshik Ahn
Department of Applied Mathematics and Statistics
Stony Brook University

Describing Data

1. Display of Data by Graphs and Tables

There are various ways to describe data. In this section, we study how to organize and describe a set of data using graphs and tables.

A. FREQUENCY DISTRIBUTIONS

A *frequency distribution* is a table that displays the frequency of observations in each interval in a sample. To build a frequency distribution, we need the following steps.

Basic Steps

1. Find the minimum and the maximum values in the data set.
2. Determine class intervals: intervals or cells of equal length that cover the range between the minimum and the maximum without overlapping
   e.g., minimum 0, maximum 100: [0, 10), [10, 20), …, [90, 100]
3. Find frequency: the number of observations in the data that belong to each class interval. Let's denote the frequencies as $f_1, f_2, \cdots$.
4. Find relative frequency:

$$\frac{\text{Class frequency}}{\text{Total number of observations}}$$

The relative frequencies are denoted as $f_1/n,\ f_2/n, \cdots$ if the total sample size is n.

EXAMPLE 1.1 Midterm scores of an introductory statistics class of 20 students are given below.

69 84 52 93 81 74 89 85 88 63 87 64 67 72 74 55 82 91 68 77

We can construct a frequency table as shown in Table 1.1.

TABLE 1.1 Frequency table for Example 1.1

Class interval	Tally	Frequency	Relative frequency
50–59	||	2	2/20 = 0.10
60–69	𝍸	5	5/20 = 0.25
70–79	||||	4	4/20 = 0.20
80–89	𝍸 ||	7	7/20 = 0.35
90–99	||	2	2/20 = 0.10
Total		20	1.00

There is no gold standard in selecting class intervals, but a rule of thumb is an integer near $\sqrt{n}$ for the number of classes.

B. HISTOGRAM

A *histogram* is a pictorial representation of a frequency distribution. Figure 1.1 is a histogram obtained from the frequency distribution in Example 1.1.

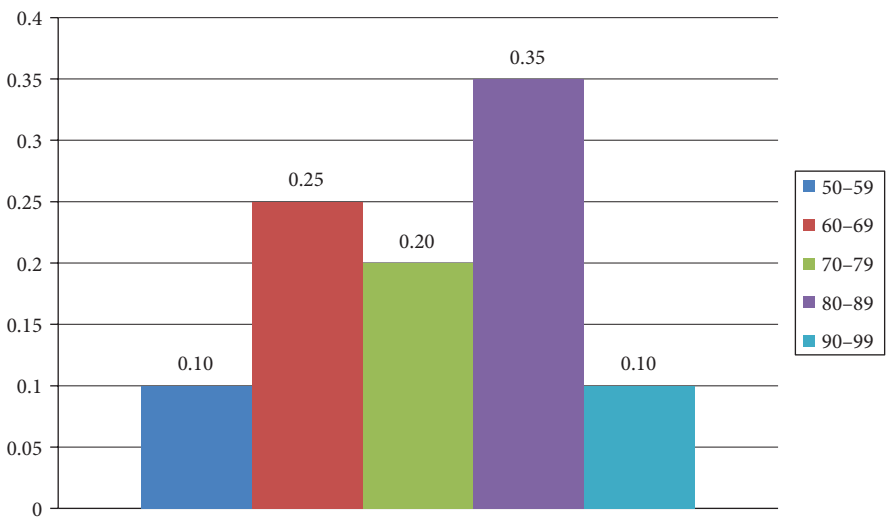

FIGURE 1.1 Histogram of the data in Example 1.1.

Figure 1.1 used the relative frequency as the height of the bar in each class. The histogram adequately visualizes the frequency distribution of the data. We expect that the class with a longer bar would have a higher count. However, the histogram may not appropriately display the frequency distribution and thus may mislead the data interpretation when we use the height as the relative frequency if the interval lengths are not equal. To avoid this, we can divide the relative frequency by the interval length for each class. Then the area of each bar becomes the relative frequency of the class, and thus the total area of the histogram becomes 1. This is necessary when the interval lengths are different. The height of this histogram is obtained as follows:

$$\text{Height} = \frac{\text{Relative frequency}}{\text{Width of the interval}}$$

For Example 1.1, the height of each bar in the histogram is:

$$\begin{aligned}
\text{Height} &= \frac{\text{Relative frequency}}{\text{Width of the interval}} \\
&= \frac{0.10}{10} = 0.010 \text{ for } [50, 60) \\
&= \frac{0.25}{10} = 0.025 \text{ for } [60, 70) \\
&= \frac{0.20}{10} = 0.020 \text{ for } [70, 80) \\
&= \frac{0.35}{10} = 0.035 \text{ for } [80, 90) \\
&= \frac{0.10}{10} = 0.010 \text{ for } [90, 100)
\end{aligned}$$

A histogram shows the shape of a distribution. Depending on the number of peaks, a distribution can be called *unimodal* (one peak), *bimodal* (two peaks) or *multimodal* (multiple peaks). A distribution can be *symmetric* or *skewed*. A skewed distribution is asymmetrical with a longer tail on one side. A distribution with a longer right tail is called skewed to the right (right skewed or positively skewed), and a distribution with a longer left tail is called skewed to the left (left skewed or negatively skewed). Figure 1.2 displays some typical shapes of distributions.

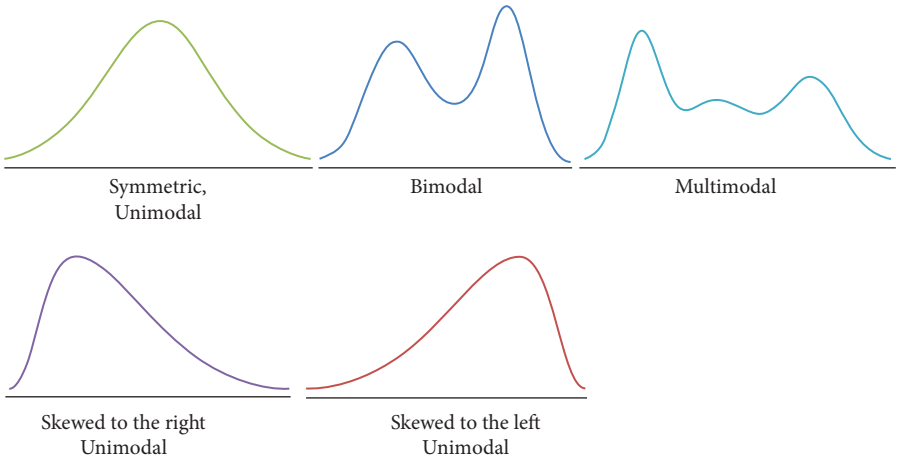

FIGURE 1.2 Shapes of distributions.

EXAMPLE 1.2 The following are the midterm exam scores of a probability and statistics course in a past semester at Stony Brook University.

30, 34, 38, 44, 45, 46, 47, 48, 50, 50, 51, 52, 53, 53, 53, 54, 55, 55, 55, 56, 56, 57, 57, 58, 58, 59, 59, 60, 60, 60, 60, 61, 61, 62, 62, 62, 62, 63, 63, 63, 63, 63, 63, 63, 64, 64, 65, 65, 65, 65, 65, 66, 66, 67, 67, 67, 68, 68, 68, 68, 68, 69, 69, 69, 69, 69, 69, 69, 70, 70, 70, 70, 70, 70, 71, 71, 71, 72, 72, 73, 73, 73, 73, 73, 73, 73, 73, 73, 74, 74, 74, 75, 75, 75, 76, 76, 76, 76, 76, 76, 77, 77, 77, 77, 77, 78, 78, 78, 78, 78, 79, 79, 79, 80, 80, 80, 80, 81, 81, 81, 81, 82, 82, 82, 82, 82, 83, 83, 83, 83, 84, 84, 84, 84, 84, 84, 84, 84, 85, 85, 86, 86, 87, 87, 87, 87, 88, 88, 88, 88, 88, 88, 89, 89, 89, 89, 89, 90, 90, 90, 91, 92, 93, 93, 94, 94, 94, 94, 95, 95, 95, 95, 96, 96, 96, 97, 97, 98, 98, 99, 100

A histogram of the above data can be obtained using R statistical software. R is a programming language and software for statistical analysis. It is freely available and can be downloaded from the Internet. You can read the data file as:

```
>midterm=read.csv("filename.csv")
```

or enter the data on R as

```
>midterm=c(30,34,38,…,100)
```

Here, > is the cursor in R. To read a data file from your computer, it must be a *comma separated values* file with extension .csv. You may need to list the directories containing the file, such as:

>midterm=read.csv("c:\\Users***\\filename.csv")

Here, *** is the name(s) of subdirectory (or subdirectories). Using the command "hist" as below,

>hist(midterm)

we obtain the histogram given in Figure 1.3.

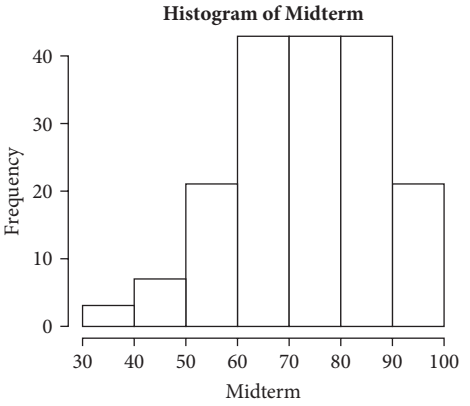

FIGURE 1.3 Histogram of the data in Example 1.2 generated by R.

C. STEM-AND-LEAF PLOT

A *stem-and-leaf plot* displays data in a graphical format, similar to a histogram. Unlike a histogram, a stem-and-leaf plot retains the original data and puts the data in order. Thus, a stem-and-leaf plot provides more details about the data than a histogram. A stem-and-leaf plot consists of two columns separated by a vertical line. The left column containing the leading digit(s) is called the *stem*, and the right column containing the trailing digit(s) is called the *leaf*. Figure 1.4 shows the shape of a stem-and-leaf plot.

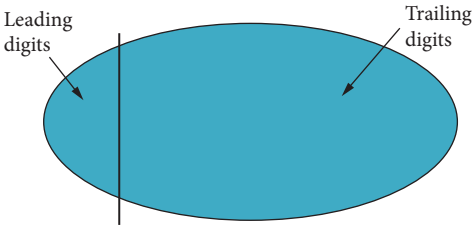

FIGURE 1.4 Shape of a stem-and-leaf plot.

To construct a stem-and-leaf plot, one or two leading digits are listed for the stem values. The trailing digit(s) become the leaf. The trailing digits in each row of the leaf are arranged in ascending order. Steps for constructing a stem-and-leaf plot are given below.

Basic Steps

1. List one or more leading digits for the stem values.
2. The trailing digit(s) become the leaves.
3. Arrange the trailing digits in each row so they are in increasing order.

EXAMPLE 1.3 Final examination scores of 26 students in an introductory statistics course are given below.

55 61 94 94 69 77 68 54 85 77 92 92 81 73 69 81 75 84 70 81 81 89 59 72 82 62

The following is a stem-and-leaf plot for the above data.

5	549		5	459
6	19892		6	12899
7	773502	→	7	023577
8	51141192		8	11112459
9	4422		9	2244

Using R, a stem-and-leaf plot for the above data can be obtained by

```
>a=c(55,61,94,94,69,77,68,54,85,77,92,92,81,73,69,81,75,84,70,81,81,89,59,72,82,62)

>stem(a)
```

D. DOT DIAGRAM

The data in Example 1.3 can be displayed using a dot diagram, as shown in Figure 1.5.

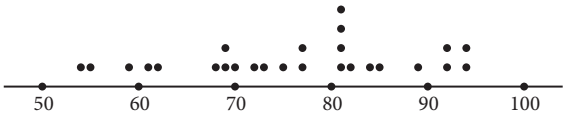

FIGURE 1.5 Dot diagram for the data in Example 1.3.

EXAMPLE 1.4 Heights of students (in inches) in a varsity wrestling team are given below.

67.2 65.0 72.5 71.1 69.1 69.0 70.2 68.2 68.5 71.3

67.5 68.6 73.1 71.3 69.4 65.5 69.5 70.8 70.0 69.2

A stem-and-leaf plot can have the tens digit or the first two digits in stem, but the latter will display the distribution more efficiently, as shown below.

65	05
66	
67	25
68	256
69	10452
70	280
71	133
72	5
73	1

$\rightarrow$

65	05
66	
67	25
68	256
69	01245
70	028
71	133
72	5
73	1

Figure 1.6 is a dot diagram of the above data.

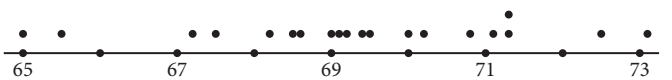

FIGURE 1.6 Dot diagram for the data in Example 1.4.

EXAMPLE 1.5 We can compare distributions of two sets of data using side-by-side stem-and-leaf plots. The same examination is given to two classes. The scores of the two classes are given below.

Class A: 78 80 60 74 85 100 51 60 40 67 100 90 58 40 89 100

Class B: 42 76 37 57 93 60 55 47 51 95 81 53 52 65 95

The following side-by-side stem-and-leaf plots compare the score distributions of the two classes.

Class A		Class B
	3	7
00	4	27
18	5	12357
007	6	05
48	7	6
059	8	1
0	9	355
000	10	

The above plots show that Class A performed better than Class B in general.

A histogram can be built for qualitative (categorical) data.

EXAMPLE 1.6 A frequency distribution of the enrollment of four classes in a high school is given in the following table.

Class	Frequency	Relative frequency
Algebra	26	0.26
English	30	0.30
Physics	19	0.19
Biology	24	0.24
Total	99	0.99

Note that the total of the relative frequencies is 0.99. This is due to a rounding error. The above table can be visualized using the bar graph in Figure 1.7.

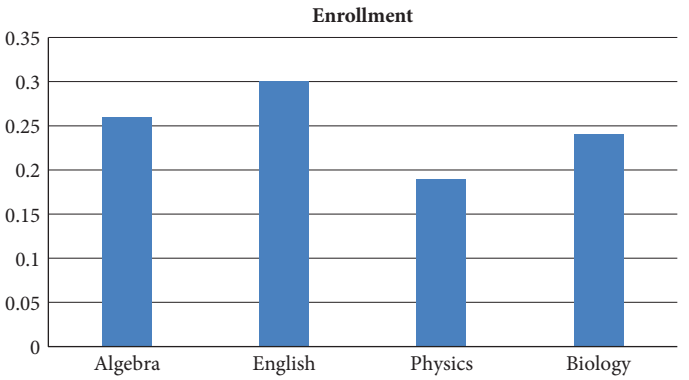

FIGURE 1.7 Bar graph for the enrollment data in Example 1.6.

The data can also be displayed using the pie chart given in Figure 1.8. In a pie chart, a circle is divided into slices according to the proportion of each group. The angle of each slice is obtained by (class frequency / sample size) $\times 360°$. It is equivalent to the relative frequency multiplied by 360°. For the above data, the central angle of each slide is obtained as follows:

Algebra: $(26/99) \times 360° = 94.5°$

English: $(30/99) \times 360° = 109.1°$

Physics: $(19/99) \times 360° = 69.1°$

Biology: $(24/99) \times 360° = 87.3°$