OpenIntro Statistics Fourth Edition

David Diez
Data Scientist
OpenIntro

Mine C¸ etinkaya-Rundel
Associate Professor of the Practice, Duke University
Professional Educator, RStudio

Christopher D Barr
Investment Analyst
Varadero Capital

Table of Contents

Preface

OpenIntro Statistics covers a first course in statistics, providing a rigorous introduction to applied statistics that is clear, concise, and accessible. This book was written with the undergraduate level in mind, but it's also popular in high schools and graduate courses.

We hope readers will take away three ideas from this book in addition to forming a foundation of statistical thinking and methods.

- Statistics is an applied field with a wide range of practical applications.
- You don't have to be a math guru to learn from real, interesting data.
- Data are messy, and statistical tools are imperfect. But, when you understand the strengths

and weaknesses of these tools, you can use them to learn about the world.

Textbook overview

The chapters of this book are as follows:

1. Introduction to data. Data structures, variables, and basic data collection techniques. 2. Summarizing data. Data summaries, graphics, and a teaser of inference using randomization. 3. Probability. Basic principles of probability.

4. Distributions of random variables. The normal model and other key distributions.
5. Foundations for inference. General ideas for statistical inference in the context of estimating the population proportion.
6. Inference for categorical data. Inference for proportions and tables using the normal and chi-square distributions.
7. Inference for numerical data. Inference for one or two sample means using the *t*-distribution, statistical power for comparing two groups, and also comparisons of many means using ANOVA.
8. Introduction to linear regression. Regression for a numerical outcome with one predictor variable. Most of this chapter could be covered after Chapter 1.
9. Multiple and logistic regression. Regression for numerical and categorical data using many predictors.

OpenIntro Statistics supports flexibility in choosing and ordering topics. If the main goal is to reach multiple regression (Chapter 9) as quickly as possible, then the following are the ideal prerequisites:

- Chapter 1, Sections 2.1, and Section 2.2 for a solid introduction to data structures and statis tical summaries that are used throughout the book.
- Section 4.1 for a solid understanding of the normal distribution.
- Chapter 5 to establish the core set of inference tools.
- Section 7.1 to give a foundation for the *t*-distribution
- Chapter 8 for establishing ideas and principles for single predictor regression.

Examples and exercises

Examples are provided to establish an understanding of how to apply methods

EXAMPLE 0.1

This is an example. When a question is asked here, where can the answer be found?

The answer can be found here, in the solution section of the example!

When we think the reader should be ready to try determining the solution to an example, we frame it as Guided Practice.

GUIDED PRACTICE 0.2

The reader may check or learn the answer to any Guided Practice problem by reviewing the full solution in a footnote.[1]

Exercises are also provided at the end of each section as well as review exercises at the end of each chapter. Solutions are given for odd-numbered exercises in Appendix A.

Additional resources

Video overviews, slides, statistical software labs, data sets used in the textbook, and much more are readily available at

openintro.org/os

We also have improved the ability to access data in this book through the addition of Appendix B, which provides additional information for each of the data sets used in the main text and is new in the Fourth Edition. Online guides to each of these data sets are also provided at openintro.org/data and through a companion R package.

We appreciate all feedback as well as reports of any typos through the website. A short-link to report a new typo or review known typos is openintro.org/os/typos.

For those focused on statistics at the high school level, consider *Advanced High School Statistics*, which is a version of *OpenIntro Statistics* that has been heavily customized by Leah Dorazio for high school courses and AP® Statistics.

Acknowledgements

This project would not be possible without the passion and dedication of many more people beyond those on the author list. The authors would like to thank the OpenIntro Staff for their involvement and ongoing contributions. We are also very grateful to the hundreds of students and instructors who have provided us with valuable feedback since we first started posting book content in 2009.

We also want to thank the many teachers who helped review this edition, including Laura Acion, Matthew E. Aiello-Lammens, Jonathan Akin, Stacey C. Behrensmeyer, Juan Gomez, Jo Hardin, Nicholas Horton, Danish Khan, Peter H.M. Klaren, Jesse Mostipak, Jon C. New, Mario Orsi, Steve Phelps, and David Rockoff. We appreciate all of their feedback, which helped us tune the text in significant ways and greatly improved this book.

[1] Guided Practice problems are intended to stretch your thinking, and you can check yourself by reviewing the footnote solution for any Guided Practice.

Chapter

1 Introduction to data

Scientists seek to answer questions using rigorous methods and careful observations. These observations – collected from the likes of field notes, surveys, and experiments – form the backbone of a statistical investigation and are called data. Statistics is the study of how best to collect, analyze, and draw conclusions from data, and in this first chapter, we focus on both the properties of data and on the collection of data.

For videos, slides, and other resources, please visit

www.openintro.org/os

1.1 Case study: using stents to prevent strokes

Section 1.1 introduces a classic challenge in statistics: evaluating the efficacy of a medical treatment. Terms in this section, and indeed much of this chapter, will all be revisited later in the text. The plan for now is simply to get a sense of the role statistics can play in practice.

In this section we will consider an experiment that studies effectiveness of stents in treating patients at risk of stroke. Stents are devices put inside blood vessels that assist in patient recovery after cardiac events and reduce the risk of an additional heart attack or death. Many doctors have hoped that there would be similar benefits for patients at risk of stroke. We start by writing the principal question the researchers hope to answer:

> Does the use of stents reduce the risk of stroke?

The researchers who asked this question conducted an experiment with 451 at-risk patients. Each volunteer patient was randomly assigned to one of two groups:

> Treatment group. Patients in the treatment group received a stent and medical management. The medical management included medications, management of risk factors, and

help in lifestyle modification.

Control group. Patients in the control group received the same medical management as the treatment group, but they did not receive stents.

Researchers randomly assigned 224 patients to the treatment group and 227 to the control group. In this study, the control group provides a reference point against which we can measure the medical impact of stents in the treatment group.

Researchers studied the effect of stents at two time points: 30 days after enrollment and 365 days after enrollment. The results of 5 patients are summarized in Figure 1.1. Patient outcomes are recorded as "stroke" or "no event", representing whether or not the patient had a stroke at the end of a time period.

Patient	group	0-30 days	0-365 days
1	treatment	no event	no event
2	treatment	stroke	stroke
3	treatment	no event	no event
⋮	⋮	⋮	
450	control	no event	no event
451	control	no event	no event

Figure 1.1: Results for five patients from the stent study.

Considering data from each patient individually would be a long, cumbersome path towards answering the original research question. Instead, performing a statistical data analysis allows us to consider all of the data at once. Figure 1.2 summarizes the raw data in a more helpful way. In this table, we can quickly see what happened over the entire study. For instance, to identify the number of patients in the treatment group who had a stroke within 30 days, we look on the left-side of the table at the intersection of the treatment and stroke: 33.

	0-30 days		0-365 days	
	stroke	no event	stroke	no event
treatment	33	191	45	179
control	13	214	28	199
Total	46	405	73	378

Figure 1.2: Descriptive statistics for the stent study.

GUIDED PRACTICE 1.1

Of the 224 patients in the treatment group, 45 had a stroke by the end of the first year. Using these two numbers, compute the proportion of patients in the treatment group who had a stroke by the end of their first year. (Please note: answers to all Guided Practice exercises are provided using footnotes.)[1]

We can compute summary statistics from the table. A summary statistic is a single number summarizing a large amount of data. For instance, the primary results of the study after 1 year could be described by two summary statistics: the proportion of people who had a stroke in the treatment and control groups.

Proportion who had a stroke in the treatment (stent) group: 45/224 = 0.20 = 20%.

Proportion who had a stroke in the control group: 28/227 = 0.12 = 12%.

These two summary statistics are useful in looking for differences in the groups, and we are in for a surprise: an additional 8% of patients in the treatment group had a stroke! This is important for two reasons. First, it is contrary to what doctors expected, which was that stents would *reduce* the rate of strokes. Second, it leads to a statistical question: do the data show a "real" difference between the groups?

This second question is subtle. Suppose you flip a coin 100 times. While the chance a coin lands heads in any given coin flip is 50%, we probably won't observe exactly 50 heads. This type of fluctuation is part of almost any type of data generating process. It is possible that the 8% difference in the stent study is due to this natural variation. However, the larger the difference we observe (for a particular sample size), the less believable it is that the difference is due to chance. So what we are really asking is the following: is the difference so large that we should reject the notion that it was due to chance?

While we don't yet have our statistical tools to fully address this question on our own, we can comprehend the conclusions of the published analysis: there was compelling evidence of harm by stents in this study of stroke patients.

Be careful: Do not generalize the results of this study to all patients and all stents. This study looked at patients with very specific characteristics who volunteered to be a part of this study and who may not be representative of all stroke patients. In addition, there are many types of stents and this study only considered the self-expanding Wingspan stent (Boston Scientific). However, this study does leave us with an important lesson: we should keep our eyes open for surprises.

[1] The proportion of the 224 patients who had a stroke within 365 days: 45/224 = 0.20.

Exercises

1.1 Migraine and acupuncture, Part I. A migraine is a particularly painful type of headache, which patients sometimes wish to treat with acupuncture. To determine whether acupuncture relieves migraine pain, researchers conducted a randomized controlled study where 89 females diagnosed with migraine headaches were randomly assigned to one of two groups: treatment or control. 43 patients in the treatment group received acupuncture that is specifically designed to treat migraines. 46 patients in the control group received placebo acupuncture (needle insertion at non-acupoint locations). 24 hours after patients received S174 Neurol Sci (2011) 32 (Suppl 1):S173–S175 acupuncture, they were asked if they were pain free. Results are summarized in the contingency table below.[2]

		Pain free		
		Yes	No	Total
Group	Treatment	10	33	43
	Control	2	44	46
	Total	12	77	89

identified on the antero-internal part of the antitragus, the anterior part of the lobe and the upper auricular concha, on the same side of pain. The majority of these points were

Fig. 1 The appropriate area (M) versus the inappropriate area (S) used in the treatment of migraine attacks

Figure from the original pa

effective very rapidly (within 1 min), while the remaining points produced a slower antalgic response, between 2 ~~and~~

5 min. The insertion of a semi-pe~~rmanent needle in these~~

zones allowed stable control of the migraine pain, which occurred within 30 min and still persisted 24 h later. Since the most active site in co~~ntrolling migraine pain~~

was the antero-internal part of the antitragus, the aim of this study was to verify the therapeutic value of this elec tive area (appropriate point) and to compare it with an area of the ear (representing the sciatic nerve) which is probably inappropriate in terms of giving a therapeutic effect on

per displaying the appropri ate area (M) versus the in appropriate area (S) used in the treatment of migraine at tacks.

(a) What percent of patients in the treatment group were pain free 24 hours after receiving acupuncture?

migraine attacks, since it has no somatotopic correlation In group B, the lower branch of the anthelix was with head pain.

pain free 24 hours after receiving acupuncture? auricular maps, this area corresponds to the representation

(b) What percent were pain free in the control group?

repeatedly tested with the algometer for about 30 s to ensure it was not sensitive. On both the French and Chinese

(c) In which group did a higher percent of patients become

Materials and methods of the sciatic nerve (Fig. 1, area S) and is specifically used

(d) Your findings so far might suggest that acupuncture is an effective treatment for migraines for all people to treat sciatic pain. Four needles were inserted in this area,

who suffer from migraines. However this is not the only possible conclusion that can be drawn based

The study enrolled 94 females, two for each ear. diagnosed as migraine

on your findings so far. What is one other possible explanation for the observed difference between the

without aura following the International Classification of In all patients, the ear acupuncture was always per Headache Disorders [5], who were subsequently examined formed by an experienced acupuncturist. The analysis of

percentages of patients that are pain free 24 hours after receiving acupuncture in the two groups?

at the Women's Headache Centre, Department of Gynae the diaries collecting VAS data was conducted by an cology and Obstetrics of Turin University. They were all impartial operator who did not know the group each patient

1.2 Sinusitis and antibiotics, Part I. Researchers studying the effect of antibiotic treatment for acute

included in the study during a migraine attack provided was in. that it started no more than 4 h previously. According to a The average values of VAS in group A and B were

sinusitis compared to symptomatic treatments randomly assigned 166 adults diagnosed with acute sinusitis to

predetermined computer-made randomization list, the eli calculated at the different times of the study, and a statis

one of two groups: treatment or control. Study participants received either a 10-day course of amoxicillin (an

gible patients were randomly and blindly assigned to the tical evaluation of the differences between the values

antibiotic) or a placebo similar in appearance and taste. The placebo consisted of symptomatic treatments

following two groups: group A (n = 46) (average age 35.93 obtained in T0, T1, T2, T3 and T4 in the two groups years, range 15–60), group B (n = 48) (average age studied was performed using an analysis of variance

such as acetaminophen, nasal decongestants, etc. At the end of the 10-day period, patients were asked if

33.2 years, range 16–58). (ANOVA) for repeated measures followed by multiple

they experienced improvement in symptoms. The distribution of responses is summarized below.[3]

Before enrollment, each patient was asked to give an informed consent to were performed using the Statistical Package for the Social participation in the study. Migraine intensity was measured by means of identify the tender points with Pain–Pressure Test (PPT). Every tender a VAS before applying NCT (T0). point located within the identified area by the pilot study (Fig. 1, area M) In group A, a specific algometer exerting a maximum pressure of 250 g was tested with NCT for 10 s (SEDATELEC, France) was chosen to

t test of Bonferroni to identify the source of variance.

Sciences (SPSS) software program. All values given in the following text are reported as arithmetic mean (±SEM).

Moreover, to evaluate the difference between group B and group A, a t test for unpaired data was always per

		Self-reported improvement in symptoms		
		Yes	No	Total
Group	Treatment	66	19	85
	Control	65	16	81
	Total	131	35	166

formed for each level of the variable "time". In the case of proportions, a Chi square ~~test was applied. All analyses~~

starting from the auricle, that was positive and the ipsilateral, to the side of prevalent cephalic pain. If the test was Results

(a) What percent of patients in the treatment group experienced improvement in symptoms?

reduction was at least 25% in respect to basis, a semi permanent needle (ASP SEDATELEC, France) was Only 89 patients out of the entire group of 94 (43 in group

(b) What percent experienced improvement in symptoms in the control group?

inserted after 1 min. On the contrary, if pain did not lessen A, 46 in group B) completed the experiment. Four patients

(c) In which group did a higher percentage of patients experience improvement in symptoms?

after 1 min, a further tender point was challenged in the withdrew from the study, because they experienced an same area and so on. When patients became aware of an unbearable exacerbation of pain in the period preceding the

(d) Your findings so far might suggest a real difference in effectiveness of antibiotic and placebo treatments

initial decrease in the pain in all the zones of the head last control at 24 h (two from group A and two from group

for improving symptoms of sinusitis. However, this is not the only possible conclusion that can be drawn

affected, they were invited to use a specific diary card to B) and were excluded from the statistical analysis since

based on your findings so far. What is one other possible explanation for the observed difference between

score the intensity of the pain with a VAS at the following they requested the removal of the needles. One patient intervals: after 10 min (T1), after 30 min (T2), after from group A did not give her consent to the implant of the

the percentages of patients in the antibiotic and placebo treatment groups that experience improvement

60 min (T3), after 120 min (T4), and after 24 h (T5). semi-permanent needles. In group A, the mean number of

in symptoms of sinusitis?

1.2 Data basics

Effective organization and description of data is a first step in most analyses. This section introduces the *data matrix* for organizing data as well as some terminology about different forms of data that will be used throughout this book.

1.2.1 Observations, variables, and data matrices

Figure 1.3 displays rows 1, 2, 3, and 50 of a data set for 50 randomly sampled loans offered through Lending Club, which is a peer-to-peer lending company. These observations will be referred to as the loan50 data set.

Each row in the table represents a single loan. The formal name for a row is a case or observational unit. The columns represent characteristics, called variables, for each of the loans. For example, the first row represents a loan of $7,500 with an interest rate of 7.34%, where the borrower is based in Maryland (MD) and has an income of $70,000.

GUIDED PRACTICE 1.2

What is the grade of the first loan in Figure 1.3? And what is the home ownership status of the borrower for that first loan? For these Guided Practice questions, you can check your answer in the footnote.[4]

In practice, it is especially important to ask clarifying questions to ensure important aspects of the data are understood. For instance, it is always important to be sure we know what each variable means and the units of measurement. Descriptions of the loan50 variables are given in Figure 1.4.

	loan amount	interest rate	term	grade	state	total income	homeownership
1	7500	7.34	36	A	MD	70000	rent
2	25000	9.43	60	B	OH	254000	mortgage
3	14500	6.08	36	A	MO	80000	mortgage
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	3000	7.96	36	A	CA	34000	rent

Figure 1.3: Four rows from the loan50 data matrix.

variable	description
loan amount	Amount of the loan received, in US dollars.
interest rate	Interest rate on the loan, in an annual percentage.
term	The length of the loan, which is always set as a whole number of months.
grade	Loan grade, which takes a values A through G and represents the quality of the loan and its likelihood of being repaid.
state	US state where the borrower resides.
total income	Borrower’s total income, including any second income, in US dollars.
homeownership	Indicates whether the person owns, owns but has a mortgage, or rents.

Figure 1.4: Variables and their descriptions for the loan50 data set.

[2] G. Allais et al. “Ear acupuncture in the treatment of migraine attacks: a randomized trial on the efficacy of appropriate versus inappropriate acupoints”. In: *Neurological Sci.* 32.1 (2011), pp. 173–175.

[3] J.M. Garbutt et al. “Amoxicillin for Acute Rhinosinusitis: A Randomized Controlled Trial”. In: *JAMA: The Journal of the American Medical Association* 307.7 (2012), pp. 685–692.

The data in Figure 1.3 represent a data matrix, which is a convenient and common way to organize data, especially if collecting data in a spreadsheet. Each row of a data matrix corresponds to a unique case (observational unit), and each column corresponds to a variable.

When recording data, use a data matrix unless you have a very good reason to use a different structure. This structure allows new cases to be added as rows or new variables as new columns.

GUIDED PRACTICE 1.3

The grades for assignments, quizzes, and exams in a course are often recorded in a gradebook that takes the form of a data matrix. How might you organize grade data using a data matrix?[5]

GUIDED PRACTICE 1.4

We consider data for 3,142 counties in the United States, which includes each county's name, the state where it resides, its population in 2017, how its population changed from 2010 to 2017, poverty rate, and six additional characteristics. How might these data be organized in a data matrix?[6]

The data described in Guided Practice 1.4 represents the county data set, which is shown as a data matrix in Figure 1.5. The variables are summarized in Figure 1.6.

[4] The loan's grade is A, and the borrower rents their residence.

[5] There are multiple strategies that can be followed. One common strategy is to have each student represented by a row, and then add a column for each assignment, quiz, or exam. Under this setup, it is easy to review a single line to understand a student's grade history. There should also be columns to include student information, such as one column to list student names.

[6] Each county may be viewed as a case, and there are eleven pieces of information recorded for each case. A table with 3,142 rows and 11 columns could hold these data, where each row represents a county and each column represents a particular piece of information.

e
m
o
c
n
i

h h

n
a
i
d
me

u
d
e

n a

i d e

o r t e

e t a r p m

e n

t i n u i t

l u

p i h s r e n w o e m o

7 1

3 5

e g e l l o c e m

s

6 8

2

5 . 7

2 6

5 2

e g e l l o c e m

s

9

3.9

6

.

22

7

.

76

8

6

3

33

a

m

o

l

p

i

d

hs no

0

5.9

1

.

11

0

.

68

4

0

4

43

a

m

o

l

p

i

d

hs yes

9

4.3

6.6

9

.

82

2

1

4

47

a

m

o

l

p

i

d

,s ,,s

2

,0

,7

0

,2

5

5

6

,9

a

m

o

l

p

i

d

,s ,o

3

,9

,9

9

,6

6

2

3

,6

a

m

o

l

p

i

d

,s ,o

9

,4

7

,3

0

.

69

6 8

6

43

e

g

e

l

l

o

c

e

som

yes

3

4.9

3

.

4

7

.

70

2

4

3

37

a

m

o

l

p

i

d

hs no

8

4.0

4.7

4

.

41

1

4

0

40

a

m

o

l

p
i
d
‚a ‚o

5
‚0

‚3

5
.
‚7
...

...

... ...

... ...

5 0

6
‚9

e
g
e
l
l
o
c
e
‚‚m

‚o

8
9
5
9
.
7
.
t
a
t
a
d
y
t
n
u
o
c
e
m
o
s
w
n
e
.
a
i
b
m
u
l
o
f
t
c
i
r
t
s
i
e
f
.
s
e
d
i
s
y
t
n
u
.
e
u
l
e
.
e
l
p
m
a
r
o
.
7
1
0
o
0
1
0
m
o
n
o
i
t
a
l
u
p
o
e
n
%
8
4
y

d e s a e r inc y t n cou s thi for n o i t a l u p po the s n a me row . . y t r e v po in n o i t a l u , r e n ow the h t wi s e llv or e m ho n ow r i the in s e llv t tha n o i t a l u . e m ho the n ow o wh s t n e r pa h t wi . s t n e m t r a ap . eg , s e r utcustrtinu-itlmuinaretthastinun.tnecrapeaserat.aarenatiloportamesniatconyt,amolpidsh,shwolebgnomameuavaletakcanhciwh,lelev.srolehcabslaequemoincs

t
p
i
.
e
m
na
y
the
e
r
e
wh
7
1 20 in
n
o
it
a
e
g
n
cha
t
n
t firs
the
n
i
7
1 20 to
0
1 20
p
po
the of
t
n
p
po
the of
t
n
g
n
i
liv
n
e
r
d
l chi
g
n
i
liv of
t
n
t
n
e
m
y
o
lp
n
cou
the
r
e
h
n
o
it
a
c
u
edn
d
an
d
l
o
,
h
e
e
s
g
u
ho
e
l
l
o
n
c
e
m
o
inc
l
a
tot
d an
s
e
l
b
a
i
r
va
e
g
u
g
iF
r

c

s

de
tn
u

o
e

t

a

t
lu
p

o
e

cr
e
8 4

.
m

or
e

c

r

e

P
e

c

r

e

P
.

g.

e
e

cr
e
me

n
te

hW
a

i

de
e m

o

s
a

i

de
e

h
:

1.6
n

a

h

c

p po

p
po
8 1.4

40
5 55
9 9.1

826
2 21
2

2 -6-

07
2 25
3 0.7

86
6 62
8 0.6

31
0 08
8

2 -6-

90
3 00
9

6 -6-

52
8 09
1

5 -6-

827
4 11
0

2 -6-

31
7,3
0

6

75
8,5

3

9

7

2,9
e

l

ba
C
S

e
P
P

e g

n

a

h

c
1
f

y t

r
p

i

h

s

r

e

n

w

o
P

t i

n

u

i
U

e t

a

r

p
o
M

u d

e

n a

M

e m

o

c

n

i

h h

n a

t

e

r

u

g

iF

e

t

a$_s$t

eman
a

m

a

b

a

$_A$l

a

g

u

a

t

$_A$u1

a

m

a

b

a

$_A$l

n

i

w

d

l$_B$a

2
a

m

a

b

a

$_A$l

r

u

o

b

r

$_B$a3

a

m

a

b

a

$_A$l

… 10 Cherokee Alabama 9 Chambers Alabama 8 Calhoun Alabama 7 Butler Alabama 6 Bullock Alabama 5 Blount Alabama 4 Bibb

1.2.2 Types of variables

Examine the unemp rate, pop, state, and median edu variables in the county data set. Each of these variables is inherently different from the other three, yet some share certain characteristics. First consider unemp rate, which is said to be a numerical variable since it can take a wide range of numerical values, and it is sensible to add, subtract, or take averages with those values. On the other hand, we would not classify a variable reporting telephone area codes as numerical since the average, sum, and difference of area codes doesn't have any clear meaning.

The pop variable is also numerical, although it seems to be a little different than unemp rate. This variable of the population count can only take whole non-negative numbers (0, 1, 2, ...). For this reason, the population variable is said to be discrete since it can only take numerical values with jumps. On the other hand, the unemployment rate variable is said to be continuous.

The variable state can take up to 51 values after accounting for Washington, DC: AL, AK, ..., and WY. Because the responses themselves are categories, state is called a categorical variable, and the possible values are called the variable's levels.

Finally, consider the median edu variable, which describes the median education level of county residents and takes values below hs, hs diploma, some college, or bachelors in each county. This variable seems to be a hybrid: it is a categorical variable but the levels have a natural ordering. A variable with these properties is called an ordinal variable, while a regular categorical variable without this type of special ordering is called a nominal variable. To simplify analyses, any ordinal variable in this book will be treated as a nominal (unordered) categorical variable.

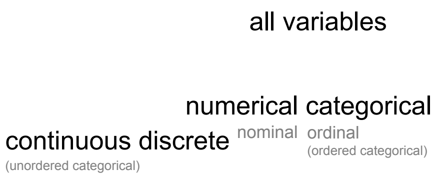


Figure 1.7: Breakdown of variables into their respective types.

EXAMPLE 1.5

Data were collected about students in a statistics course. Three variables were recorded for each student: number of siblings, student height, and whether the student had previously taken a statistics course. Classify each of the variables as continuous numerical, discrete numerical, or categorical.

The number of siblings and student height represent numerical variables. Because the number of siblings is a count, it is discrete. Height varies continuously, so it is a continuous numerical variable. The last variable classifies students into two categories – those who have and those who have not taken a statistics course – which makes this variable categorical.

GUIDED PRACTICE 1.6

An experiment is evaluating the effectiveness of a new drug in treating migraines. A group variable is used to indicate the experiment group for each patient: treatment or control. The num migraines variable represents the number of migraines the patient experienced during a 3-month period. Classify each variable as either numerical or categorical?[7]

[7] There group variable can take just one of two group names, making it categorical. The num migraines variable describes a count of the number of migraines, which is an outcome where basic arithmetic is sensible, which means this is numerical outcome; more specifically, since it represents a count, num migraines is a discrete numerical variable.

1.2.3 Relationships between variables

Many analyses are motivated by a researcher looking for a relationship between two or more variables. A social scientist may like to answer some of the following questions:

(1) If homeownership is lower than the national average in one county, will the percent of multi-unit structures in that county tend to be above or below the national average?

(2) Does a higher than average increase in county population tend to correspond to counties with higher or lower median household incomes?

(3) How useful a predictor is median education level for the median household income for US counties?

To answer these questions, data must be collected, such as the county data set shown in Figure 1.5. Examining summary statistics could provide insights for each of the three questions about counties. Additionally, graphs can be used to visually explore data.

Scatterplots are one type of graph used to study the relationship between two numerical variables. Figure 1.8 compares the variables homeownership and multi unit, which is the percent of units in multi-unit structures (e.g. apartments, condos). Each point on the plot represents a single county. For instance, the highlighted dot corresponds to County 413 in the county data set: Chattahoochee County, Georgia, which has 39.4% of units in multi-unit structures and a homeownership rate of 31.3%. The scatterplot suggests a relationship between the two variables:

counties with a higher rate of multi-units tend to have lower homeownership rates. We might brainstorm as to why this relationship exists and investigate each idea to determine which are the most reasonable explanations.

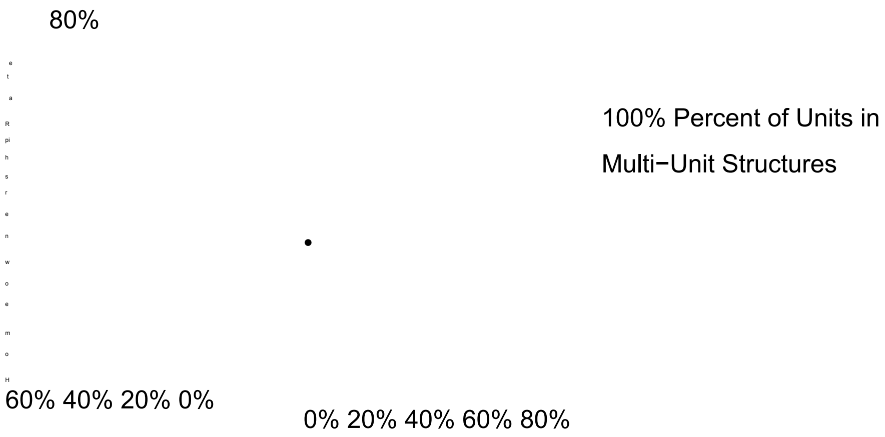


Figure 1.8: A scatterplot of homeownership versus the percent of units that are in multi-unit structures for US counties. The highlighted dot represents Chatta hoochee County, Georgia, which has a multi-unit rate of 39.4% and a homeowner ship rate of 31.3%.

The multi-unit and homeownership rates are said to be associated because the plot shows a discernible pattern. When two variables show some connection with one another, they are called associated variables. Associated variables can also be called dependent variables and vice-versa.

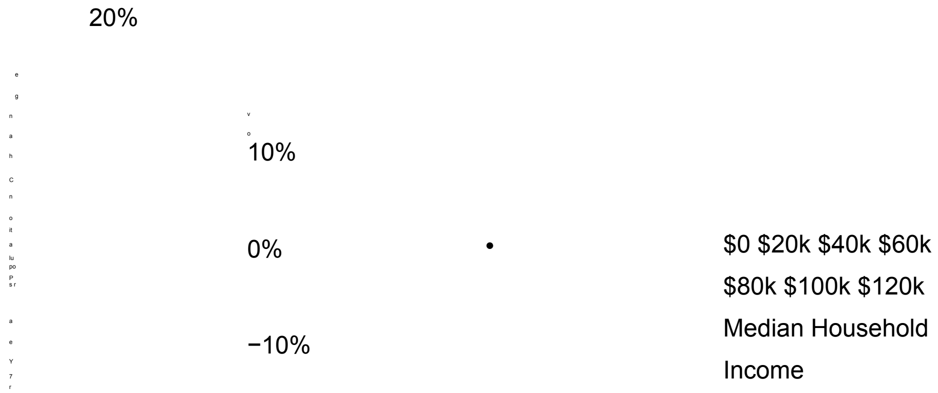


Figure 1.9: A scatterplot showing pop change against median hh income. Owsley County of Kentucky, is highlighted, which lost 3.63% of its population from 2010 to 2017 and had median household income of $22,736.

GUIDED PRACTICE 1.7

Examine the variables in the loan50 data set, which are described in Figure 1.4 on page 12. Create two questions about possible relationships between variables in loan50 that are of interest to you.[8]

EXAMPLE 1.8

This example examines the relationship between a county's population change from 2010 to 2017 and median household income, which is visualized as a scatterplot in Figure 1.9. Are these variables associated?

The larger the median household income for a county, the higher the population growth observed for the county. While this trend isn't true for every county, the trend in the plot is evident. Since there is some relationship between the variables, they are associated.

Because there is a downward trend in Figure 1.8 – counties with more units in multi-unit structures are associated with lower homeownership – these variables are said to be negatively associated. A positive association is shown in the relationship between the median hh income and pop change in Figure 1.9, where counties with higher median household income tend to have higher rates of population growth.

If two variables are not associated, then they are said to be independent. That is, two variables are independent if there is no evident relationship between the two.

ASSOCIATED OR INDEPENDENT, NOT BOTH

A pair of variables are either related in some way (associated) or not (independent). No pair of variables is both associated and independent.

[8] Two example questions: (1) What is the relationship between loan amount and total income? (2) If someone's income is above the average, will their interest rate tend to be above or below the average?

1.2.4 Explanatory and response variables

When we ask questions about the relationship between two variables, we sometimes also want to determine if the change in one variable causes a change in the other. Consider the following rephrasing of an earlier question about the county data set:

> *If there is an increase in the median household income in a county, does this drive an increase in its population?*

In this question, we are asking whether one variable affects another. If this is our underlying belief, then *median household income* is the explanatory variable and the *population change* is the response variable in the hypothesized relationship.[9]

EXPLANATORY AND RESPONSE VARIABLES

When we suspect one variable might causally affect another, we label the first variable the explanatory variable and the second the response variable.

response variable
explanatory might affect
variable

For many pairs of variables, there is no hypothesized relationship, and these labels would not be applied to either variable in such cases.

Bear in mind that the act of labeling the variables in this way does nothing to guarantee that a causal relationship exists. A formal evaluation to check whether one variable causes a change in another requires an experiment.

1.2.5 Introducing observational studies and experiments

There are two primary types of data collection: observational studies and experiments. Researchers perform an observational study when they collect data in a way that does not directly interfere with how the data arise. For instance, researchers may collect information via surveys, review medical or company records, or follow a cohort of many similar individuals to form hypotheses about why certain diseases might develop. In each of these situations, researchers merely observe the data that arise. In general, observational studies can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection. When researchers want to investigate the possibility of a causal connection, they conduct an experiment. Usually there will be both an explanatory and a response variable. For instance, we may suspect administering a drug will reduce mortality in heart attack patients over the following year. To check if there really is a causal connection between the explanatory variable and the response, researchers will collect a sample of individuals and split them into groups. The individuals in each group are *assigned* a treatment. When individuals are randomly assigned to a group, the experiment is called a randomized experiment. For example, each heart attack patient in the drug trial could be randomly assigned, perhaps by flipping a coin, into one of two groups: the first group receives a placebo (fake treatment) and the second group receives the drug. See the case study in Section 1.1 for another example of an experiment, though that study did not employ a placebo.

ASSOCIATION 6= CAUSATION

In general, association does not imply causation, and causation can only be inferred from a randomized experiment.

[9] Sometimes the explanatory variable is called the independent variable and the response variable is called the dependent variable. However, this becomes confusing since a *pair* of variables might be independent or dependent, so we avoid this language.

Exercises

1.3 Air pollution and birth outcomes, study components. Researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Specifically, levels of carbon monoxide were recorded in parts per million, nitrogen dioxide and ozone in parts per hundred million, and coarse particulate matter (PM_{10}) in $\mu g/m^3$. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth. The analysis suggested that increased ambient PM_{10} and, to a lesser degree, CO concentrations may be associated with the occurrence of preterm births.[10]

(a) Identify the main research question of the study.

(b) Who are the subjects in this study, and how many are included?

(c) What are the variables in the study? Identify each variable as numerical or categorical. If numerical, state whether the variable is discrete or continuous. If categorical, state whether the variable is ordinal.

1.4 Buteyko method, study components. The Buteyko method is a shallow breathing technique developed by Konstantin Buteyko, a Russian doctor, in 1952. Anecdotal evidence suggests that the Buteyko method can reduce asthma symptoms and improve quality of life. In a scientific study to determine the effectiveness of this method, researchers recruited 600 asthma patients aged 18-69 who relied on medication for asthma treatment. These patients were randomnly split into two research groups: one practiced the Buteyko method and the other did not. Patients were scored on quality of life, activity, asthma symptoms, and medication reduction on a scale from 0 to 10. On average, the participants in the Buteyko group experienced a significant reduction in asthma symptoms and an improvement in quality of life.[11]

(a) Identify the main research question of the study.

(b) Who are the subjects in this study, and how many are included?

(c) What are the variables in the study? Identify each variable as numerical or categorical. If numerical, state whether the variable is discrete or continuous. If categorical, state whether the variable is ordinal.

1.5 Cheaters, study components. Researchers studying the relationship between honesty, age and self control conducted an experiment on 160 children between the ages of 5 and 15. Participants reported their age, sex, and whether they were an only child or not. The researchers asked each child to toss a fair coin in private and to record the outcome (white or black) on a paper sheet, and said they would only reward children who report white. The study's findings can be summarized as follows: "Half the students were explicitly told not to cheat and the others were not given any explicit instructions. In the no instruction group probability of cheating was found to be uniform across groups based on child's characteristics. In the group that was explicitly told to not cheat, girls were less likely to cheat, and while rate of cheating didn't vary by age for boys, it decreased with age for girls."[12]

(a) Identify the main research question of the study.

(b) Who are the subjects in this study, and how many are included?

(c) How many variables were recorded for each subject in the study in order to conclude these findings? State the variables and their types.

[10] B. Ritz et al. "Effect of air pollution on preterm birth among children born in Southern California between 1989 and 1993". In: *Epidemiology* 11.5 (2000), pp. 502–511.

[11] J. McGowan. "Health Education: Does the Buteyko Institute Method make a difference?" In: *Thorax* 58 (2003).

[12] Alessandro Bucciol and Marco Piovesan. "Luck or cheating? A field experiment on honesty with children". In: *Journal of Economic Psychology* 32.1 (2011), pp. 73–78.

1.6 Stealers, study components. In a study of the relationship between socio-economic class and unethical behavior, 129 University of California undergraduates at Berkeley were asked to identify themselves as having low or high social-class by comparing themselves to others with the most (least) money, most (least) education, and most (least) respected jobs. They were also presented with a jar of individually wrapped candies and informed that the candies were for children in a nearby laboratory, but that they could take some if they wanted. After completing some unrelated tasks, participants reported the number of candies they had taken.[13]

(a) Identify the main research question of the study.

(b) Who are the subjects in this study, and how many are included?

(c) The study found that students who were identified as upper-class took more candy than others. How many variables were recorded for each subject in the study in order to conclude these findings? State the variables and their types.

1.7 Migraine and acupuncture, Part II. Exercise 1.1 introduced a study exploring whether acupuncture had any effect on migraines. Researchers conducted a randomized controlled study where patients were randomly assigned to one of two groups: treatment or control. The patients in the treatment group re ceived acupuncture that was specifically designed to treat migraines. The patients in the control group received placebo acupuncture (needle insertion at non-acupoint locations). 24 hours after patients received acupuncture, they were asked if they were pain free. What are the explanatory and response variables in this study?

1.8 Sinusitis and antibiotics, Part II. Exercise 1.2 introduced a study exploring the effect of antibiotic treatment for acute sinusitis. Study participants either received either a 10-day course of an antibiotic (treatment) or a placebo similar in appearance and taste (control). At the end of the 10-day period, patients were asked if they experienced improvement in symptoms. What are the explanatory and response variables in this study?

1.9 Fisher's irises. Sir Ronald Aylmer Fisher was an English statistician, evolutionary biologist, and geneticist who worked on a data set that contained sepal length and width, and petal length and width from three species of iris flowers (*setosa*, *versicolor* and *virginica*). There were 50 flowers from each species in the data set.[14]

(a) How many cases were included in the data?

(b) How many numerical variables are included in the data? Indicate what they are, and if they are continuous or discrete.

(c) How many categorical variables are included in the data, and what are they? List the corre sponding

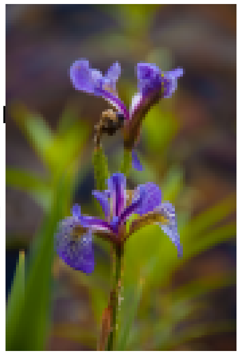


Photo by Ryan Claussen (http://flic.kr/p/6QTcuX) CC BY-SA 2.0 license

1.10 Smoking habits of UK residents. A survey was conducted to study the smoking habits of UK residents. Below is a data matrix displaying a portion of the data collected in this survey. Note that "£" stands for British Pounds Sterling, "cig" stands for cigarettes, and "N/A" refers to a missing component of the data.[15]

	sex	age	marital	grossIncome	smoke	amtWeekends	amtWeekdays
1	Female	42	Single	Under £2,600	Yes	12 cig/day	12 cig/day
2	Male	44	Single	£10,400 to £15,600	No	N/A	N/A
3	Male	53	Married	Above £36,400	Yes	6 cig/day	6 cig/day
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1691	Male	40	Single	£2,600 to £5,200	Yes	8 cig/day	8 cig/day

(a) What does each row of the data matrix represent?

(b) How many participants were included in the survey?

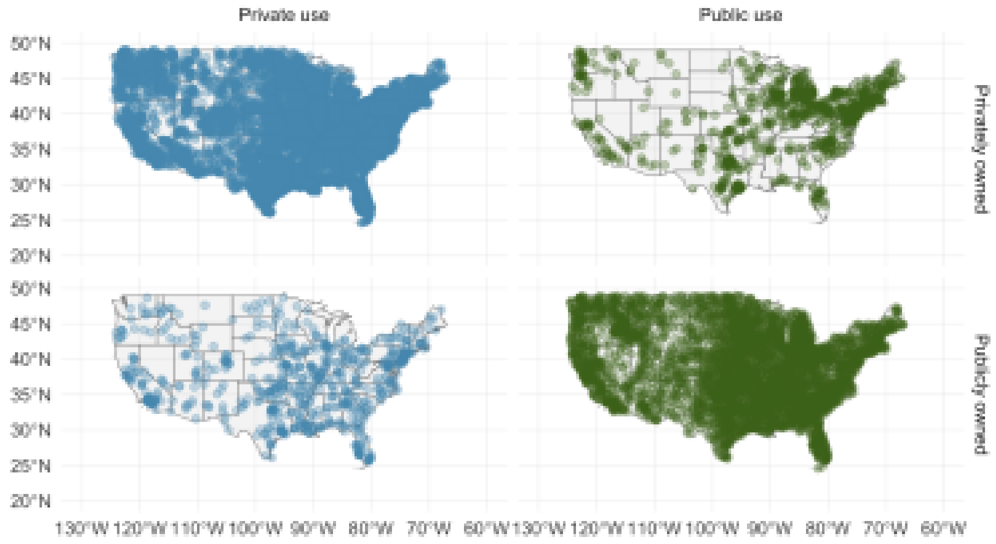
(c) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as contin uous or discrete. If categorical, indicate if the variable is ordinal.

[13] P.K. Piff et al. "Higher social class predicts increased unethical behavior". In: *Proceedings of the National Academy of Sciences* (2012).

[14] R.A Fisher. "The Use of Multiple Measurements in Taxonomic Problems". In: *Annals of Eugenics* 7 (1936), pp. 179–188.

[15] National STEM Centre, Large Datasets from stats4schools.

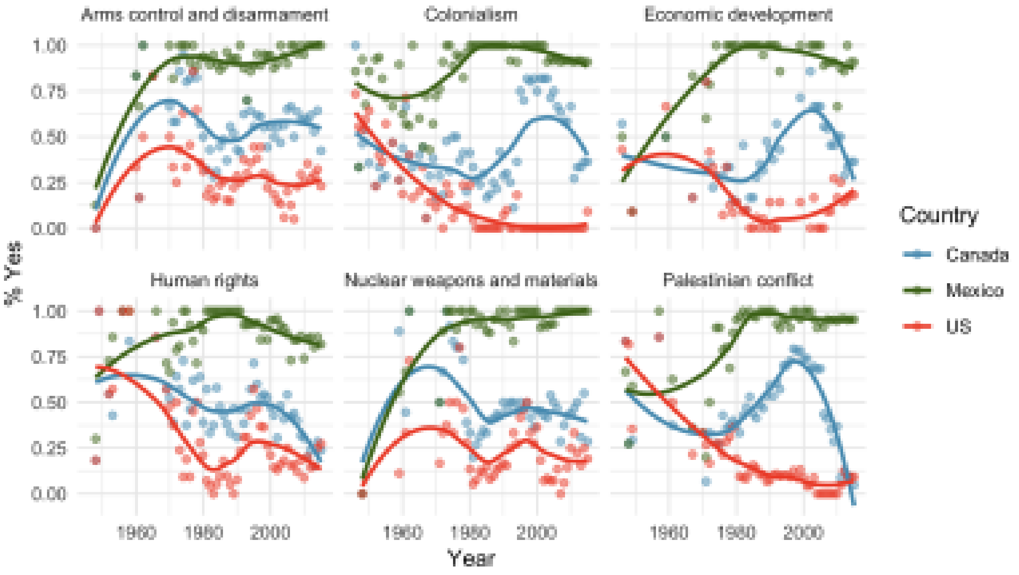
1.11 US Airports. The visualization below shows the geographical distribution of airports in the contiguous United States and Washington, DC. This visualization was constructed based on a dataset where each observation is an airport.[16]


Private use
Public use
Privately owned
Publicly owned
50°N
45°N
40°N
35°N
30°N
25°N
20°N
130°W 120°W 110°W 100°W 90°W 80°W 70°W 60°W
130°W 120°W 110°W 100°W 90°W 80°W 70°W 60°W

(a) List the variables used in creating this visualization.

(b) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as contin uous or discrete. If categorical, indicate if the variable is ordinal.

1.12 UN Votes. The visualization below shows voting patterns in the United States, Canada, and Mexico in the United Nations General Assembly on a variety of issues. Specifically, for a given year between 1946 and 2015, it displays the percentage of roll calls in which the country voted yes for each issue. This visualization was constructed based on a dataset where each observation is a country/year pair.[17]



(a) List the variables used in creating this visualization.

(b) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as contin uous or discrete. If categorical, indicate if the variable is ordinal.

[16] Federal Aviation Administration, www.faa.gov/airports/airport safety/airportdata 5010.

[17] David Robinson. *unvotes: United Nations General Assembly Voting Data*. R package version 0.2.0. 2017. url: https://CRAN.R-project.org/package=unvotes.

1.3 Sampling principles and

strategies

The first step in conducting research is to identify topics or questions that are to be investigated. A clearly laid out research question is helpful in identifying what subjects or cases should be studied and what variables are important. It is also important to consider *how* data are collected so that they are reliable and help achieve the research goals.

1.3.1 Populations and samples

Consider the following three research questions:

1. What is the average mercury content in swordfish in the Atlantic Ocean?
2. Over the last 5 years, what is the average time to complete a degree for Duke undergrads? 3. Does a new drug reduce the number of deaths in patients with severe heart disease?

Each research question refers to a target population. In the first question, the target population is all swordfish in the Atlantic ocean, and each fish represents a case. Often times, it is too expensive to collect data for every case in a population. Instead, a sample is taken. A sample represents a subset of the cases and is often a small fraction of the population. For instance, 60 swordfish (or some other number) in the population might be selected, and this sample data may be used to provide an estimate of the population average and answer the research question.

GUIDED PRACTICE 1.9

For the second and third questions above, identify the target population and what represents an individual case.[18]

1.3.2 Anecdotal evidence

Consider the following possible responses to the three research questions:

1. A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.
2. I met two students who took more than 7 years to graduate from Duke, so it must take longer to graduate at Duke than at many other colleges.
3. My friend’s dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

Each conclusion is based on data. However, there are two problems. First, the data only represent one or two cases. Second, and more importantly, it is unclear whether these cases are actually representative of the population. Data collected in this haphazard fashion are called anecdotal evidence.

ANECDOTAL EVIDENCE

Be careful of data collected in a haphazard fashion. Such evidence may be true and verifiable, but it may only represent extraordinary cases.

[18](2) The first question is only relevant to students who complete their degree; the average cannot be computed using a student who never finished her degree. Thus, only Duke undergrads who graduated in the last five years represent

under consideration. Each such student is an individual case. (3) A person with severe heart . The population includes all people with severe heart disease.

Figure 1.10: In February 2010, some media pundits cited one large snow storm as valid evidence against global warming. As comedian Jon Stewart pointed out, “It’s one storm, in one region, of one country.”

Anecdotal evidence typically is composed of unusual cases that we recall based on their striking characteristics. For instance, we are more likely to remember the two people we met who took 7 years to graduate than the six others who graduated in four years. Instead of looking at the most unusual cases, we should examine a sample of many cases that represent the population.

1.3.3 Sampling from a population

We might try to estimate the time to graduation for Duke undergraduates in the last 5 years by collecting a sample of students. All graduates in the last 5 years represent the *population*, and graduates who are selected for review are collectively called the *sample*. In general, we always seek to *randomly* select a sample from a population. The most basic type of random selection is equivalent to how raffles are conducted. For example, in selecting graduates, we could write each graduate’s name on a raffle ticket and draw 100 tickets. The selected names would represent a random sample of 100 graduates. We pick samples randomly to reduce the chance we introduce biases.

all graduates

sample

Figure 1.11: In this graphic, five graduates are randomly selected from the popu lation to be included in the sample.

EXAMPLE 1.10

Suppose we ask a student who happens to be majoring in nutrition to select several graduates for the study. What kind of students do you think she might collect? Do you think her sample would be representative of all graduates?

Perhaps she would pick a disproportionate number of graduates from health-related fields. Or perhaps her selection would be a good representation of the population. When selecting samples by hand, we run the risk of picking a biased sample, even if their bias isn't intended.

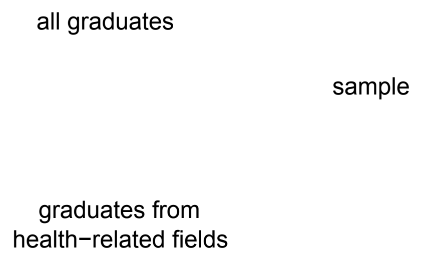


Figure 1.12: Asked to pick a sample of graduates, a nutrition major might inadvertently pick a disproportionate number of graduates from health-related majors.

If someone was permitted to pick and choose exactly which graduates were included in the sample, it is entirely possible that the sample could be skewed to that person's interests, which may be entirely unintentional. This introduces bias into a sample. Sampling randomly helps resolve this problem. The most basic random sample is called a simple random sample, and which is equivalent to using a raffle to select cases. This means that each case in the population has an equal chance of being included and there is no implied connection between the cases in the sample.

The act of taking a simple random sample helps minimize bias. However, bias can crop up in other ways. Even when people are picked at random, e.g. for surveys, caution must be exercised if the non-response rate is high. For instance, if only 30% of the people randomly sampled for a survey actually respond, then it is unclear whether the results are representative of the entire population. This non-response bias can skew results.

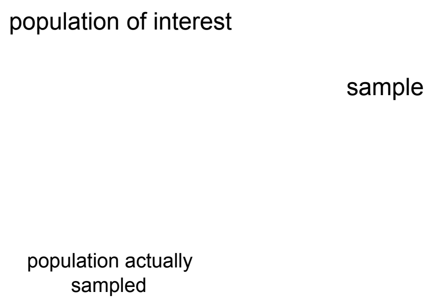


Figure 1.13: Due to the possibility of non-response, surveys studies may only reach a certain group within the population. It is difficult, and often times impossible, to completely fix this problem.

Another common downfall is a convenience sample, where individuals who are easily accessible are more likely to be included in the sample. For instance, if a political survey is done by stopping people walking in the Bronx, this will not represent all of New York City. It is often difficult to discern what sub-population a convenience sample represents.

GUIDED PRACTICE 1.11

We can easily access ratings for products, sellers, and companies through websites. These ratings are based only on those people who go out of their way to provide a rating. If 50% of online reviews for a product are negative, do you think this means that 50% of buyers are dissatisfied with the product?[19]

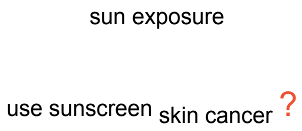
1.3.4 Observational studies

Data where no treatment has been explicitly applied (or explicitly withheld) is called observational data. For instance, the loan data and county data described in Section 1.2 are both examples of observational data. Making causal conclusions based on experiments is often reasonable. However, making the same causal conclusions based on observational data can be treacherous and is not recommended. Thus, observational studies are generally only sufficient to show associations or form hypotheses that we later check using experiments.

GUIDED PRACTICE 1.12

Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen *causes* skin cancer?[20]

Some previous research tells us that using sunscreen actually reduces skin cancer risk, so maybe there is another variable that can explain this hypothetical association between sunscreen usage and skin cancer. One important piece of information that is absent is sun exposure. If someone is out in the sun all day, she is more likely to use sunscreen *and* more likely to get skin cancer. Exposure to the sun is unaccounted for in the simple investigation.



Sun exposure is what is called a confounding variable,[21] which is a variable that is correlated with both the explanatory and response variables. While one method to justify making causal conclusions from observational studies is to exhaust the search for confounding variables, there is no guarantee that all confounding variables can be examined or measured.

GUIDED PRACTICE 1.13

Figure 1.8 shows a negative association between the homeownership rate and the percentage of multi unit structures in a county. However, it is unreasonable to conclude that there is a causal relationship between the two variables. Suggest a variable that might explain the negative relationship.[22]

Observational studies come in two forms: prospective and retrospective studies. A prospective study identifies individuals and collects information as events unfold. For instance, medical researchers may identify and follow a group of patients over many years to assess the possible influences of behavior on cancer risk. One example of such a study is The Nurses' Health Study, started in 1976 and expanded in 1989. This prospective study recruits registered nurses and then collects data from them using questionnaires. Retrospective studies collect data after events have taken place, e.g. researchers may review past events in medical records. Some data sets may contain both prospectively- and retrospectively-collected variables.

1.3.5 Four sampling methods

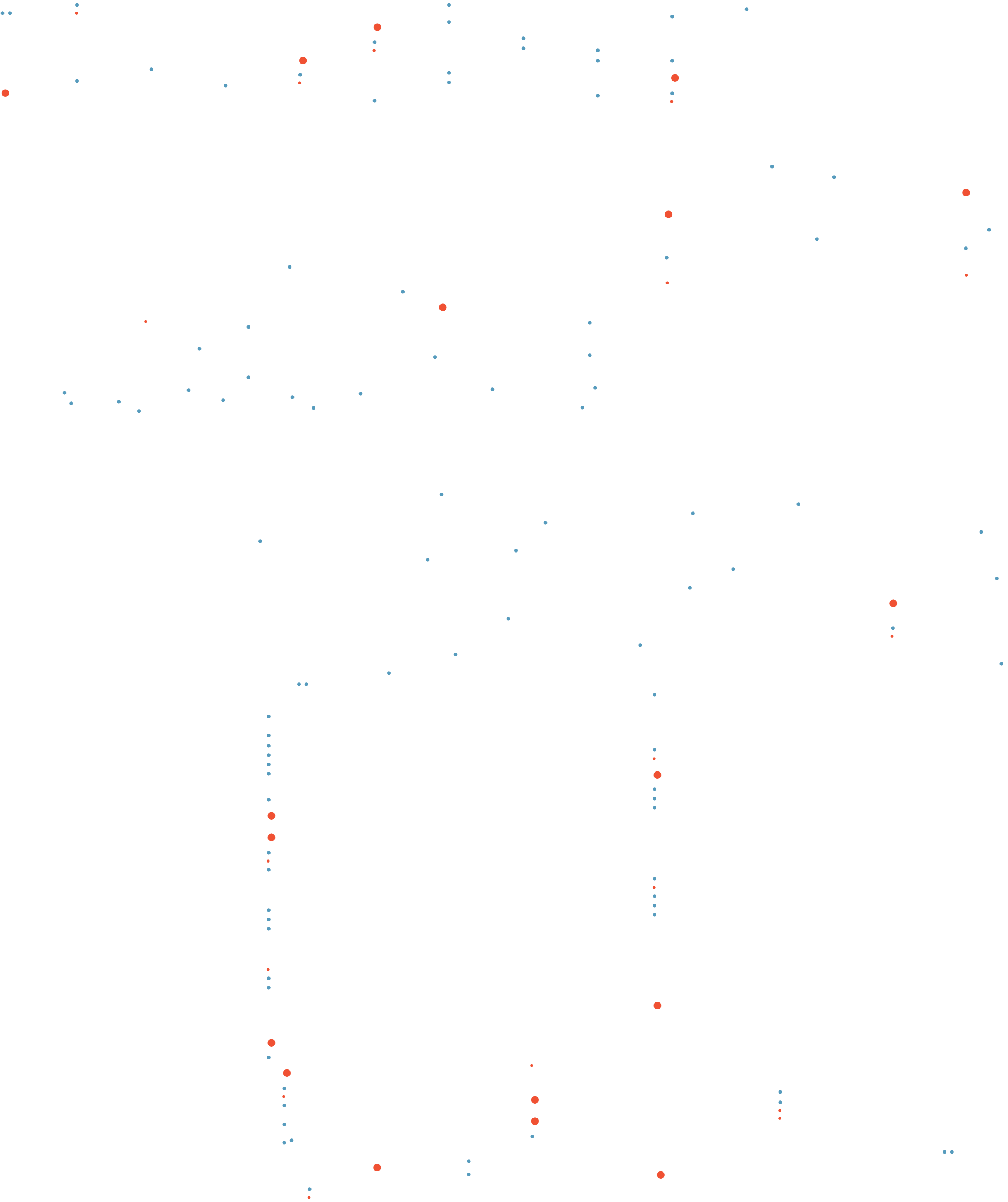
[19] Answers will vary. From our own anecdotal experiences, we believe people tend to rant more about products that fell below expectations than rave about those that perform as expected. For this reason, we suspect there is a negative bias in product ratings on sites like Amazon. However, since our experiences may not be representative, we also keep an open mind.

Almost all statistical methods are based on the notion of implied randomness. If observational data are not collected in a random framework from a population, these statistical methods – the estimates and errors associated with the estimates – are not reliable. Here we consider four random sampling techniques: simple, stratified, cluster, and multistage sampling. Figures 1.14 and 1.15 provide graphical representations of these techniques.

[20] No. See the paragraph following the exercise for an explanation.

[21] Also called a lurking variable, confounding factor, or a confounder.

[22] Answers will vary. Population density may be important. If a county is very dense, then this may require a larger fraction of residents to live in multi-unit structures. Additionally, the high density may contribute to increases in property value, making homeownership infeasible for many residents.



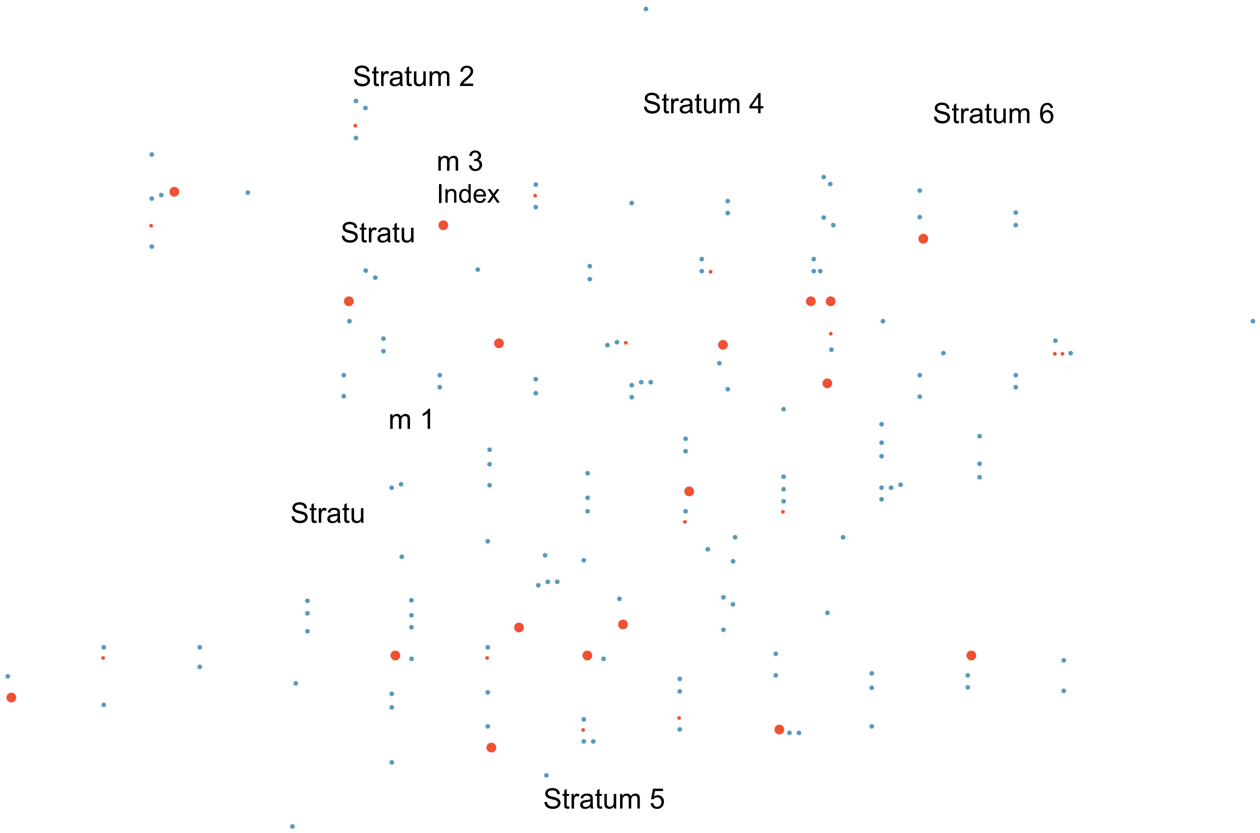


Figure 1.14: Examples of simple random and stratified sampling. In the top panel, simple random sampling was used to randomly select the 18 cases. In the bottom panel, stratified sampling was used: cases were grouped into strata, then simple random sampling was employed within each stratum.

Simple random sampling is probably the most intuitive form of random sampling. Consider the salaries of Major League Baseball (MLB) players, where each player is a member of one of the league's 30 teams. To take a simple random sample of 120 baseball players and their salaries, we could write the names of that season's several hundreds of players onto slips of paper, drop the slips into a bucket, shake the bucket around until we are sure the names are all mixed up, then draw out slips until we have the sample of 120 players. In general, a sample is referred to as “simple random” if each case in the population has an equal chance of being included in the final sample *and* knowing that a case is included in a sample does not provide useful information about which other cases are included.

Stratified sampling is a divide-and-conquer sampling strategy. The population is divided into groups called strata. The strata are chosen so that similar cases are grouped together, then a second sampling method, usually simple random sampling, is employed within each stratum. In the baseball salary example, the teams could represent the strata, since some teams have a lot more money (up to 4 times as much!). Then we might randomly sample 4 players from each team for a total of 120 players.

Stratified sampling is especially useful when the cases in each stratum are very similar with respect to the outcome of interest. The downside is that analyzing data from a stratified sample is a more complex task than analyzing data from a simple random sample. The analysis methods introduced in this book would need to be extended to analyze data collected using stratified sampling.

EXAMPLE 1.14

Why would it be good for cases within each stratum to be very similar?

We might get a more stable estimate for the subpopulation in a stratum if the cases are very similar, leading to more precise estimates within each group. When we combine these estimates into a single estimate for the full population, that population estimate will tend to be more precise since each individual group estimate is itself more precise.

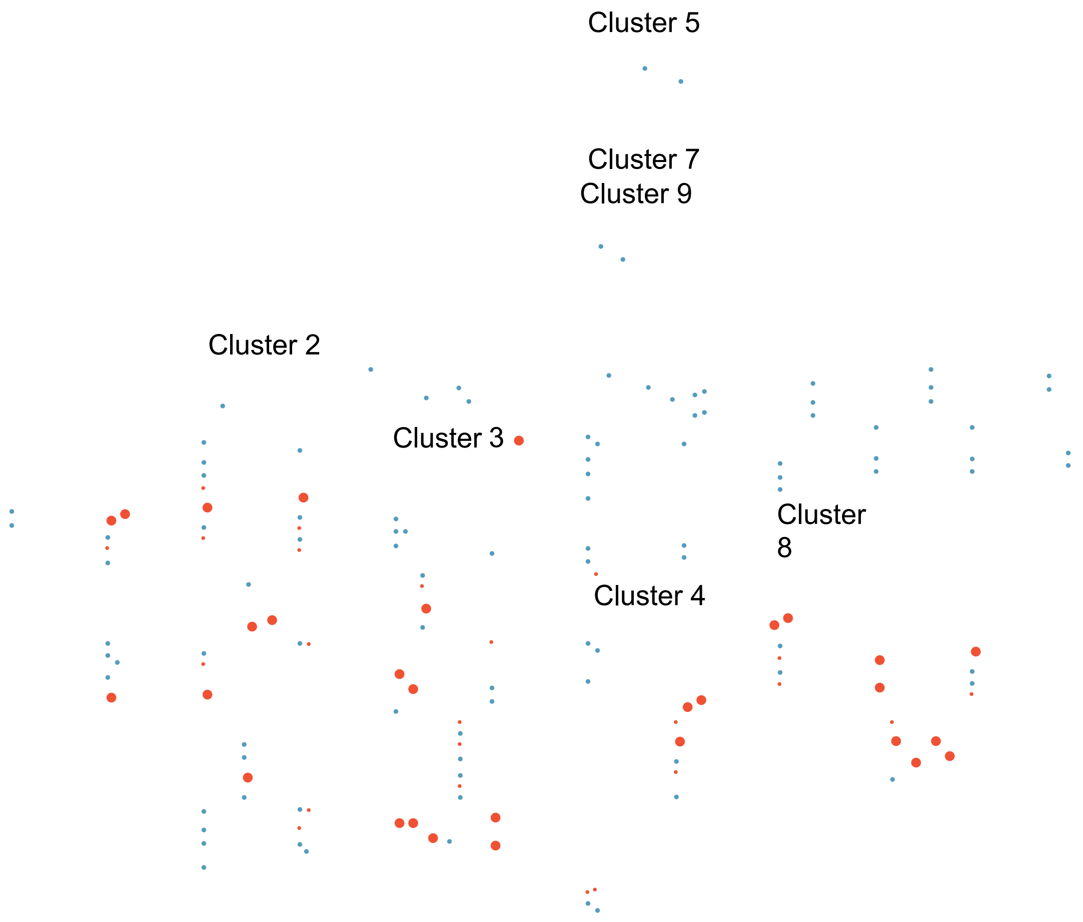
In a cluster sample, we break up the population into many groups, called clusters. Then we sample a fixed number of clusters and include all observations from each of those clusters in the sample. A multistage sample is like a cluster sample, but rather than keeping all observations in each cluster, we collect a random sample within each selected cluster.

Sometimes cluster or multistage sampling can be more economical than the alternative sampling techniques. Also, unlike stratified sampling, these approaches are most helpful when there is a lot of case-to-case variability within a cluster but the clusters themselves don't look very different from one another. For example, if neighborhoods represented clusters, then cluster or multistage sampling work best when the neighborhoods are very diverse. A downside of these methods is that more advanced techniques are typically required to analyze the data, though the methods in this book can be extended to handle such data.

EXAMPLE 1.15

Suppose we are interested in estimating the malaria rate in a densely tropical portion of rural Indonesia. We learn that there are 30 villages in that part of the Indonesian jungle, each more or less similar to the next. Our goal is to test 150 individuals for malaria. What sampling method should be employed?

A simple random sample would likely draw individuals from all 30 villages, which could make data collection extremely expensive. Stratified sampling would be a challenge since it is unclear how we would build strata of similar individuals. However, cluster sampling or multistage sampling seem like very good ideas. If we decided to use multistage sampling, we might randomly select half of the villages, then randomly select 10 people from each. This would probably reduce our data collection costs substantially in comparison to a simple random sample, and the cluster sample would still give us reliable information, even if we would need to analyze the data with slightly more advanced methods than we discuss in this book.



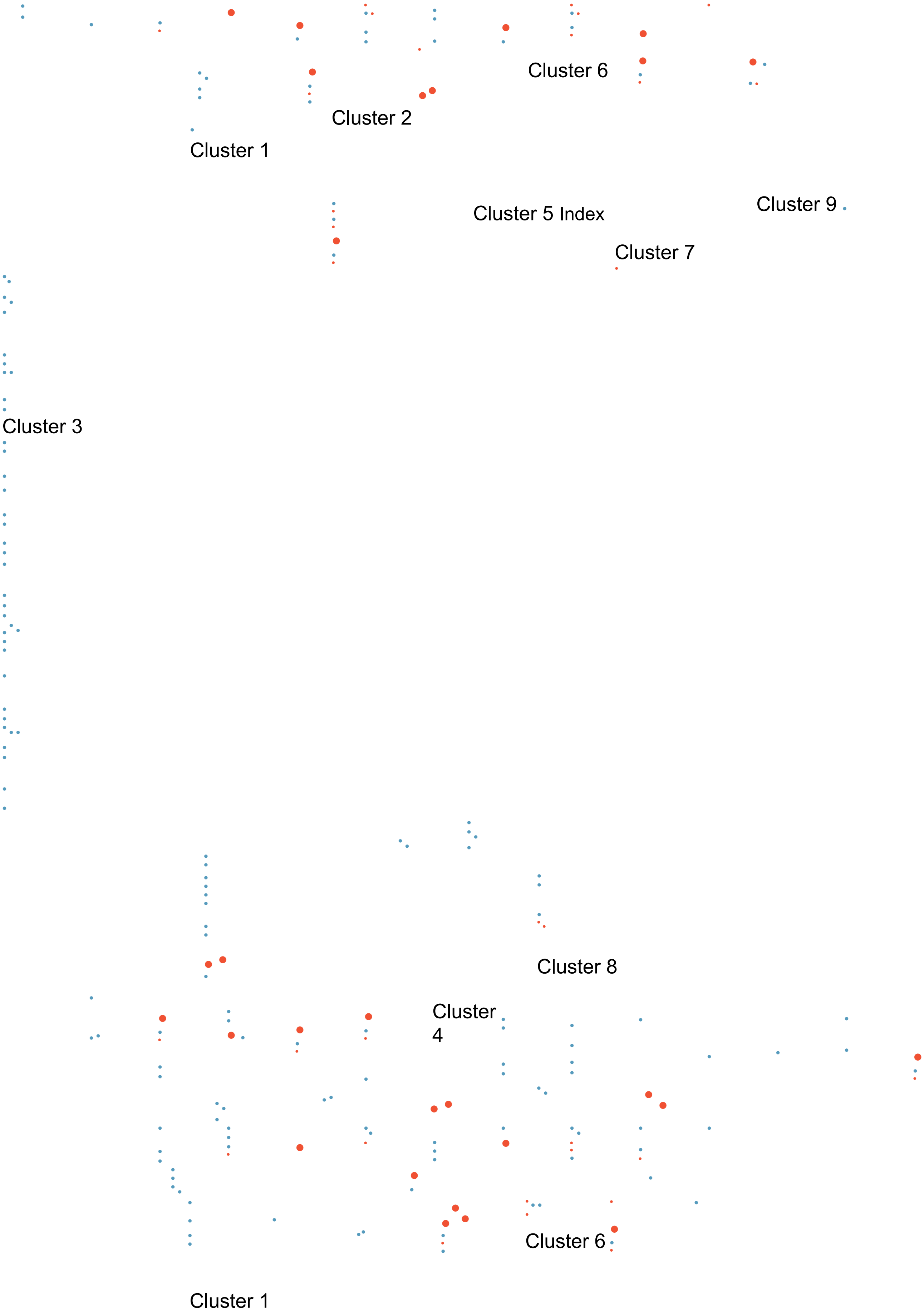


Figure 1.15: Examples of cluster and multistage sampling. In the top panel,

cluster sampling was used: data were binned into nine clusters, three of these clusters were sampled, and all observations within these three cluster were included in the sample. In the bottom panel, multistage sampling was used, which differs from cluster sampling only in that we randomly select a subset of each cluster to be included in the sample rather than measuring every case in each sampled cluster.

Exercises

1.13 Air pollution and birth outcomes, scope of inference. Exercise 1.3 introduces a study where researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth.

(a) Identify the population of interest and the sample in this study.

(b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

1.14 Cheaters, scope of inference. Exercise 1.5 introduces a study where researchers studying the relationship between honesty, age, and self-control conducted an experiment on 160 children between the ages of 5 and 15. The researchers asked each child to toss a fair coin in private and to record the outcome (white or black) on a paper sheet, and said they would only reward children who report white. Half the students were explicitly told not to cheat and the others were not given any explicit instructions. Differences were observed in the cheating rates in the instruction and no instruction groups, as well as some differences across children's characteristics within each group.

(a) Identify the population of interest and the sample in this study.

(b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

1.15 Buteyko method, scope of inference. Exercise 1.4 introduces a study on using the Buteyko shallow breathing technique to reduce asthma symptoms and improve quality of life. As part of this study 600 asthma patients aged 18-69 who relied on medication for asthma treatment were recruited and randomly assigned to two groups: one practiced the Buteyko method and the other did not. Those in the Buteyko group experienced, on average, a significant reduction in asthma symptoms and an improvement in quality of life.

(a) Identify the population of interest and the sample in this study.

(b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

1.16 Stealers, scope of inference. Exercise 1.6 introduces a study on the relationship between socioeconomic class and unethical behavior. As part of this study 129 University of California Berkeley undergraduates were asked to identify themselves as having low or high social-class by comparing themselves to others with the most (least) money, most (least) education, and most (least) respected jobs. They were also presented with a jar of individually wrapped candies and informed that the candies were for children in a nearby laboratory, but that they could take some if they wanted. After completing some unrelated tasks, participants reported the number of candies they had taken. It was found that those who were identified as upper-class took more candy than others.

(a) Identify the population of interest and the sample in this study.

(b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

1.17 Relaxing after work. The General Social Survey asked the question, "After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?" to a random sample of 1,155 Americans. The average relaxing time was found to be 1.65 hours. Determine which of the following is an observation, a variable, a sample statistic (value calculated based on the observed sample), or a population parameter.

(a) An American in the sample.

(b) Number of hours spent relaxing after an average work day.

(c) 1.65.

(d) Average number of hours all Americans spend relaxing after an average work day.

1.18 Cats on YouTube. Suppose you want to estimate the percentage of videos on YouTube that are cat videos. It is impossible for you to watch all videos on YouTube so you use a random video picker to select 1000 videos for you. You find that 2% of these videos are cat videos. Determine which of the following is an observation, a variable, a sample statistic (value calculated based on the observed sample), or a population parameter.

(a) Percentage of all videos on YouTube that are cat videos.

(b) 2%.

(c) A video in your sample.

(d) Whether or not a video is a cat video.

1.19 Course satisfaction across sections. A large college class has 160 students. All 160 students attend the lectures together, but the students are divided into 4 groups, each of 40 students, for lab sections administered by different teaching assistants. The professor wants to conduct a survey about how satisfied the students are with the course, and he believes that the lab section a student is in might affect the student's overall satisfaction with the course.

(a) What type of study is this?

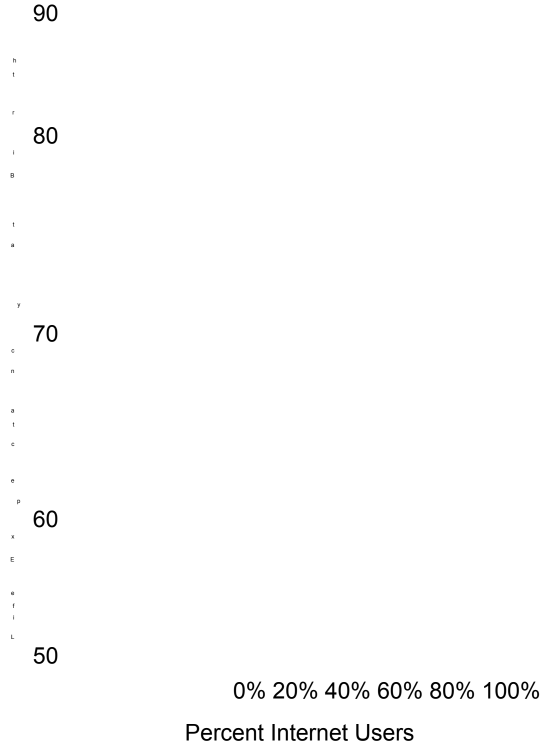
(b) Suggest a sampling strategy for carrying out this study.

1.20 Housing proposal across dorms. On a large college campus first-year students and sophomores live in dorms located on the eastern part of the campus and juniors and seniors live in dorms located on the western part of the campus. Suppose you want to collect student opinions on a new housing structure the college administration is proposing and you want to make sure your survey equally represents opinions from students from all years.

(a) What type of study is this?

(b) Suggest a sampling strategy for carrying out this study.

1.21 Internet use and life expectancy. The following scatterplot was created as part of a study evaluating the relationship between estimated life expectancy at birth (as of 2014) and percentage of internet users (as of 2009) in 208 countries for which such data were available.[23]



(a) Describe the relationship between life expectancy and percentage of internet users.

(b) What type of study is this?

(c) State a possible confounding variable that might explain this relationship and describe its potential effect.

1.22 Stressed out, Part I. A study that surveyed a random sample of otherwise healthy high school students found that they are more likely to get muscle cramps when they are stressed. The study also noted that students drink more coffee and sleep less when they are stressed.

(a) What type of study is this?

(b) Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

(c) State possible confounding variables that might explain the observed relationship between increased stress and muscle cramps.

1.23 Evaluate sampling methods. A university wants to determine what fraction of its undergraduate student body support a new $25 annual fee to improve the student union. For each proposed method below, indicate whether the method is reasonable or not.

(a) Survey a simple random sample of 500 students.

(b) Stratify students by their field of study, then sample 10% of students from each stratum. (c) Cluster students by their ages (e.g. 18 years old in one cluster, 19 years old in one cluster, etc.), then randomly sample three clusters and survey all students in those clusters.

[23] CIA Factbook, Country Comparisons, 2014.

1.24 Random digit dialing. The Gallup Poll uses a procedure called random digit dialing, which creates phone numbers based on a list of all area codes in America in conjunction with the associated number of residential households in each area code. Give a possible reason the Gallup Poll chooses to use random digit dialing instead of picking phone numbers from the phone book.

1.25 Haters are gonna hate, study confirms. A study published in the *Journal of Personality and Social Psychology* asked a group of 200 randomly sampled men and women to evaluate how they felt about various subjects, such as camping, health care, architecture, taxidermy, crossword puzzles, and Japan in order to measure their attitude towards mostly independent stimuli. Then, they presented the participants with information about a new product: a microwave oven. This microwave oven does not exist, but the participants didn't know this, and were given three positive and three negative fake reviews. People who reacted positively to the subjects on the dispositional attitude measurement also tended to react positively to the microwave oven, and those who reacted negatively tended to react negatively to it. Researchers concluded that "some people tend to like things, whereas others tend to dislike things, and a more thorough understanding of this tendency will lead to a more thorough understanding of the psychology of attitudes."[24]

(a) What are the cases?

(b) What is (are) the response variable(s) in this study?

(c) What is (are) the explanatory variable(s) in this study?

(d) Does the study employ random sampling?

(e) Is this an observational study or an experiment? Explain your reasoning.

(f) Can we establish a causal link between the explanatory and response variables?

(g) Can the results of the study be generalized to the population at large?

1.26 Family size. Suppose we want to estimate household size, where a "household" is defined as people living together in the same dwelling, and sharing living accommodations. If we select students at random at an elementary school and ask them what their family size is, will this be a good measure of household size? Or will our average be biased? If so, will it overestimate or underestimate the true value?

1.27 Sampling strategies. A statistics student who is curious about the relationship between the amount of time students spend on social networking sites and their performance at school decides to conduct a survey. Various research strategies for collecting data are described below. In each, name the sampling method proposed and any bias you might expect.

(a) He randomly samples 40 students from the study's population, gives them the survey, asks them to fill it out and bring it back the next day.

(b) He gives out the survey only to his friends, making sure each one of them fills out the survey. (c) He posts a link to an online survey on Facebook and asks his friends to fill out the survey. (d) He randomly samples 5 classes and asks a random sample of students from those classes to fill out the survey.

1.28 Reading the paper. Below are excerpts from two articles published in the *NY Times*: (a)

An article titled *Risks: Smokers Found More Prone to Dementia* states the following:[25]

> "Researchers analyzed data from 23,123 health plan members who participated in a voluntary exam and health behavior survey from 1978 to 1985, when they were 50-60 years old. 23 years later, about 25% of the group had dementia, including 1,136 with Alzheimer's disease and 416 with vascular dementia. After adjusting for other factors, the researchers concluded that pack-a-day smokers were 37% more likely than nonsmokers to develop

> dementia, and the risks went up with increased smoking; 44% for one to two packs a day; and twice the risk for more than two packs."

Based on this study, can we conclude that smoking causes dementia later in life? Explain your reasoning. (b) Another article titled *The School Bully Is Sleepy* states the following:[26]

> "The University of Michigan study, collected survey data from parents on each child's sleep habits and asked both parents and teachers to assess behavioral concerns. About a third of the students studied were identified by parents or teachers as having problems with disruptive behavior or bullying. The researchers found that children who had behavioral issues and those who were identified as bullies were twice as likely to have shown symptoms of sleep disorders."

A friend of yours who read the article says, "The study shows that sleep disorders lead to bullying in school children." Is this statement justified? If not, how best can you describe the conclusion that can be drawn from this study?

[24] Justin Hepler and Dolores Albarrac´ın. "Attitudes without objects - Evidence for a dispositional attitude, its measurement, and its consequences". In: *Journal of personality and social psychology* 104.6 (2013), p. 1060. [25] R.C. Rabin. "Risks: Smokers Found More Prone to Dementia". In: *New York Times* (2010).

[26] T. Parker-Pope. "The School Bully Is Sleepy". In: *New York Times* (2011).

1.4 Experiments

Studies where the researchers assign treatments to cases are called experiments. When this assignment includes randomization, e.g. using a coin flip to decide which treatment a patient receives, it is called a randomized experiment. Randomized experiments are fundamentally important when trying to show a causal connection between two variables.

1.4.1 Principles of experimental design

Randomized experiments are generally built on four principles.

Controlling. Researchers assign treatments to cases, and they do their best to control any other differences in the groups.[27] For example, when patients take a drug in pill form, some patients take the pill with only a sip of water while others may have it with an entire glass of water. To control for the effect of water consumption, a doctor may ask all patients to drink a 12 ounce glass of water with the pill.

Randomization. Researchers randomize patients into treatment groups to account for variables that cannot be controlled. For example, some patients may be more susceptible to a disease than others due to their dietary habits. Randomizing patients into the treatment or control group helps even out such differences, and it also prevents accidental bias from entering the study.

Replication. The more cases researchers observe, the more accurately they can estimate the effect of the explanatory variable on the response. In a single study, we replicate by collecting a sufficiently large sample. Additionally, a group of scientists may replicate an entire study to verify an earlier finding.

Blocking. Researchers sometimes know or suspect that variables, other than the treatment, influ ence the response. Under these circumstances, they may first group individuals based on this variable into blocks and then randomize cases within each block to the treatment groups. This strategy is often referred to as blocking. For instance, if we are looking at the effect of a drug on heart attacks, we might first split patients in the study into low-risk and high-risk blocks, then randomly assign half the patients from each block to the control group and the other half to the treatment group, as shown in Figure 1.16. This strategy ensures each treatment group has an equal number of low-risk and high-risk patients.

It is important to incorporate the first three experimental design principles into any study, and this book describes applicable methods for analyzing data from such experiments. Blocking is a slightly more advanced technique, and statistical methods in this book may be extended to analyze data collected using blocking.

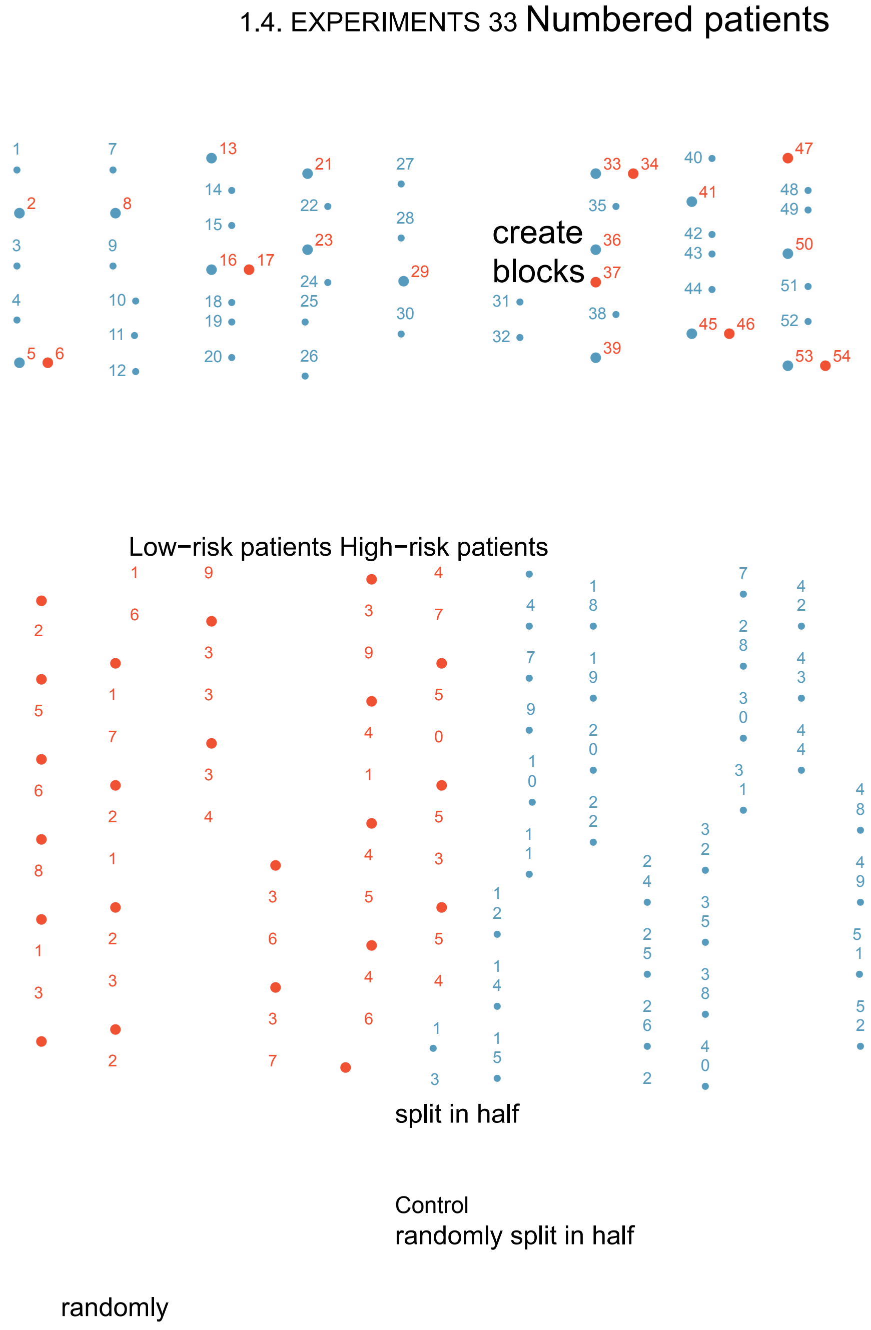
1.4.2 Reducing bias in human experiments

Randomized experiments are the gold standard for data collection, but they do not ensure an unbiased perspective into the cause and effect relationship in all cases. Human studies are perfect examples where bias can unintentionally arise. Here we reconsider a study where a new drug was used to treat heart attack patients. In particular, researchers wanted to know if the drug reduced deaths in patients.

These researchers designed a randomized experiment because they wanted to draw causal con clusions about the drug's effect. Study volunteers[28] were randomly placed into two study groups. One group, the treatment group, received the drug. The other group, called the control group, did not receive any drug treatment.

[27] This is a different concept than a *control group*, which we discuss in the second principle and in Section 1.4.2.

[28] Human subjects are often called patients, volunteers, or study participants.



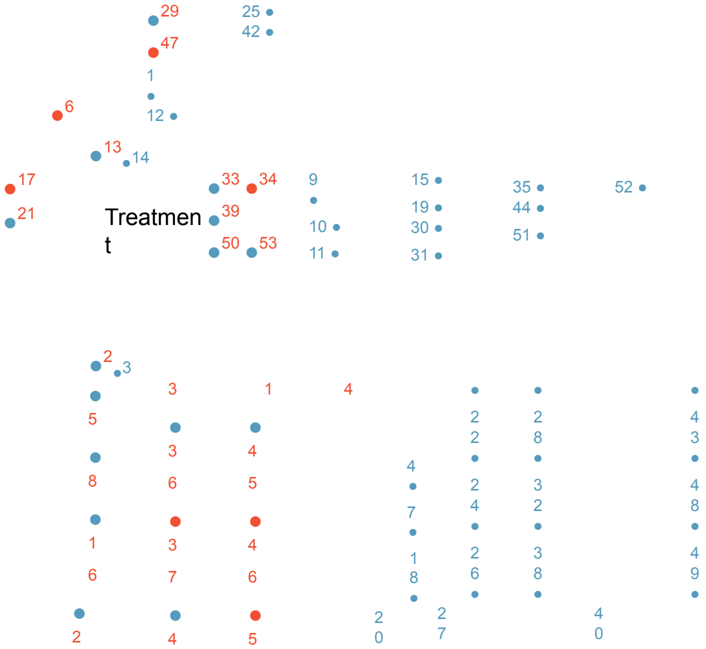


Figure 1.16: Blocking using a variable depicting patient risk. Patients are first divided into low-risk and high-risk blocks, then each block is evenly separated into the treatment groups using randomization. This strategy ensures an equal representation of patients in each treatment group from both the low-risk and high-risk categories.

Put yourself in the place of a person in the study. If you are in the treatment group, you are given a fancy new drug that you anticipate will help you. On the other hand, a person in the other group doesn't receive the drug and sits idly, hoping her participation doesn't increase her risk of death. These perspectives suggest there are actually two effects: the one of interest is the effectiveness of the drug, and the second is an emotional effect that is difficult to quantify.

Researchers aren't usually interested in the emotional effect, which might bias the study. To circumvent this problem, researchers do not want patients to know which group they are in. When researchers keep the patients uninformed about their treatment, the study is said to be blind. But there is one problem: if a patient doesn't receive a treatment, she will know she is in the control group. The solution to this problem is to give fake treatments to patients in the control group. A fake treatment is called a placebo, and an effective placebo is the key to making a study truly blind. A classic example of a placebo is a sugar pill that is made to look like the actual treatment pill. Often times, a placebo results in a slight but real improvement in patients. This effect has been dubbed the placebo effect.

The patients are not the only ones who should be blinded: doctors and researchers can accidentally bias a study. When a doctor knows a patient has been given the real treatment, she might inadvertently give that patient more attention or care than a patient that she knows is on the placebo. To guard against this bias, which again has been found to have a measurable effect in some instances, most modern studies employ a double-blind setup where doctors or researchers who interact with patients are, just like the patients, unaware of who is or is not receiving the treatment.[29]

GUIDED PRACTICE 1.16

Look back to the study in Section 1.1 where researchers were testing whether stents were effective at reducing strokes in at-risk patients. Is this an experiment? Was the study blinded? Was it double-blinded?[30]

GUIDED PRACTICE 1.17

For the study in Section 1.1, could the researchers have employed a placebo? If so, what would that placebo have looked like?[31]

You may have many questions about the ethics of sham surgeries to create a placebo after reading Guided Practice 1.17. These questions may have even arisen in your mind when in the general experiment context, where a possibly helpful treatment was withheld from individuals in the control group; the main difference is that a sham surgery tends to create additional risk, while withholding a treatment only maintains a person's risk.

There are always multiple viewpoints of experiments and placebos, and rarely is it obvious which is ethically "correct". For instance, is it ethical to use a sham surgery when it creates a risk to the patient? However, if we don't use sham surgeries, we may promote the use of a costly treatment that has no real effect; if this happens, money and other resources will be diverted away from other treatments that are known to be helpful. Ultimately, this is a difficult situation where we cannot perfectly protect both the patients who have volunteered for the study and the patients who may benefit (or not) from the treatment in the future.

[29] There are always some researchers involved in the study who do know which patients are receiving which treat ment. However, they do not interact with the study's patients and do not tell the blinded health care professionals who is receiving which treatment.

[30] The researchers assigned the patients into their treatment groups, so this study was an experiment. However, the patients could distinguish what treatment they received, so this study was not blind. The study could not be double-blind since it was not blind.

[31] Ultimately, can we make patients think they got treated from a surgery? In fact, we can, and some experiments use what's called a sham surgery. In a sham surgery, the patient does undergo surgery, but the patient does not receive the full treatment, though they will still get a placebo effect.

Exercises

1.29 Light and exam performance. A study is designed to test the effect of light level on exam performance of students. The researcher believes that light levels might have different effects on males and females, so wants to make sure both are equally represented in each treatment. The treatments are fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps).

(a) What is the response variable?

(b) What is the explanatory variable? What are its levels?

(c) What is the blocking variable? What are its levels?

1.30 Vitamin supplements. To assess the effectiveness of taking large doses of vitamin C in reducing the duration of the common cold, researchers recruited 400 healthy volunteers from staff and students at a university. A quarter of the patients were assigned a placebo, and the rest were evenly divided between 1g Vitamin C, 3g Vitamin C, or 3g Vitamin C plus additives to be taken at onset of a cold for the following two days. All tablets had identical appearance and packaging. The nurses who handed the prescribed pills to the patients knew which patient received which treatment, but the researchers assessing the patients when they were sick did not. No significant differences were observed in any measure of cold duration or severity between the four groups, and the placebo group had the shortest duration of symptoms.[32]

(a) Was this an experiment or an observational study? Why?

(b) What are the explanatory and response variables in this study?

(c) Were the patients blinded to their treatment?

(d) Was this study double-blind?

(e) Participants are ultimately able to choose whether or not to use the pills prescribed to them. We might expect that not all of them will adhere and take their pills. Does this introduce a confounding variable to the study? Explain your reasoning.

1.31 Light, noise, and exam performance. A study is designed to test the effect of light level and noise level on exam performance of students. The researcher believes that light and noise levels might have different effects on males and females, so wants to make sure both are equally represented in each treatment. The light treatments considered are fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps). The noise treatments considered are no noise, construction noise, and human chatter noise.

(a) What type of study is this?

(b) How many factors are considered in this study? Identify them, and describe their levels.

(c) What is the role of the sex variable in this study?

1.32 Music and learning. You would like to conduct an experiment in class to see if students learn better if they study without any music, with music that has no lyrics (instrumental), or with music that has lyrics. Briefly outline a design for this study.

1.33 Soda preference. You would like to conduct an experiment in class to see if your classmates prefer the taste of regular Coke or Diet Coke. Briefly outline a design for this study.

1.34 Exercise and mental health. A researcher is interested in the effects of exercise on mental health and he proposes the following study: Use stratified random sampling to ensure representative proportions of 18-30, 31-40 and 41- 55 year olds from the population. Next, randomly assign half the subjects from each age group to exercise twice a week, and instruct the rest not to exercise. Conduct a mental health exam at the beginning and at the end of the study, and compare the results.

(a) What type of study is this?

(b) What are the treatment and control groups in this study?

(c) Does this study make use of blocking? If so, what is the blocking variable?

(d) Does this study make use of blinding?

(e) Comment on whether or not the results of the study can be used to establish a causal relationship between exercise and mental health, and indicate whether or not the conclusions can be generalized to the population at large.

(f) Suppose you are given the task of determining if this proposed study should get funding. Would you have any reservations about the study proposal?

[32] C. Audera et al. "Mega-dose vitamin C in treatment of the common cold: a randomised controlled trial". In: *Medical Journal of Australia* 175.7 (2001), pp. 359–362.

Chapter exercises

1.35 Pet names. The city of Seattle, WA has an open data portal that includes pets registered in the city. For each registered pet, we have information on the pet's name and species. The following visualization plots the proportion of dogs with a given name versus the proportion of cats with the same name. The 20 most common cat and dog names are displayed. The diagonal line on the plot is the *x* = *y* line; if a name appeared on this line, the name's popularity would be exactly the same for dogs and cats.

(a) Are these data collected as part of an experiment or an observational study?

(b) What is the most common dog name? What is the most common cat name?

(c) What names are more common for cats than dogs?

(d) Is the relationship between the two variables positive or negative? What does this mean in context of the data?

1.36 Stressed out, Part II. In a study evaluating the relationship between stress and muscle cramps, half the subjects are randomly assigned to be exposed to increased stress by being placed into an elevator that falls rapidly and stops abruptly and the other half are left at no or baseline stress.

(a) What type of study is this?

(b) Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

1.37 Chia seeds and weight loss. Chia Pets – those terra-cotta figurines that sprout fuzzy green hair – made the chia plant a household name. But chia has gained an entirely new reputation as a diet supplement. In one 2009 study, a team of researchers recruited 38 men and divided them randomly into two groups: treatment or control. They also recruited 38 women, and they randomly placed half of these participants into the treatment group and the other half into the control group. One group was given 25 grams of chia seeds twice a day, and the other was given a placebo. The subjects volunteered to be a part of the study. After 12 weeks, the scientists found no significant difference between the groups in appetite or weight loss.[33]

(a) What type of study is this?

(b) What are the experimental and control treatments in this study?

(c) Has blocking been used in this study? If so, what is the blocking variable?

(d) Has blinding been used in this study?

(e) Comment on whether or not we can make a causal statement, and indicate whether or not we can generalize the conclusion to the population at large.

1.38 City council survey. A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures. For each part below, identify the sampling methods described, and describe the statistical pros and cons of the method in the city's context.

(a) Randomly sample 200 households from the city.

(b) Divide the city into 20 neighborhoods, and sample 10 households from each neighborhood. (c) Divide the city into 20 neighborhoods, randomly sample 3 neighborhoods, and then sample all households from those 3 neighborhoods.

(d) Divide the city into 20 neighborhoods, randomly sample 8 neighborhoods, and then randomly sample 50 households from those neighborhoods.

(e) Sample the 200 households closest to the city council offices.

[33] D.C. Nieman et al. "Chia seed does not promote weight loss or alter disease risk factors in overweight adults". In: *Nutrition Research* 29.6 (2009), pp. 414–418.

1.39 Flawed reasoning. Identify the flaw(s) in reasoning in the following scenarios. Explain what the individuals in the study should have done differently if they wanted to make such strong conclusions.

(a) Students at an elementary school are given a questionnaire that they are asked to return after their parents have completed it. One of the questions asked is, "Do you find that your work schedule makes it difficult for you to spend time with your kids after school?" Of the parents who replied, 85% said "no". Based on these results, the school officials conclude that a great majority of the parents have no difficulty spending time with their kids after school.

(b) A survey is conducted on a simple random sample of 1,000 women who recently gave birth, asking them about whether or not they smoked during pregnancy. A follow-up survey asking if the children have respiratory problems is conducted 3 years later. However, only 567 of these women are reached at the same address. The researcher reports that these 567 women are representative of all mothers.

(c) An orthopedist administers a questionnaire to 30 of his patients who do not have any joint problems and finds that 20 of them regularly go running. He concludes that running decreases the risk of joint problems.

1.40 Income and education in US counties. The scatterplot below shows the relationship between per capita income (in thousands of dollars) and percent of population with a bachelor's degree in 3,143 counties in the US in 2010.

(a) What are the explanatory and response variables?

(b) Describe the relationship between the two variables. Make sure to discuss unusual observations, if any.

(c) Can we conclude that having a bachelor's degree increases one's income?

$60k

$40k

$20k

$0

0% 20% 40% 60% 80% Percent with Bachelor's Degree

1.41 Eat better, feel better? In a public health study on the effects of consumption of fruits and vegetables on psychological well-being in young adults, participants were randomly assigned to three groups: (1) diet as-usual, (2) an ecological momentary intervention involving text message reminders to increase their fruits and vegetable consumption plus a voucher to purchase them, or (3) a fruit and vegetable intervention in which participants were given two additional daily servings of fresh fruits and vegetables to consume on top of their normal diet. Participants were asked to take a nightly survey on their smartphones. Participants were student volunteers at the University of Otago, New Zealand. At the end of the 14-day study, only participants in the third group showed improvements to their psychological well-being across the 14-days relative to the other groups.[34]

(a) What type of study is this?

(b) Identify the explanatory and response variables.

(c) Comment on whether the results of the study can be generalized to the population. (d) Comment on whether the results of the study can be used to establish causal relationships. (e) A newspaper article reporting on the study states, "The results of this study provide proof that giving young adults fresh fruits and vegetables to eat can have psychological benefits, even over a brief period of time." How would you suggest revising this statement so that it can be supported by the study?

[34] Tamlin S Conner et al. "Let them eat fruit! The effect of fruit and vegetable consumption on psychological well-being in young adults: A randomized controlled trial". In: *PloS one* 12.2 (2017), e0171206.

1.42 Screens, teens, and psychological well-being. In a study of three nationally representative large scale data sets from Ireland, the United States, and the United Kingdom (n = 17,247), teenagers between the ages of 12 to 15 were asked to keep a diary of their screen time and answer questions about how they felt or acted. The answers to these questions were then used to compute a psychological well-being score. Additional data were collected and included in the analysis, such as each child's sex and age, and on the mother's education, ethnicity, psychological distress, and employment. The study concluded that there is little clear-cut evidence that screen time decreases adolescent well-being.[35]

(a) What type of study is this?

(b) Identify the explanatory variables.

(c) Identify the response variable.

(d) Comment on whether the results of the study can be generalized to the population, and why. (e) Comment on whether the results of the study can be used to establish causal relationships.

1.43 Stanford Open Policing. The Stanford Open Policing project gathers, analyzes, and releases records from traffic stops by law enforcement agencies across the United States. Their goal is to help researchers, journalists, and policymakers investigate and improve interactions between police and the public.[36] The following is an excerpt from a summary table created based off of the data collected as part of this project.

		Driver's	No. of stops	% of	stopped
County	State	race	per year	cars searched	drivers arrested
Apaice County	Arizona	Black	266	0.08	0.02
Apaice County	Arizona	Hispanic	1008	0.05	0.02
Apaice County	Arizona	White	6322	0.02	0.01
Cochise County	Arizona	Black	1169	0.05	0.01
Cochise County	Arizona	Hispanic	9453	0.04	0.01
Cochise County	Arizona	White	10826	0.02	0.01
. . .	. . .	. . .	. . .	. . .	. . .
Wood County	Wisconsin	Black	16	0.24	0.10
Wood County	Wisconsin	Hispanic	27	0.04	0.03
Wood County	Wisconsin	White	1157	0.03	0.03

(a) What variables were collected on each individual traffic stop in order to create to the summary table above?

(b) State whether each variable is numerical or categorical. If numerical, state whether it is continuous or discrete. If categorical, state whether it is ordinal or not.

(c) Suppose we wanted to evaluate whether vehicle search rates are different for drivers of different races. In this analysis, which variable would be the response variable and which variable would be the explanatory variable?

1.44 Space launches. The following summary table shows the number of space launches in the US by the type of launching agency and the outcome of the launch (success or failure).[37]

	1957 - 1999		2000 - 2018	
	Failure	Success	Failure	Success
Private	13	295	10	562
State	281	3751	33	711
Startup	-	-	5	65

(a) What variables were collected on each launch in order to create to the summary table above? (b) State whether each variable is numerical or categorical. If numerical, state whether it is continuous or discrete. If categorical, state whether it is ordinal or not.

(c) Suppose we wanted to study how the success rate of launches vary between launching agencies and over time. In this analysis, which variable would be the response variable and which variable would be the explanatory variable?

[35] Amy Orben and AK Baukney-Przybylski. “Screens, Teens and Psychological Well-Being: Evidence from three time-use diary studies”. In: *Psychological Science* (2018).

[36] Emma Pierson et al. “A large-scale analysis of racial disparities in police stops across the United States”. In: *arXiv preprint arXiv:1706.05678* (2017).

[37] JSR Launch Vehicle Database, A comprehensive list of suborbital space launches, 2019 Feb 10 Edition.

Chapter 2 Summarizing data

This chapter focuses on the mechanics and construction of summary statistics and graphs. We use statistical software for generating the summaries and graphs presented in this chapter and book. However, since this might be your first exposure to these concepts, we take our time in this chapter to detail how to create them. Mastery of the content presented in this chapter will be crucial for understanding the methods and techniques introduced in rest of the book.

For videos, slides, and other resources, please visit
www.openintro.org/os

2.1 Examining numerical data

In this section we will explore techniques for summarizing numerical variables. For example, consider the loan amount variable from the loan50 data set, which represents the loan size for all 50 loans in the data set. This variable is numerical since we can sensibly discuss the numerical difference of the size of two loans. On the other hand, area codes and zip codes are not numerical, but rather they are categorical variables.

Throughout this section and the next, we will apply these methods using the loan50 and county data sets, which were introduced in Section 1.2. If you'd like to review the variables from either data set, see Figures 1.3 and 1.5.

2.1.1 Scatterplots for paired data

A scatterplot provides a case-by-case view of data for two numerical variables. In Figure 1.8 on page 16, a scatterplot was used to examine the homeownership rate against the fraction of housing units that were part of multi-unit properties (e.g. apartments) in the county data set. Another scatterplot is shown in Figure 2.1, comparing the total income of a borrower (total income) and the amount they borrowed (loan amount) for the loan50 data set. In any scatterplot, each point represents a single case. Since there are 50 cases in loan50, there are 50 points in Figure 2.1.

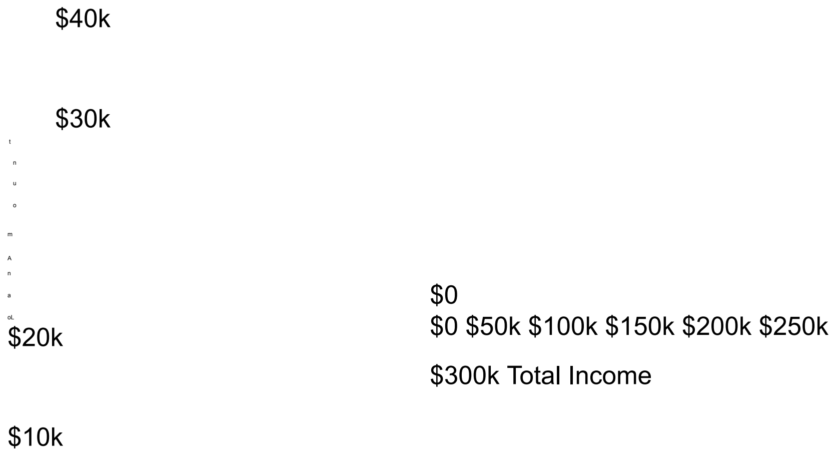


Figure 2.1: A scatterplot of total income versus loan amount for the loan50 data set.

Looking at Figure 2.1, we see that there are many borrowers with an income below $100,000 on the left side of the graph, while there are a handful of borrowers with income above $250,000.

EXAMPLE 2.1

Figure 2.2 shows a plot of median household income against the poverty rate for 3,142 counties. What can be said about the relationship between these variables?

The relationship is evidently nonlinear, as highlighted by the dashed line. This is different from previous scatterplots we've seen, which show relationships that do not show much, if any, curvature in the trend.

$120k

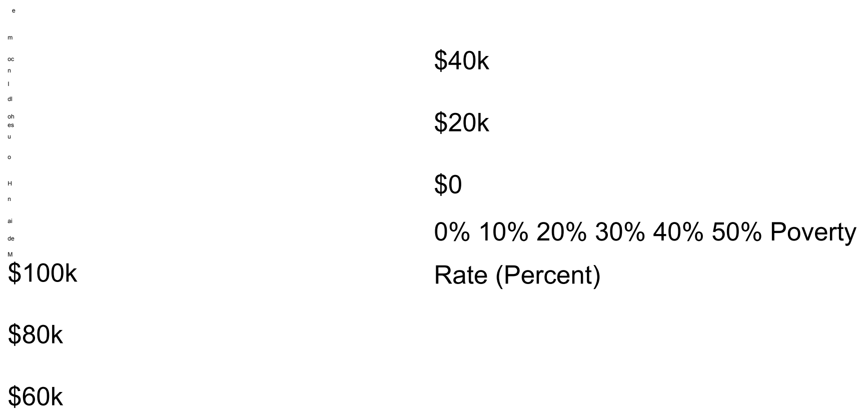


Figure 2.2: A scatterplot of the median household income against the poverty rate for the county data set. A statistical model has also been fit to the data and is shown as a dashed line.

GUIDED PRACTICE 2.2

What do scatterplots reveal about the data, and how are they useful?[1]

GUIDED PRACTICE 2.3

Describe two variables that would have a horseshoe-shaped association in a scatterplot (∩ or _).[2]

2.1.2 Dot plots and the mean

Sometimes two variables are one too many: only one variable may be of interest. In these cases, a dot plot provides the most basic of displays. A dot plot is a one-variable scatterplot; an example using the interest rate of 50 loans is shown in Figure 2.3. A stacked version of this dot plot is shown in Figure 2.4.

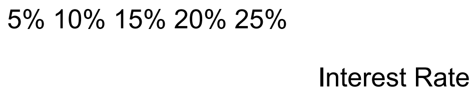


Figure 2.3: A dot plot of interest rate for the loan50 data set. The distribu tion's mean is shown as a red triangle.

[1] Answers may vary. Scatterplots are helpful in quickly spotting associations relating variables, whether those associations come in the form of simple trends or whether those relationships are more complex.

[2] Consider the case where your vertical axis represents something "good" and your horizontal axis represents something that is only good in moderation. Health and water consumption fit this description: we require some water to survive, but consume too much and it becomes toxic and can kill a person.

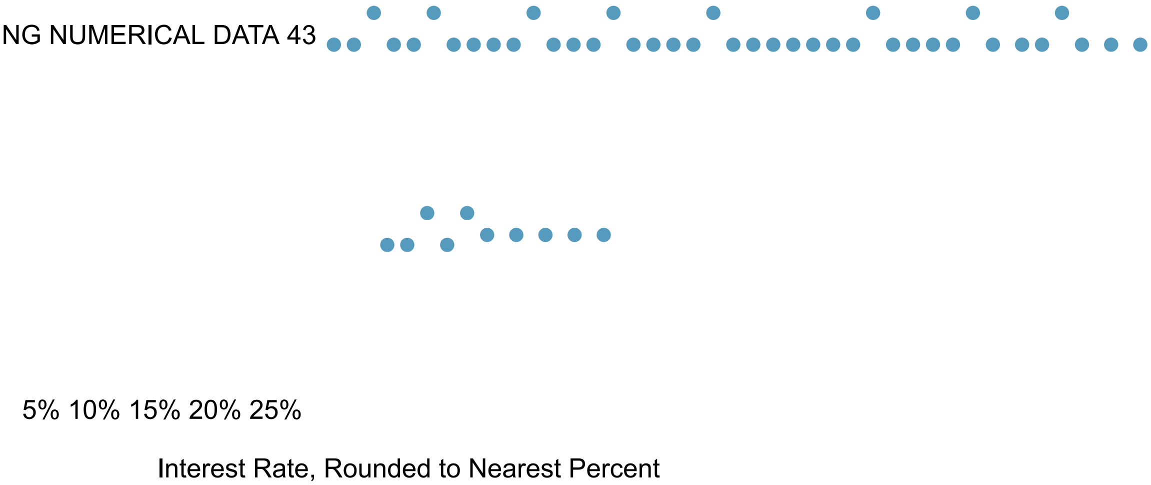


Figure 2.4: A stacked dot plot of interest rate for the loan50 data set. The rates have been rounded to the nearest percent in this plot, and the distribution's mean is shown as a red triangle.

The mean, often called the average, is a common way to measure the center of a distribution of data. To compute the mean interest rate, we add up all the interest rates and divide by the number of observations:

$$\bar{x} = \frac{10.90\% + 9.92\% + 26.30\% + \cdots + 6.08\%}{50} = 11.57\%$$

The sample mean is often labeled $\bar{x}$. The letter x is being used as a generic placeholder for the variable of interest, interest rate, and the bar over the x communicates we're looking at the average interest rate, which for these 50 loans was 11.57%. It is useful to think of the mean as the balancing point of the distribution, and it's shown as a triangle in Figures 2.3 and 2.4.

MEAN

The sample mean can be computed as the sum of the observed values divided by the number of observations:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

where $x_1, x_2, \ldots, x_n$ represent the n observed values.

GUIDED PRACTICE 2.4

Examine the equation for the mean. What does x_1 correspond to? And x_2? Can you infer a general meaning to what x_i might represent?[3]

GUIDED PRACTICE 2.5

What was n in this sample of loans?[4]

The loan50 data set represents a sample from a larger population of loans made through Lending Club. We could compute a mean for this population in the same way as the sample mean. However, the population mean has a special label: μ. The symbol μ is the Greek letter *mu* and represents the average of all observations in the population. Sometimes a subscript, such as $_x$, is used to represent which variable the population mean refers to, e.g. μ_x. Often times it is too expensive to measure the population mean precisely, so we often estimate μ using the sample mean, $\bar{x}$.

[3] x_1 corresponds to the interest rate for the first loan in the sample (10.90%), x_2 to the second loan's interest rate (9.92%), and x_i corresponds to the interest rate for the i^{th} loan in the data set. For example, if $i = 4$, then we're examining x_4, which refers to the fourth observation in the data set.

[4] The sample size was $n = 50$.

EXAMPLE 2.6

The average interest rate across all loans in the population can be estimated using the sample data. Based on the sample of 50 loans, what would be a reasonable estimate of μ_x, the mean interest rate for all loans in the full data set?

The sample mean, 11.57%, provides a rough estimate of μ_x. While it's not perfect, this is our single best guess of the average interest rate of all the loans in the population under study.

In Chapter 5 and beyond, we will develop tools to characterize the accuracy of *point estimates* like the sample mean. As you might have guessed, point estimates based on larger samples tend to be more accurate than those based on smaller samples.

EXAMPLE 2.7

The mean is useful because it allows us to rescale or standardize a metric into something more easily interpretable and comparable. Provide 2 examples where the mean is useful for making comparisons.

1. We would like to understand if a new drug is more effective at treating asthma attacks than the standard drug. A trial of 1500 adults is set up, where 500 receive the new drug, and 1000 receive a standard drug in the control group:

	New drug	Standard drug
Number of patients	500	1000
Total asthma attacks	200	300

Comparing the raw counts of 200 to 300 asthma attacks would make it appear that the new drug is better, but this is an artifact of the imbalanced group sizes. Instead, we should look at the average number of asthma attacks per patient in each group:

New drug: 200/500 = 0.4 Standard drug: 300/1000 = 0.3

The standard drug has a lower average number of asthma attacks per patient than the average in the treatment group.

2. Emilio opened a food truck last year where he sells burritos, and his business has stabilized over the last 3 months. Over that 3 month period, he has made $11,000 while working 625 hours. Emilio's average hourly earnings provides a useful statistic for evaluating whether his venture is, at least from a financial perspective, worth it:

$$\frac{\$11000}{625 \text{ hours}} = \$17.60 \text{ per hour}$$

By knowing his average hourly wage, Emilio now has put his earnings into a standard unit that is easier to compare with many other jobs that he might consider.

EXAMPLE 2.8

Suppose we want to compute the average income per person in the US. To do so, we might first think to take the mean of the per capita incomes across the 3,142 counties in the county data set. What would be a better approach?

The county data set is special in that each county actually represents many individual people. If we were to simply average across the income variable, we would be treating counties with 5,000 and 5,000,000 residents equally in the calculations. Instead, we should compute the total income for each county, add up all the counties' totals, and then divide by the number of people in all the counties. If we completed these steps with the county data, we would find that the per capita income for the US is $30,861. Had we computed the *simple* mean of per capita income across counties, the result would have been just $26,093!

This example used what is called a weighted mean. For more information on this topic, check out the following online supplement regarding weighted means openintro.org/d?file=stat wtd mean.

2.1.3 Histograms and shape

Dot plots show the exact value for each observation. This is useful for small data sets, but they can become hard to read with larger samples. Rather than showing the value of each observation, we prefer to think of the value as belonging to a *bin*. For example, in the loan50 data set, we created a table of counts for the number of loans with interest rates between 5.0% and 7.5%, then the number of loans with rates between 7.5% and 10.0%, and so on. Observations that fall on the boundary of a bin (e.g. 10.00%) are allocated to the lower bin. This tabulation is shown in Figure 2.5. These binned counts are plotted as bars in Figure 2.6 into what is called a histogram, which resembles a more heavily binned version of the stacked dot plot shown in Figure 2.4.

Interest Rate 5.0% - 7.5% 7.5% - 10.0% 10.0% - 12.5% 12.5% - 15.0% · · · 25.0% - 27.5% Count 11 15 8 4 · · · 1

Figure 2.5: Counts for the binned interest rate data.

15

y c

0

5% 10% 15% 20% 25% Interest

Rate

n e u q e r F

10

5

Figure 2.6: A histogram of interest rate. This distribution is strongly skewed to the right.

Histograms provide a view of the data density. Higher bars represent where the data are relatively more common. For instance, there are many more loans with rates between 5% and 10% than loans with rates between 20% and 25% in the data set. The bars make it easy to see how the density of the data changes relative to the interest rate.

Histograms are especially convenient for understanding the shape of the data distribution. Figure 2.6 suggests that most loans have rates under 15%, while only a handful of loans have rates above 20%. When data trail off to the right in this way and has a longer right tail, the shape is said to be right skewed.[5]

Data sets with the reverse characteristic – a long, thinner tail to the left – are said to be left skewed. We also say that such a distribution has a long left tail. Data sets that show roughly equal trailing off in both directions are called symmetric.

LONG TAILS TO IDENTIFY SKEW

When data trail off in one direction, the distribution has a long tail. If a distribution has a long left tail, it is left skewed. If a distribution has a long right tail, it is right skewed.

[5] Other ways to describe data that are right skewed: skewed to the right, skewed to the high end, or skewed to the positive end.

GUIDED PRACTICE 2.9

Take a look at the dot plots in Figures 2.3 and 2.4. Can you see the skew in the data? Is it easier to see the skew in this histogram or the dot plots?[6]

GUIDED PRACTICE 2.10

Besides the mean (since it was labeled), what can you see in the dot plots that you cannot see in the histogram?[7]

In addition to looking at whether a distribution is skewed or symmetric, histograms can be used to identify modes. A mode is represented by a prominent peak in the distribution. There is only one prominent peak in the histogram of loan amount.

A definition of *mode* sometimes taught in math classes is the value with the most occurrences in the data set. However, for many real-world data sets, it is common to have *no* observations with the same value in a data set, making this definition impractical in data analysis.

Figure 2.7 shows histograms that have one, two, or three prominent peaks. Such distributions are called unimodal, bimodal, and multimodal, respectively. Any distribution with more than 2 prominent peaks is called multimodal. Notice that there was one prominent peak in the unimodal distribution with a second less prominent peak that was not counted since it only differs from its neighboring bins by a few observations.

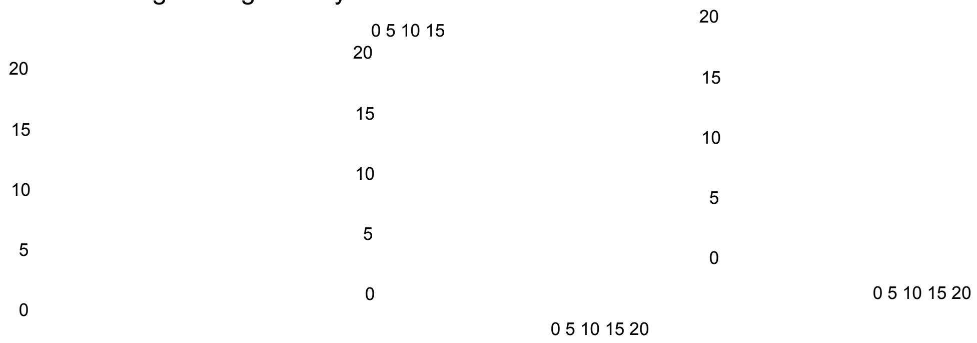


Figure 2.7: Counting only prominent peaks, the distributions are (left to right) unimodal, bimodal, and multimodal. Note that we've said the left plot is unimodal intentionally. This is because we are counting *prominent* peaks, not just any peak.

EXAMPLE 2.11

Figure 2.6 reveals only one prominent mode in the interest rate. Is the distribution unimodal, bimodal, or multimodal?

Unimodal. Remember that *uni* stands for 1 (think *uni*cycles). Similarly, *bi* stands for 2 (think *bi*cycles). We're hoping a *multicycle* will be invented to complete this analogy.

GUIDED PRACTICE 2.12

Height measurements of young students and adult teachers at a K-3 elementary school were taken. How many modes would you expect in this height data set?[8]

Looking for modes isn't about finding a clear and correct answer about the number of modes in a distribution, which is why *prominent* is not rigorously defined in this book. The most important part of this examination is to better understand your data.

[6] The skew is visible in all three plots, though the flat dot plot is the least useful. The stacked dot plot and histogram are helpful visualizations for identifying skew.

[7] The interest rates for individual loans.

[8] There might be two height groups visible in the data set: one of the students and one of the adults. That is, the data are probably bimodal.

2.1.4 Variance and standard deviation

The mean was introduced as a method to describe the center of a data set, and variability in the data is also important. Here, we introduce two measures of variability: the variance and the standard deviation. Both of these are very useful in data analysis, even though their formulas are a bit tedious to calculate by hand. The standard deviation is the easier of the two to comprehend, and it roughly describes how far away the typical observation is from the mean.

We call the distance of an observation from its mean its deviation. Below are the deviations for the 1st, 2nd, 3rd, and 50th observations in the interest rate variable:

$$
\begin{aligned}
x_1 - \bar{x} &= 10.90 - 11.57 = -0.67 \\
x_2 - \bar{x} &= 9.92 - 11.57 = -1.65 \\
x_3 - \bar{x} &= 26.30 - 11.57 = 14.73 \\
&\vdots \\
x_{50} - \bar{x} &= 6.08 - 11.57 = -5.49
\end{aligned}
$$

If we square these deviations and then take an average, the result is equal to the sample variance, denoted by s^2:

$$
\begin{aligned}
s^2 &= \frac{(-0.67)^2 + (-1.65)^2 + (14.73)^2 + \cdots + (-5.49)^2}{50 - 1} \\
&= \frac{0.45 + 2.72 + 216.97 + \cdots + 30.14}{49} \\
&= 25.52
\end{aligned}
$$

We divide by $n - 1$, rather than dividing by n, when computing a sample's variance; there's some mathematical nuance here, but the end result is that doing this makes this statistic slightly more reliable and useful.

Notice that squaring the deviations does two things. First, it makes large values relatively much larger, seen by comparing $(-0.67)^2$, $(-1.65)^2$, $(14.73)^2$, and $(-5.49)^2$. Second, it gets rid of any negative signs.

The standard deviation is defined as the square root of the variance:

$$s = \sqrt{25.52} = 5.05$$

While often omitted, a subscript of $_x$ may be added to the variance and standard deviation, i.e. s^2_x and s_x, if it is useful as a reminder that these are the variance and standard deviation of the observations represented by $x_1, x_2, ..., x_n$.

VARIANCE AND STANDARD DEVIATION

The variance is the average squared distance from the mean. The standard deviation is the square root of the variance. The standard deviation is useful when considering how far the data are distributed from the mean.

The standard deviation represents the typical deviation of observations from the mean. Usually about 70% of the data will be within one standard deviation of the mean and about 95% will be within two standard deviations. However, as seen in Figures 2.8 and 2.9, these percentages are not strict rules.

Like the mean, the population values for variance and standard deviation have special symbols: $\sigma^2$for the variance and σ for the standard deviation. The symbol σ is the Greek letter *sigma*.

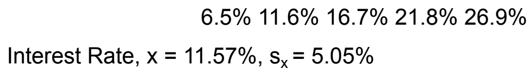


Figure 2.8: For the interest rate variable, 34 of the 50 loans (68%) had interest rates within 1 standard deviation of the mean, and 48 of the 50 loans (96%) had rates within 2 standard deviations. Usually about 70% of the data are within 1 standard deviation of the mean and 95% within 2 standard deviations, though this is far from a hard rule.

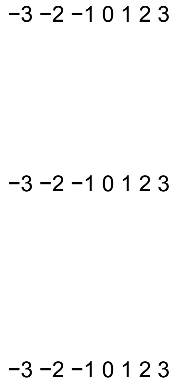


Figure 2.9: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

GUIDED PRACTICE 2.13

On page 45, the concept of shape of a distribution was introduced. A good description of the shape of a distribution should include modality and whether the distribution is symmetric or skewed to one side. Using Figure 2.9 as an example, explain why such a description is important.[9]

EXAMPLE 2.14

Describe the distribution of the interest rate variable using the histogram in Figure 2.6. The description should incorporate the center, variability, and shape of the distribution, and it should also be placed in context. Also note any especially unusual cases.

The distribution of interest rates is unimodal and skewed to the high end. Many of the rates fall near the mean at 11.57%, and most fall within one standard deviation (5.05%) of the mean. There are a few exceptionally large interest rates in the sample that are above 20%.

In practice, the variance and standard deviation are sometimes used as a means to an end, where the "end" is being able to accurately estimate the uncertainty associated with a sample statistic. For example, in Chapter 5 the standard deviation is used in calculations that help us understand how much a sample mean varies from one sample to the next.

[9] Figure 2.9 shows three distributions that look quite different, but all have the same mean, variance, and standard deviation. Using modality, we can distinguish between the first plot (bimodal) and the last two (unimodal). Using skewness, we can distinguish between the last plot (right skewed) and the first two. While a picture, like a histogram, tells a more complete story, we can use modality and shape (symmetry/skew) to characterize basic information about a distribution.

2.1.5 Box plots, quartiles, and the median

A box plot summarizes a data set using five statistics while also plotting unusual observations. Figure 2.10 provides a vertical dot plot alongside a box plot of the interest rate variable from the loan50 data set.

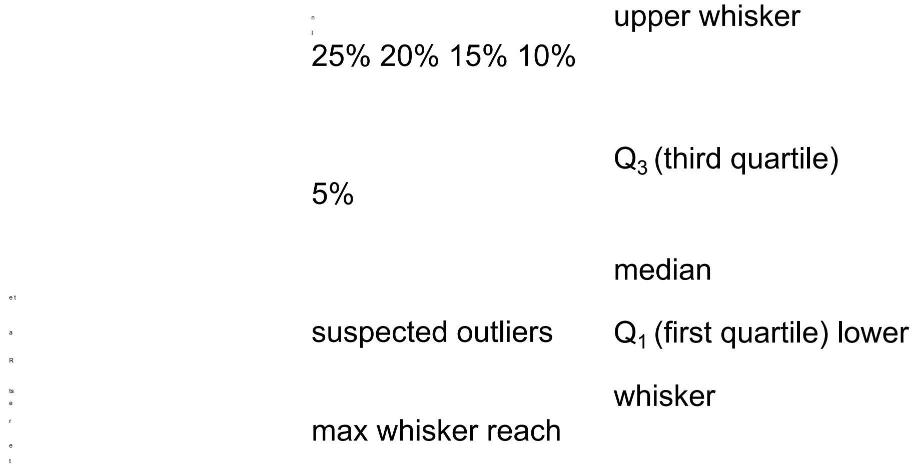


Figure 2.10: A vertical dot plot, where points have been horizontally stacked, next to a labeled box plot for the interest rates of the 50 loans.

The first step in building a box plot is drawing a dark line denoting the median, which splits the data in half. Figure 2.10 shows 50% of the data falling below the median and other 50% falling above the median. There are 50 loans in the data set (an even number) so the data are perfectly split into two groups of 25. We take the median in this case to be the average of the two observations closest to the 50^{th} percentile, which happen to be the same value in this data set: (9.93%+ 9.93%)⁄2 = 9.93%. When there are an odd number of observations, there will be exactly one observation that splits the data into two halves, and in such a case that observation is the median (no average needed).

MEDIAN: THE NUMBER IN THE MIDDLE

If the data are ordered from smallest to largest, the median is the observation right in the middle. If there are an even number of observations, there will be two values in the middle, and the median is taken as their average.

The second step in building a box plot is drawing a rectangle to represent the middle 50% of the data. The total length of the box, shown vertically in Figure 2.10, is called the interquartile range (IQR, for short). It, like the standard deviation, is a measure of variability in data. The more variable the data, the larger the standard deviation and IQR tend to be. The two boundaries of the box are called the first quartile (the 25^{th} percentile, i.e. 25% of the data fall below this value) and the third quartile (the 75^{th} percentile), and these are often labeled Q_1 and Q_3, respectively.

INTERQUARTILE RANGE (IQR)

The IQR is the length of the box in a box plot. It is computed as

$$IQR = Q_3 - Q_1$$

where Q_1 and Q_3 are the 25^{th} and 75^{th} percentiles.

GUIDED PRACTICE 2.15

What percent of the data fall between Q_1 and the median? What percent is between the median and Q_3?[10]

Extending out from the box, the whiskers attempt to capture the data outside of the box. However, their reach is never allowed to be more than $1.5 \times IQR$. They capture everything within this reach. In Figure 2.10, the upper whisker does not extend to the last two points, which is beyond $Q_3 + 1.5 \times IQR$, and so it extends only to the last point below this limit. The lower whisker stops at the lowest value, 5.31%, since there is no additional data to reach; the lower whisker's limit is not shown in the figure because the plot does not extend down to $Q_1 - 1.5 \times IQR$. In a sense, the box is like the body of the box plot and the whiskers are like its arms trying to reach the rest of the data.

Any observation lying beyond the whiskers is labeled with a dot. The purpose of labeling these points – instead of extending the whiskers to the minimum and maximum observed values – is to help identify any observations that appear to be unusually distant from the rest of the data. Unusually distant observations are called outliers. In this case, it would be reasonable to classify the interest rates of 24.85% and 26.30% as outliers since they are numerically distant from most of the data.

OUTLIERS ARE EXTREME

An outlier is an observation that appears extreme relative to the rest of the data.

Examining data for outliers serves many useful purposes, including

1. Identifying strong skew in the distribution.
2. Identifying possible data collection or data entry errors.
3. Providing insight into interesting properties of the data.

GUIDED PRACTICE 2.16

Using Figure 2.10, estimate the following values for interest rate in the loan50 data set: (a) Q_1, (b) Q_3, and (c) IQR.[11]

2.1.6 Robust statistics

How are the sample statistics of the interest rate data set affected by the observation, 26.3%? What would have happened if this loan had instead been only 15%? What would happen to these summary statistics if the observation at 26.3% had been even larger, say 35%? These scenarios are plotted alongside the original data in Figure 2.11, and sample statistics are computed under each scenario in Figure 2.12.

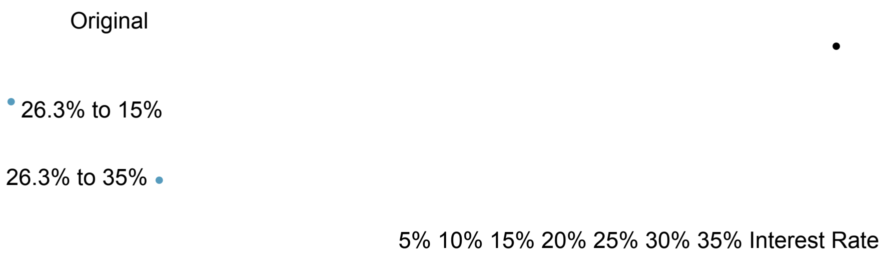


Figure 2.11: Dot plots of the original interest rate data and two modified data sets.

	robust		not robust	
scenario	median	IQR	$\bar{x}$	s
original interest rate data	9.93%	5.76%	11.57%	5.05%
move 26.3% → 15%	9.93%	5.76%	11.34%	4.61%
move 26.3% → 35%	9.93%	5.76%	11.74%	5.68%

Figure 2.12: A comparison of how the median, IQR, mean ($\bar{x}$), and standard deviation (s) change had an extreme observations from the interest rate variable been different.

GUIDED PRACTICE 2.17

(a) Which is more affected by extreme observations, the mean or median? Figure 2.12 may be helpful. (b) Is the standard deviation or IQR more affected by extreme observations?[12]

The median and IQR are called robust statistics because extreme observations have little effect on their values: moving the most extreme value generally has little influence on these

[10] Since Q_1 and Q_3 capture the middle 50% of the data and the median splits the data in the middle, 25% of the data fall between Q_1 and the median, and another 25% falls between the median and Q_3.

[11] These visual estimates will vary a little from one person to the next: Q_1 = 8%, Q_3 = 14%, IQR = Q_3 −Q_1 = 6%. (The true values: Q_1 = 7.96%, Q_3 = 13.72%, IQR = 5.76%.)

statistics. On the other hand, the mean and standard deviation are more heavily influenced by changes in extreme observations, which can be important in some situations.

EXAMPLE 2.18

The median and IQR did not change under the three scenarios in Figure 2.12. Why might this be the case?

The median and IQR are only sensitive to numbers near Q_1, the median, and Q_3. Since values in these regions are stable in the three data sets, the median and IQR estimates are also stable.

GUIDED PRACTICE 2.19

The distribution of loan amounts in the loan50 data set is right skewed, with a few large loans lingering out into the right tail. If you were wanting to understand the typical loan size, should you be more interested in the mean or median?[13]

[12] (a) Mean is affected more. (b) Standard deviation is affected more. Complete explanations are provided in the material following Guided Practice 2.17.

[13] Answers will vary! If we're looking to simply understand what a typical individual loan looks like, the median is probably more useful. However, if the goal is to understand something that scales well, such as the total amount of money we might need to have on hand if we were to offer 1,000 loans, then the mean would be more useful.

2.1.7 Transforming data (special topic)

When data are very strongly skewed, we sometimes transform them so they are easier to model.

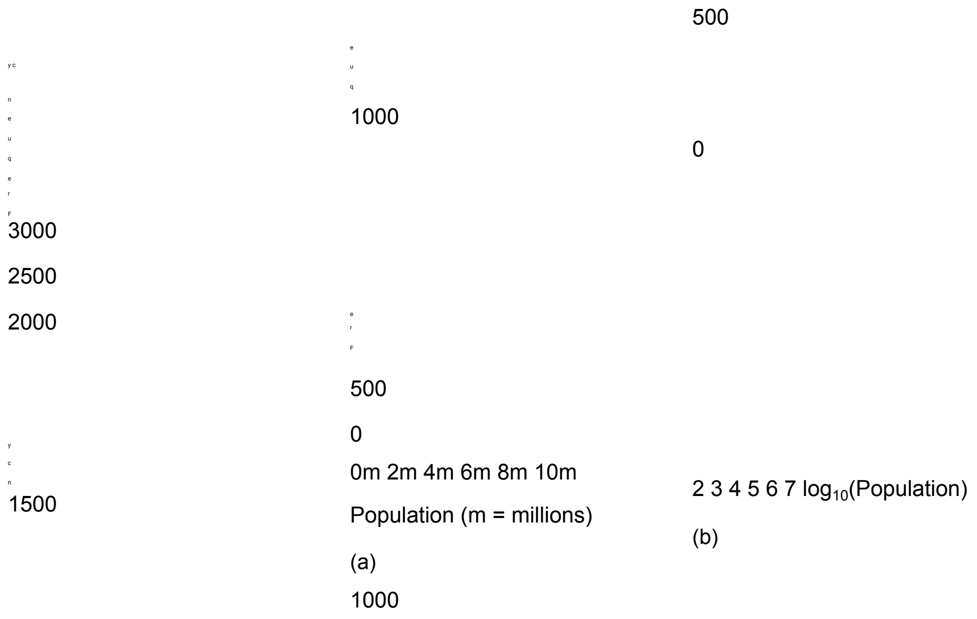


Figure 2.13: (a) A histogram of the populations of all US counties. (b) A histogram of $\log_{10}$-transformed county populations. For this plot, the x-value corresponds to the power of 10, e.g. "4" on the x-axis corresponds to 10^4 = 10,000.

EXAMPLE 2.20

Consider the histogram of county populations shown in Figure 2.13(a), which shows extreme skew. What isn’t useful about this plot?

Nearly all of the data fall into the left-most bin, and the extreme skew obscures many of the potentially interesting details in the data.

There are some standard transformations that may be useful for strongly right skewed data where much of the data is positive but clustered near zero. A transformation is a rescaling of the data using a function. For instance, a plot of the logarithm (base 10) of county populations results in the new histogram in Figure 2.13(b). This data is symmetric, and any potential outliers appear much less extreme than in the original data set. By reigning in the outliers and extreme skew, transformations like this often make it easier to build statistical models against the data.

Transformations can also be applied to one or both variables in a scatterplot. A scatterplot of the population change from 2010 to 2017 against the population in 2010 is shown in Figure 2.14(a). In this first scatterplot, it’s hard to decipher any interesting patterns because the population variable is so strongly skewed. However, if we apply a $\log_{10}$ transformation to the population variable, as shown in Figure 2.14(b), a positive association between the variables is revealed. In fact, we may be interested in fitting a trend line to the data when we explore methods around fitting regression lines in Chapter 8.

Transformations other than the logarithm can be useful, too. For instance, the square root ($\sqrt{\text{original observation}}$) and inverse ($\frac{1}{\text{original observation}}$) are commonly used by data scientists. Common goals in transforming data are to see the data structure differently, reduce skew, assist in modeling, or straighten a nonlinear relationship in a scatterplot.

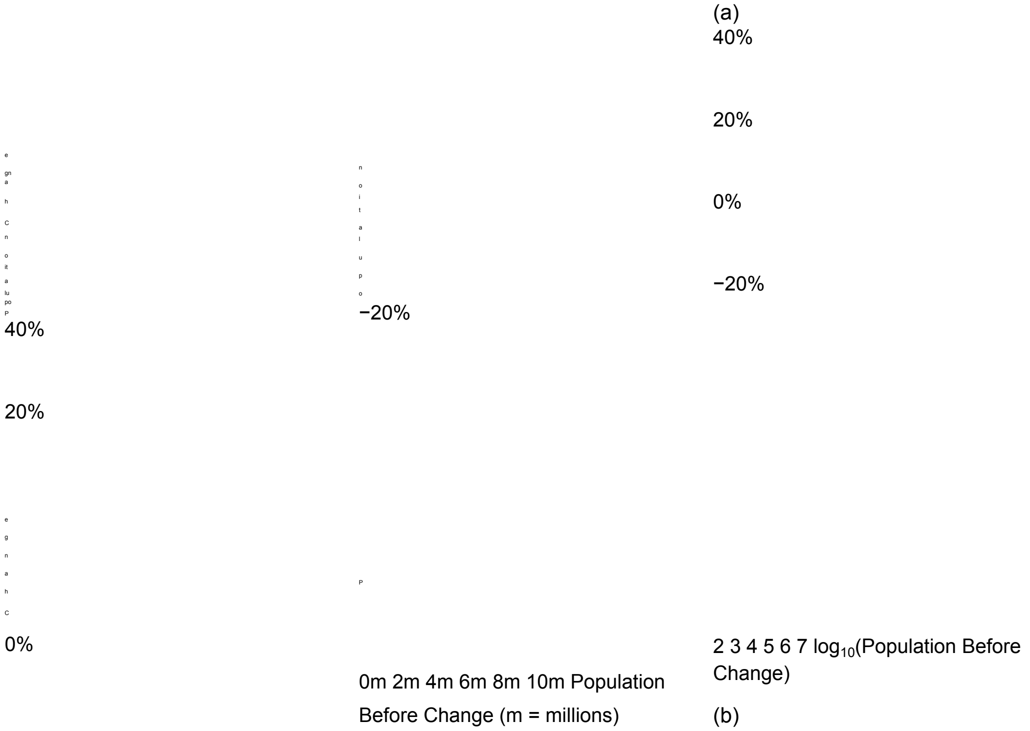


Figure 2.14: (a) Scatterplot of population change against the population before the change. (b) A scatterplot of the same data but where the population size has been log-transformed.

2.1.8 Mapping data (special topic)

The county data set offers many numerical variables that we could plot using dot plots, scatter plots, or box plots, but these miss the true nature of the data. Rather, when we encounter geographic data, we should create an intensity map, where colors are used to show higher and lower values of a variable. Figures 2.15 and 2.16 shows intensity maps for poverty rate in percent (poverty), unemployment rate (unemployment rate), homeownership rate in percent (homeownership), and median household income (median hh income). The color key indicates which colors correspond to which values. The intensity maps are not generally very helpful for getting precise values in any given county, but they are very helpful for seeing geographic trends and generating interesting research questions or hypotheses.

EXAMPLE 2.21

What interesting features are evident in the poverty and unemployment rate intensity maps?

Poverty rates are evidently higher in a few locations. Notably, the deep south shows higher poverty rates, as does much of Arizona and New Mexico. High poverty rates are evident in the Mississippi flood plains a little north of New Orleans and also in a large section of Kentucky. The unemployment rate follows similar trends, and we can see correspondence between the two variables. In fact, it makes sense for higher rates of unemployment to be closely related to poverty rates. One observation that stand out when comparing the two maps: the poverty rate is much higher than the unemployment rate, meaning while many people may be working, they are not making enough to break out of poverty.

GUIDED PRACTICE 2.22

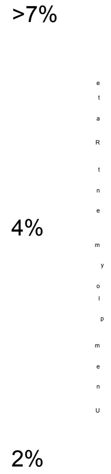
What interesting features are evident in the median hh income intensity map in Figure 2.16(b)?[14]

[14] Note: answers will vary. There is some correspondence between high earning and metropolitan areas, where we can see darker spots (higher median household income), though there are several exceptions. You might look for large cities you are familiar with and try to spot them on the map as dark spots.

>25%

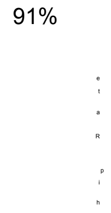
14%

(a)



(b)

Figure 2.15: (a) Intensity map of poverty rate (percent). (b) Map of the unemployment rate (percent).



median household income ($1000s).

73% <55%

(a)

>$75 $47

$19

Homeowners

Median Household income

(b)

Figure 2.16: (a) Intensity map of homeownership rate (percent). (b) Intensity map of

Exercises

2.1 Mammal life spans. Data were collected on life spans (in years) and gestation lengths (in days) for 62 mammals. A scatterplot of life span versus length of gestation is shown below.[15]

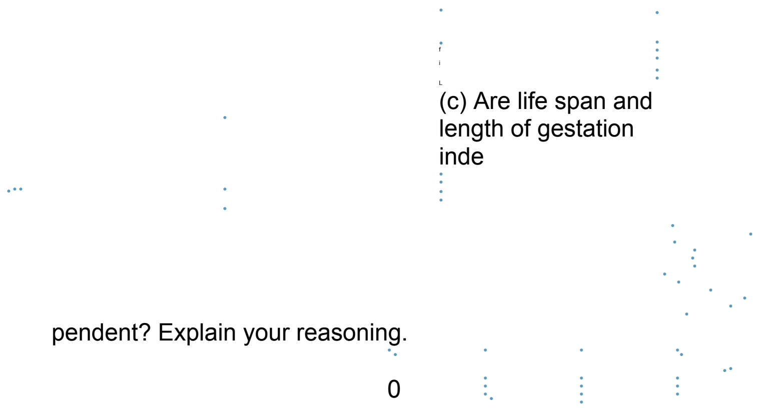
(a) What type of an association is apparent between life span and length of gestation?

(b) What type of an association would you expect to see if the axes of the plot were reversed, i.e. if we plotted length of gestation versus life span?

100 75

50

25



(c) Are life span and length of gestation independent? Explain your reasoning.

2.2 Associations. Indicate which of the plots show (a) a positive association, (b) a negative association, or (c) no association. Also determine if the positive and negative associations are linear or nonlinear. Each part may refer to more than one plot.

(1) (2) (3) (4)

2.3 Reproducing bacteria. Suppose that there is only sufficient space and nutrients to support one million bacterial cells in a petri dish. You place a few bacterial cells in this petri dish, allow them to reproduce freely, and record the number of bacterial cells in the dish over time. Sketch a plot representing the relationship between number of bacterial cells and time.

2.4 Office productivity. Office productivity is relatively low when the employees feel no stress about their work or job security. However, high levels of stress can also lead to reduced employee productivity. Sketch a plot to represent the relationship between stress and productivity.

2.5 Parameters and statistics. Identify which value represents the sample mean and which value represents the claimed population mean.

(a) American households spent an average of about $52 in 2007 on Halloween merchandise such as costumes, decorations and candy. To see if this number had changed, researchers conducted a new survey in 2008 before industry numbers were reported. The survey included 1,500 households and found that average Halloween spending was $58 per household.

(b) The average GPA of students in 2001 at a private university was 3.37. A survey on a sample of 203 students from this university yielded an average GPA of 3.59 a decade later.

2.6 Sleeping in college. A recent article in a college newspaper stated that college students get an average of 5.5 hrs of sleep each night. A student who was skeptical about this value decided to conduct a survey by randomly sampling 25 students. On average, the sampled students slept 6.25 hours per night. Identify which value represents the sample mean and which value represents the claimed population mean.

[15] T. Allison and D.V. Cicchetti. "Sleep in mammals: ecological and constitutional correlates". In: *Arch. Hydrobiol* 75 (1975), p. 442.

2.7 Days off at a mining plant. Workers at a particular mining site receive an average of 35 days paid vacation, which is lower than the national average. The manager of this plant is under pressure from a local union to increase the amount of paid time off. However, he does not want to give more days off to the workers because that would be costly. Instead he decides he should fire 10 employees in such a way as to raise the average number of days off that are reported by his employees. In order to achieve this goal, should he fire employees who have the most number of days off, least number of days off, or those who have about the average number of days off?

2.8 Medians and IQRs. For each part, compare distributions (1) and (2) based on their medians and IQRs. You do not need to calculate these statistics; simply state how the medians and IQRs compare. Make sure to explain your reasoning.

(a) (1) 3, 5, 6, 7, 9 (2) 3, 5, 6, 7, 20 (b) (1) 3, 5, 6, 7, 9 (2) 3, 5, 7, 8, 9
(c) (1) 1, 2, 3, 4, 5

(2) 6, 7, 8, 9, 10
(d) (1) 0, 10, 50, 60, 100 (2) 0, 100, 500, 600, 1000

2.9 Means and SDs. For each part, compare distributions (1) and (2) based on their means and standard deviations. You do not need to calculate these statistics; simply state how the means and the standard deviations compare. Make sure to explain your reasoning. *Hint:* It may be useful to sketch dot plots of the distributions.

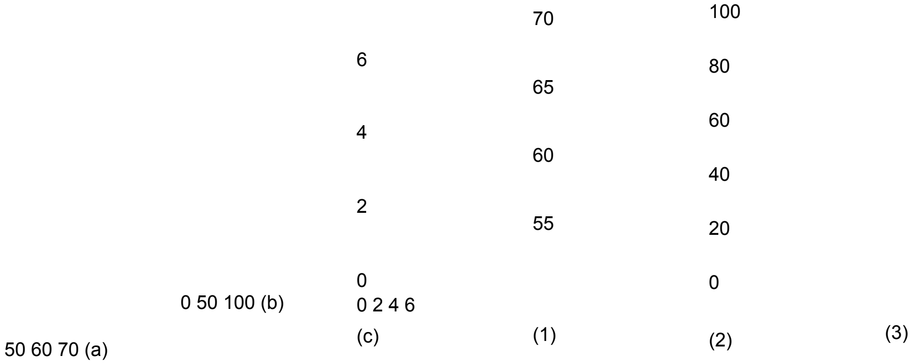
(a) (1) 3, 5, 5, 5, 8, 11, 11, 11, 13
(2) 3, 5, 5, 5, 8, 11, 11, 11, 20

(b) (1) -20, 0, 0, 0, 15, 25, 30, 30
(2) -40, 0, 0, 0, 15, 25, 30, 30

(c) (1) 0, 2, 4, 6, 8, 10
(2) 20, 22, 24, 26, 28, 30

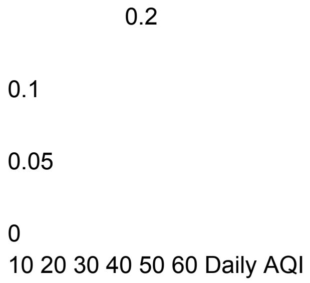
(d) (1) 100, 200, 300, 400, 500
(2) 0, 50, 300, 550, 600

2.10 Mix-and-match. Describe the distribution in the histograms below and match them to the box plots.



2.11 Air quality. Daily air quality is measured by the air quality index (AQI) reported by the Environ mental Protection Agency. This index reports the pollution level and what associated health effects might be a concern. The index is calculated for five major air pollutants regulated by the Clean Air Act and takes values from 0 to 300, where a higher value indicates lower air quality. AQI was reported for a sample of 91 days in 2011 in Durham, NC. The relative frequency histogram below shows the distribution of the AQI values on these days.[16]

(a) Estimate the median AQI value of this sample.
(b) Would you expect the mean AQI value of this sample to be higher or lower than the median? Explain your reasoning.
(c) Estimate Q1, Q3, and IQR for the distribution.
(d) Would any of the days in this sample be considered to have an unusually low or high AQI? Explain your reasoning.



2.12 Median vs. mean. Estimate the median for the 400 observations shown in the histogram, and note whether you expect the mean to be higher or lower than the median.

80

60

40

20

0

40 50 60 70 80 90 100

2.13 Histograms vs. box plots. Compare the two plots below. What characteristics of the distribution are apparent in the histogram and not in the box plot? What characteristics are apparent in the box plot but not in the histogram?

200

150

100

50

0

5 10 15 20 25

25 20 15 10 5

2.14 Facebook friends. Facebook data indicate that 50% of Facebook users have 100 or more friends, and that the average friend count of users is 190. What do these findings suggest about the shape of the distribution of number of friends of Facebook users?[17]

2.15 Distributions and appropriate statistics, Part I. For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

(a) Number of pets per household.

(b) Distance to work, i.e. number of miles between work and home.

(c) Heights of adult males.

[16] US Environmental Protection Agency, AirData, 2011.

[17] Lars Backstrom. "Anatomy of Facebook". In: *Facebook Data Team's Notes* (2011).

2.16 Distributions and appropriate statistics, Part II. For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

(a) Housing prices in a country where 25% of the houses cost below $350,000, 50% of the houses cost below $450,000, 75% of the houses cost below $1,000,000 and there are a meaningful number of houses that cost more than $6,000,000.

(b) Housing prices in a country where 25% of the houses cost below $300,000, 50% of the houses cost below $600,000, 75% of the houses cost below $900,000 and very few houses that cost more than $1,200,000. (c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

(d) Annual salaries of the employees at a Fortune 500 company where only a few high level executives earn much higher salaries than all the other employees.

2.17 Income at the coffee shop. The first histogram below shows the distribution of the yearly incomes of 40 patrons at a college coffee shop. Suppose two new people walk into the coffee shop: one making $225,000 and the other $250,000. The second histogram shows the new income distribution. Summary statistics are also provided.

12

8

4

0

$60k $62.5k $65k $67.5k $70k (1)

12

8

4

0

$60k $110k $160k $210k $260k (2)

(1) (2)
n 40 42 Min. 60,680 60,680 1st Qu. 63,620
63,710 Median 65,240 65,350 Mean 65,090
73,300 3rd Qu. 66,160 66,540 Max. 69,890
250,000 SD 2,122 37,321

(a) Would the mean or the median best represent what we might think of as a typical income for the 42 patrons at this coffee shop? What does this say about the robustness of the two measures? (b) Would the standard deviation or the IQR best represent the amount of variability in the incomes of the 42 patrons at this coffee shop? What does this say about the robustness of the two measures?

2.18 Midrange. The *midrange* of a distribution is defined as the average of the maximum and the minimum of that distribution. Is this statistic robust to outliers and extreme skew? Explain your reasoning

2.19 Commute times. The US census collects data on time it takes Americans to commute to work, among many other variables. The histogram below shows the distribution of average commute times in 3,142 US counties in 2010. Also shown below is a spatial intensity map of the same data.

200

100

0

10 20 30 40

Mean work travel (in min)

>33

19

4

(a) Describe the numerical distribution and comment on whether or not a log transformation may be advisable for these data.

(b) Describe the spatial distribution of commuting times using the map below.

2.20 Hispanic population. The US census collects data on race and ethnicity of Americans, among many other variables. The histogram below shows the distribution of the percentage of the population that is Hispanic in 3,142 counties in the US in 2010. Also shown is a histogram of logs of these values.

2000

1500

1000

500

0

0 20 40 60 80 100 Hispanic %

100

50

0

−2 −1 0 1 2 3 4 log(% Hispanic)

250

200

150

>40

20

0

(a) Describe the numerical distribution and comment on why we might want to use log-transformed values in analyzing or modeling these data.

(b) What features of the distribution of the Hispanic population in US counties are apparent in the map but not in the histogram? What features are apparent in the histogram but not the map? (c) Is one visualization more appropriate or helpful than the other? Explain your reasoning.

2.2 Considering categorical data

In this section, we will introduce tables and other basic tools for categorical data that are used throughout this book. The loan50 data set represents a sample from a larger loan data set called loans. This larger data set contains information on 10,000 loans made through Lending Club. We will examine the relationship between homeownership, which for the loans data can take a value of rent, mortgage (owns but has a mortgage), or own, and app type, which indicates whether the loan application was made with a partner or whether it was an individual application.

2.2.1 Contingency tables and bar plots

Figure 2.17 summarizes two variables: app type and homeownership. A table that summarizes data for two categorical variables in this way is called a contingency table. Each value in the table represents the number of times a particular combination of variable outcomes occurred. For example, the value 3496 corresponds to the number of loans in the data set where the borrower rents their home and the application type was by an individual. Row and column totals are also included. The row totals provide the total counts across each row (e.g. 3496 + 3839 + 1170 = 8505), and column totals are total counts down each column. We can also create a table that shows only the overall percentages or proportions for each combination of categories, or we can create a table for a single variable, such as the one shown in Figure 2.18 for the homeownership variable.

		homeownership			
		rent	mortgage	own	Total
app type	individual	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

Figure 2.17: A contingency table for app type and homeownership.

homeownership	Count
rent	3858
mortgage	4789
own	1353
Total	10000

Figure 2.18: A table summarizing the frequencies of each value for the homeownership variable.

A bar plot is a common way to display a single categorical variable. The left panel of Figure 2.19 shows a bar plot for the homeownership variable. In the right panel, the counts are converted into proportions, showing the proportion of observations that are in each level (e.g. 3858/10000 = 0.3858 for rent).

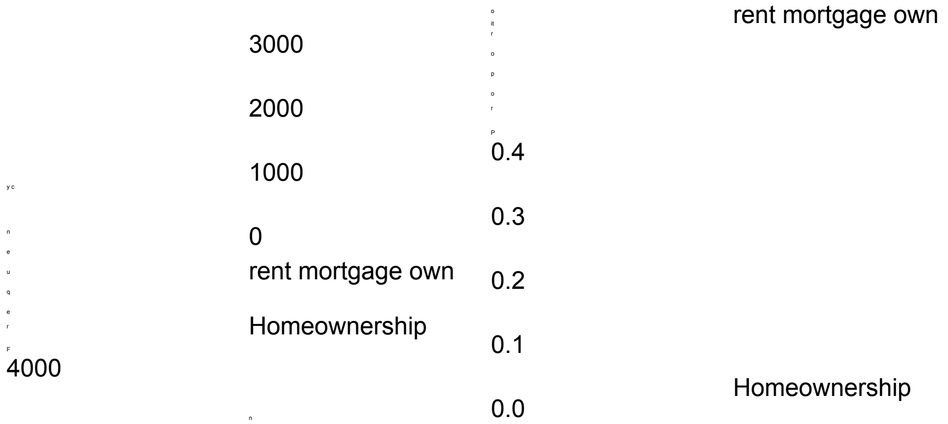


Figure 2.19: Two bar plots of number. The left panel shows the counts, and the right panel shows the proportions in each group.

2.2.2 Row and column proportions

Sometimes it is useful to understand the fractional breakdown of one variable in another, and we can modify our contingency table to provide such a view. Figure 2.20 shows the row proportions for Figure 2.17, which are computed as the counts divided by their row totals. The value 3496 at the intersection of individual and rent is replaced by 3496/8505 = 0.411, i.e. 3496 divided by its row total, 8505. So what does 0.411 represent? It corresponds to the proportion of individual applicants who rent.

	rent	mortgage	own	Total
individual	0.411	0.451	0.138	1.000
joint	0.242	0.635	0.122	1.000
Total	0.386	0.479	0.135	1.000

Figure 2.20: A contingency table with row proportions for the app type and homeownership variables. The row total is off by 0.001 for the joint row due to a rounding error.

A contingency table of the column proportions is computed in a similar way, where each column proportion is computed as the count divided by the corresponding column total. Figure 2.21 shows such a table, and here the value 0.906 indicates that 90.6% of renters applied as individuals for the loan. This rate is higher compared to loans from people with mortgages (80.2%) or who own their home (86.5%). Because these rates vary between the three levels of homeownership (rent, mortgage, own), this provides evidence that the app type and homeownership variables are associated.

	rent	mortgage	own	Total
individual	0.906	0.802	0.865	0.851
joint	0.094	0.198	0.135	0.150
Total	1.000	1.000	1.000	1.000

Figure 2.21: A contingency table with column proportions for the app type and homeownership variables. The total for the last column is off by 0.001 due to a rounding error.

We could also have checked for an association between app type and homeownership in Figure 2.20 using row proportions. When comparing these row proportions, we would look down columns to see if the fraction of loans where the borrower rents, has a mortgage, or owns varied across the individual to joint application types.

GUIDED PRACTICE 2.23

(a) What does 0.451 represent in Figure 2.20?
(b) What does 0.802 represent in Figure 2.21?[18]

GUIDED PRACTICE 2.24

(a) What does 0.122 at the intersection of joint and own represent in Figure 2.20?
(b) What does 0.135 represent in the Figure 2.21?[19]

EXAMPLE 2.25

Data scientists use statistics to filter spam from incoming email messages. By noting specific characteristics of an email, a data scientist may be able to classify some emails as spam or not spam with high accuracy. One such characteristic is whether the email contains no numbers, small numbers, or big numbers. Another characteristic is the email format, which indicates whether or not an email has any HTML content, such as bolded text. We'll focus on email format and spam status using the email data set, and these variables are summarized in a contingency table in Figure 2.22. Which would be more helpful to someone hoping to classify email as spam or regular email for this table: row or column proportions?

A data scientist would be interested in how the proportion of spam changes within each email format. This corresponds to column proportions: the proportion of spam in plain text emails and the proportion of spam in HTML emails.

If we generate the column proportions, we can see that a higher fraction of plain text emails are spam (209/1195 = 17.5%) than compared to HTML emails (158/2726 = 5.8%). This information on its own is insufficient to classify an email as spam or not spam, as over 80% of plain text emails are not spam. Yet, when we carefully combine this information with many other characteristics, we stand a reasonable chance of being able to classify some emails as spam or not spam with confidence.

	text	HTML	Total
spam	209	158	367
not spam	986	2568	3554
Total	1195	2726	3921

Figure 2.22: A contingency table for spam and format.

Example 2.25 points out that row and column proportions are not equivalent. Before settling on one form for a table, it is important to consider each to ensure that the most useful table is constructed. However, sometimes it simply isn't clear which, if either, is more useful.

EXAMPLE 2.26

Look back to Tables 2.20 and 2.21. Are there any obvious scenarios where one might be more useful than the other?

None that we thought were obvious! What is distinct about app type and homeownership vs the email example is that these two variables don't have a clear explanatory-response variable relationship that we might hypothesize (see Section 1.2.4 for these terms). Usually it is most useful to "condition" on the explanatory variable. For instance, in the email example, the email format was seen as a possible explanatory variable of whether the message was spam, so we would find it more interesting to compute the relative frequencies (proportions) for each email format.

[18] (a) 0.451 represents the proportion of individual applicants who have a mortgage. (b) 0.802 represents the fraction of applicants with mortgages who applied as individuals.

[19] (a) 0.122 represents the fraction of joint borrowers who own their home. (b) 0.135 represents the home-owning borrowers who had a joint application for the loan.

2.2.3 Using a bar plot with two variables

Contingency tables using row or column proportions are especially useful for examining how two categorical variables are related. Stacked bar plots provide a way to visualize the information in these tables.

A stacked bar plot is a graphical display of contingency table information. For example, a stacked bar plot representing Figure 2.21 is shown in Figure 2.23(a), where we have first created a bar plot using the homeownership variable and then divided each group by the levels of app type.

One related visualization to the stacked bar plot is the side-by-side bar plot, where an example is shown in Figure 2.23(b).

For the last type of bar plot we introduce, the column proportions for the app type and homeownership contingency table have been translated into a standardized stacked bar plot in Figure 2.23(c). This type of visualization is helpful in understanding the fraction of individual or joint loan applications for borrowers in each level of homeownership. Additionally, since the proportions of joint and individual vary across the groups, we can conclude that the two variables are associated.

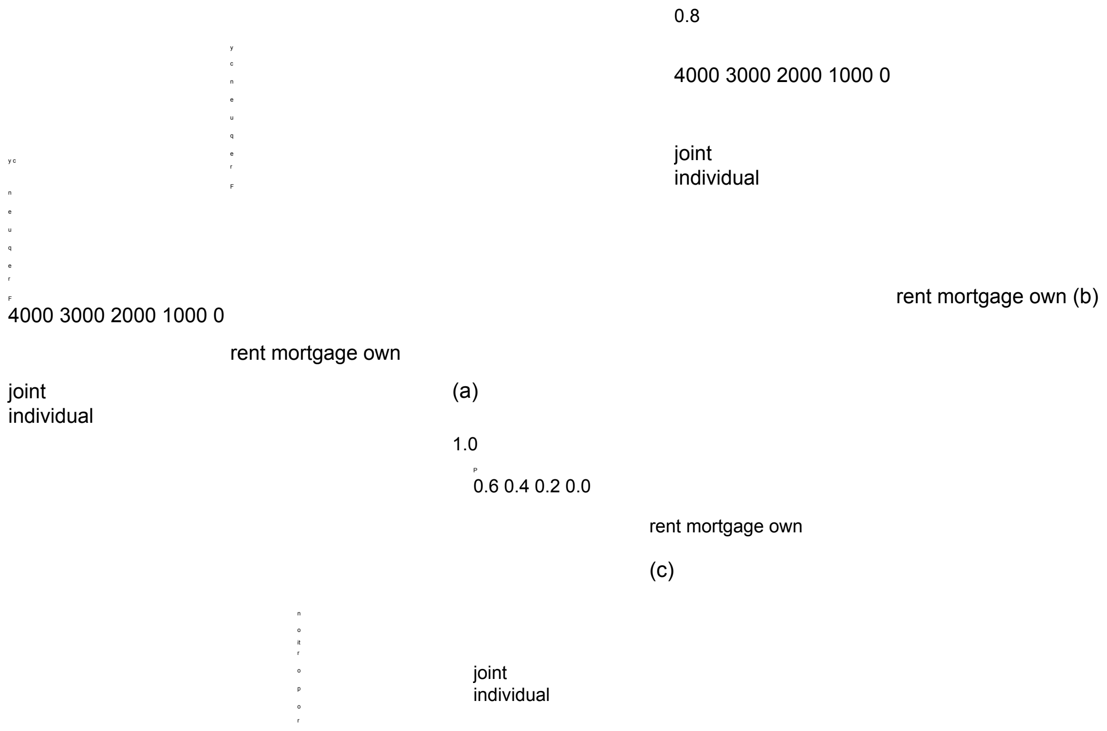


Figure 2.23: (a) Stacked bar plot for homeownership, where the counts have been further broken down by app type. (b) Side-by-side bar plot for homeownership and app type. (c) Standardized version of the stacked bar plot.

EXAMPLE 2.27

Examine the three bar plots in Figure 2.23. When is the stacked, side-by-side, or standardized stacked bar plot the most useful?

The stacked bar plot is most useful when it’s reasonable to assign one variable as the explanatory variable and the other variable as the response, since we are effectively grouping by one variable first and then breaking it down by the others.

Side-by-side bar plots are more agnostic in their display about which variable, if any, represents the explanatory and which the response variable. It is also easy to discern the number of cases in of the six different group combinations. However, one downside is that it tends to require more horizontal space; the narrowness of Figure 2.23(b) makes the plot feel a bit cramped. Additionally, when two groups are of very different sizes, as we see in the own group relative to either of the other two groups, it is difficult to discern if there is an association between the variables.

The standardized stacked bar plot is helpful if the primary variable in the stacked bar plot is relatively imbalanced, e.g. the own category has only a third of the observations in the mortgage category, making the simple stacked bar plot less useful for checking for an association. The major downside of the standardized version is that we lose all sense of how many cases each of the bars represents.

2.2.4 Mosaic plots

A mosaic plot is a visualization technique suitable for contingency tables that resembles a standardized stacked bar plot with the benefit that we still see the relative group sizes of the primary variable as well.

To get started in creating our first mosaic plot, we'll break a square into columns for each category of the homeownership variable, with the result shown in Figure 2.24(a). Each column represents a level of homeownership, and the column widths correspond to the proportion of loans in each of those categories. For instance, there are fewer loans where the borrower is an owner than where the borrower has a mortgage. In general, mosaic plots use box *areas* to represent the number of cases in each category.

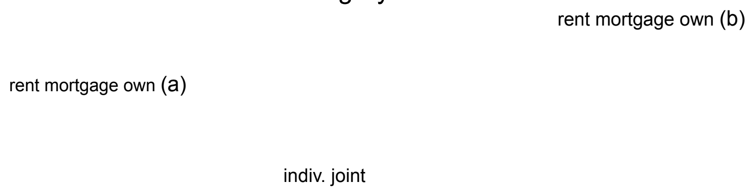


Figure 2.24: (a) The one-variable mosaic plot for homeownership. (b) Two-variable mosaic plot for both homeownership and app type.

To create a completed mosaic plot, the single-variable mosaic plot is further divided into pieces in Figure 2.24(b) using the app type variable. Each column is split proportional to the number of loans from individual and joint borrowers. For example, the second column represents loans where the borrower has a mortgage, and it was divided into individual loans (upper) and joint loans (lower). As another example, the bottom segment of the third column represents loans where the borrower owns their home and applied jointly, while the upper segment of this column represents borrowers who are homeowners and filed individually. We can again use this plot to see that the homeownership and app type variables are associated, since some columns are divided in different

vertical locations than others, which was the same technique used for checking an association in the standardized stacked bar plot.

In Figure 2.24, we chose to first split by the homeowner status of the borrower. However, we could have instead first split by the application type, as in Figure 2.25. Like with the bar plots, it's common to use the explanatory variable to represent the first split in a mosaic plot, and then for the response to break up each level of the explanatory variable, if these labels are reasonable to attach to the variables under consideration.

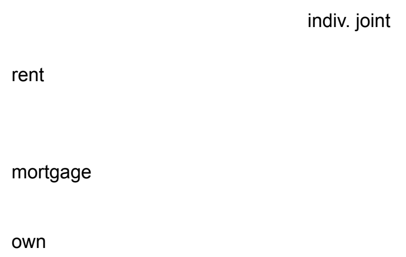


Figure 2.25: Mosaic plot where loans are grouped by the homeownership variable after they've been divided into the individual and joint application types.

2.2.5 The only pie chart you will see in this book

A pie chart is shown in Figure 2.26 alongside a bar plot representing the same information. Pie charts can be useful for giving a high-level overview to show how a set of cases break down. However, it is also difficult to decipher details in a pie chart. For example, it takes a couple seconds longer to recognize that there are more loans where the borrower has a mortgage than rent when looking at the pie chart, while this detail is very obvious in the bar plot. While pie charts can be useful, we prefer bar plots for their ease in comparing groups.

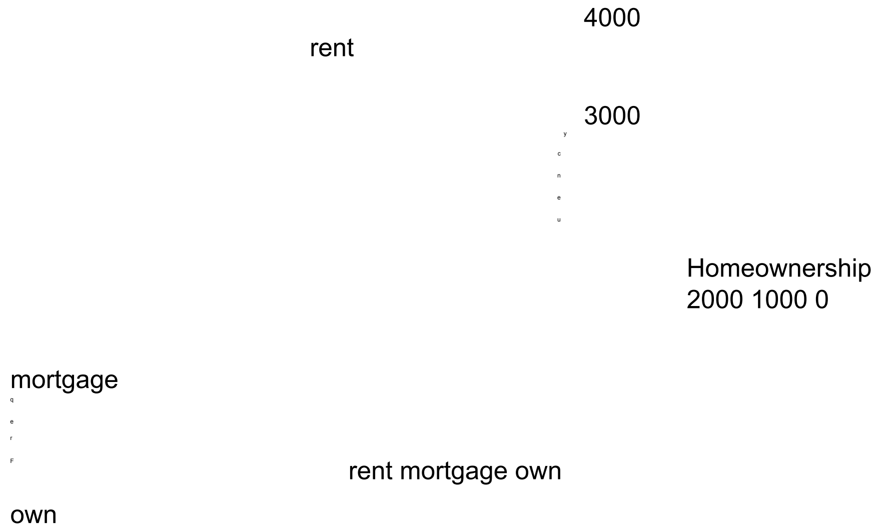


Figure 2.26: A pie chart and bar plot of homeownership.

2.2.6 Comparing numerical data across groups

Some of the more interesting investigations can be considered by examining numerical data across groups. The methods required here aren't really new: all that's required is to make a numerical plot for each group in the same graph. Here two convenient methods are introduced: side-by-side box plots and hollow histograms.

We will take a look again at the county data set and compare the median household income for counties that gained population from 2010 to 2017 versus counties that had no gain. While we might like to make a causal connection here, remember that these are observational data and so such an interpretation would be, at best, half-baked.

There were 1,454 counties where the population increased from 2010 to 2017, and there were 1,672 counties with no gain (all but one were a loss). A random sample of 100 counties from the first group and 50 from the second group are shown in Figure 2.27 to give a better sense of some of the raw median income data.

Median Income for 150 Counties, in $1000s

Population Gain No Population Gain

38.2 43.6 42.2 61.5 51.1 45.7 48.3 60.3 50.7
44.6 51.8 40.7 48.1 56.4 41.9 39.3 40.4 40.3
40.6 63.3 52.1 60.3 49.8 51.7 57 47.2 45.9
51.1 34.1 45.5 52.8 49.1 51 42.3 41.5 46.1
80.8 46.3 82.2 43.6 39.7 49.4 44.9 51.7 46.4
75.2 40.6 46.3 62.4 44.1 51.3 29.1 51.8 50.5
51.9 34.7 54 42.9 52.2 45.1 27 30.9 34.9
61 51.4 56.5 62 46 46.4 40.7 51.8 61.1
53.8 57.6 69.2 48.4 40.5 48.6 43.4 34.7 45.7
53.1 54.6 55 46.4 39.9 56.7 33.1 21 37
63 49.1 57.2 44.1 50 38.9 52 31.9 45.7
46.6 46.5 38.9 50.9 56 34.6 56.3 38.7 45.7
74.2 63 49.6 53.7 77.5 60 56.2 43 21.7
63.2 47.6 55.9 39.1 57.8 42.6 44.5 34.5 48.9
50.4 49 45.6 39 38.8 37.1 50.9 42.1 43.2
57.2 44.7 71.7 35.3 100.2 35.4 41.3 33.6
42.6 55.5 38.6 52.7 63 43.4 56.5

Figure 2.27: In this table, median household income (in $1000s) from a random sample of 100 counties that had population gains are shown on the left. Median incomes from a random sample of 50 counties that had no population gain are shown on the right.

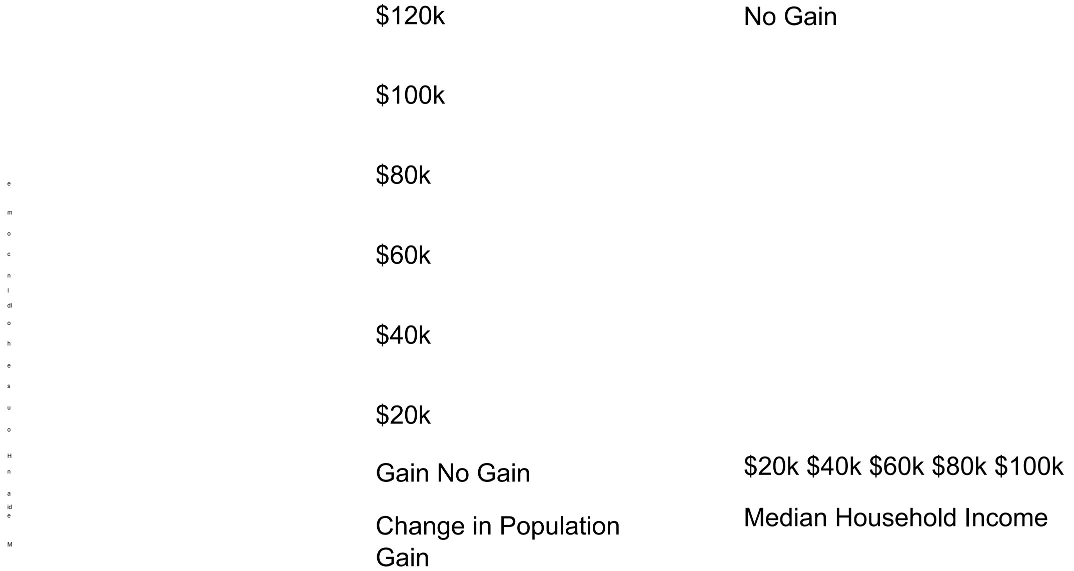


Figure 2.28: Side-by-side box plot (left panel) and hollow histograms (right panel) for med hh income, where the counties are split by whether there was a population gain or loss.

The side-by-side box plot is a traditional tool for comparing across groups. An example is shown in the left panel of Figure 2.28, where there are two box plots, one for each group, placed into one plotting window and drawn on the same scale.

Another useful plotting method uses hollow histograms to compare numerical data across groups. These are just the outlines of histograms of each group put on the same plot, as shown in the right panel of Figure 2.28.

GUIDED PRACTICE 2.28

Use the plots in Figure 2.28 to compare the incomes for counties across the two groups. What do you notice about the approximate center of each group? What do you notice about the variability between groups? Is the shape relatively consistent between groups? How many *prominent* modes are there for each group?[20]

GUIDED PRACTICE 2.29

What components of each plot in Figure 2.28 do you find most useful?[21]

[20] Answers may vary a little. The counties with population gains tend to have higher income (median of about $45,000) versus counties without a gain (median of about $40,000). The variability is also slightly larger for the population gain group. This is evident in the IQR, which is about 50% bigger in the *gain* group. Both distributions show slight to moderate right skew and are unimodal. The box plots indicate there are many observations far above the median in each group, though we should anticipate that many observations will fall beyond the whiskers when examining any data set that contain more than a couple hundred data points.

[21] Answers will vary. The side-by-side box plots are especially useful for comparing centers and spreads, while the hollow histograms are more useful for seeing distribution shape, skew, and potential anomalies.

Exercises

2.21 Antibiotic use in children. The bar plot and the pie chart below show the distribution of pre-existing medical conditions of children involved in a study on the optimal duration of antibiotic use in treatment of tracheitis, which is an upper respiratory infection.

Prematurity
Cardiovascular
Respiratory
Trauma
Neuromuscular
Genetic/metabolic

Immunocompromised
Gastrointestinal

0.00 0.10 0.20 0.30 Relative frequency

Trauma

Respiratory

Cardiovascular

Neuromuscular

Genetic/metabolic

Immunocompromised
Gastrointestinal

Prematurity

(a) What features are apparent in the bar plot but not in the pie chart?

(b) What features are apparent in the pie chart but not in the bar plot?

(c) Which graph would you prefer to use for displaying these categorical data?

2.22 Views on immigration. 910 randomly sampled registered voters from Tampa, FL were asked if they thought workers who have illegally entered the US should be (i) allowed to keep their jobs and apply for US citizenship, (ii) allowed to keep their jobs as temporary guest workers but not allowed to apply for US citizenship, or (iii) lose their jobs and have to leave the country. The results of the survey by political ideology are shown below.[22]

Political ideology
Conservative Moderate Liberal Total
(i) Apply for citizenship 57 120 101 278
(ii) Guest worker 121 113 28 262 *Response* (iii) Leave the country 179 126 45 350 (iv) Not sure 15 4 1 20

Total 372 363 175 910

(a) What percent of these Tampa, FL voters identify themselves as conservatives? (b) What percent of these Tampa, FL voters are in favor of the citizenship option? (c) What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option?

(d) What percent of these Tampa, FL voters who identify themselves as conservatives are also in favor of the citizenship option? What percent of moderates share this view? What percent of liberals share this view?

(e) Do political ideology and views on immigration appear to be independent? Explain your reasoning. [22]SurveyUSA,

News Poll #18927, data collected Jan 27-29, 2012.

2.23 Views on the DREAM Act. A random sample of registered voters from Tampa, FL were asked if they support the DREAM Act, a proposed law which would provide a path to citizenship for people brought illegally to the US as children. The survey also collected information on the political ideology of the respondents. Based on the mosaic plot shown below, do views on the DREAM Act and political ideology appear to be independent? Explain your reasoning.[23]

Conservative Moderate Liberal

Support

Not support

Not sure

2.24 Raise taxes. A random sample of registered voters nationally were asked whether they think it's better to raise taxes on the rich or raise taxes on the poor. The survey also collected information on the political party affiliation of the respondents. Based on the mosaic plot shown below, do views on raising taxes and political affiliation appear to be independent? Explain your reasoning.[24]

Democrat Republican Indep / Other

Raise taxes on the rich

Raise taxes on the poor
Not sure

[23] SurveyUSA, News Poll #18927, data collected Jan 27-29, 2012.

[24] Public Policy Polling, Americans on College Degrees, Classic Literature, the Seasons, and More, data collected Feb 20-22, 2015.

2.3 Case study: malaria vaccine

EXAMPLE 2.30

Suppose your professor splits the students in class into two groups: students on the left and students on the right. If $\hat{p}_L$ and $\hat{p}_R$ represent the proportion of students who own an Apple product on the left and right, respectively, would you be surprised if $\hat{p}_L$ did not exactly equal $\hat{p}_R$?

While the proportions would probably be close to each other, it would be unusual for them to be exactly the same. We would probably observe a small difference due to chance.

GUIDED PRACTICE 2.31

If we don't think the side of the room a person sits on in class is related to whether the person owns an Apple product, what assumption are we making about the relationship between these two variables?[25]

2.3.1 Variability within data

We consider a study on a new malaria vaccine called PfSPZ. In this study, volunteer patients were randomized into one of two experiment groups: 14 patients received an experimental vaccine or 6 patients received a placebo vaccine. Nineteen weeks later, all 20 patients were exposed to a drug-sensitive malaria virus strain; the motivation of using a drug-sensitive strain of virus here is for ethical considerations, allowing any infections to be treated effectively. The results are summarized in Figure 2.29, where 9 of the 14 treatment patients remained free of signs of infection while all of the 6 patients in the control group patients showed some baseline signs of infection.

		outcome		
		infection	no infection	Total
treatment	vaccine	5	9	14
	placebo	6	0	6
	Total	11	9	20

Figure 2.29: Summary results for the malaria vaccine experiment.

GUIDED PRACTICE 2.32

Is this an observational study or an experiment? What implications does the study type have on what can be inferred from the results?[26]

In this study, a smaller proportion of patients who received the vaccine showed signs of an infection (35.7% versus 100%). However, the sample is very small, and it is unclear whether the difference provides *convincing evidence* that the vaccine is effective.

[25] We would be assuming that these two variables are independent.

[26] The study is an experiment, as patients were randomly assigned an experiment group. Since this is an experiment, the results can be used to evaluate a causal relationship between the malaria vaccine and whether patients showed signs of an infection.

EXAMPLE 2.33

Data scientists are sometimes called upon to evaluate the strength of evidence. When looking at the rates of infection for patients in the two groups in this study, what comes to mind as we try to determine whether the data show convincing evidence of a real difference?

The observed infection rates (35.7% for the treatment group versus 100% for the control group) suggest the vaccine may be effective. However, we cannot be sure if the observed difference represents the vaccine's efficacy or is just from random chance. Generally there is a little bit of fluctuation in sample data, and we wouldn't expect the sample proportions to be *exactly* equal, even if the truth was that the infection rates were independent of getting the vaccine. Additionally, with such small samples, perhaps it's common to observe such large differences when we randomly split a group due to chance alone!

Example 2.33 is a reminder that the observed outcomes in the data sample may not perfectly reflect the true relationships between variables since there is random noise. While the observed difference in rates of infection is large, the sample size for the study is small, making it unclear if this observed difference represents efficacy of the vaccine or whether it is simply due to chance. We label these two competing claims, H_0 and H_A, which are spoken as "H-nought" and "H-A":

H_0: Independence model. The variables treatment and outcome are independent. They have no relationship, and the observed difference between the proportion of patients who developed an infection in the two groups, 64.3%, was due to chance.

H_A: Alternative model. The variables are *not* independent. The difference in infection rates of 64.3% was not due to chance, and vaccine affected the rate of infection.

What would it mean if the independence model, which says the vaccine had no influence on the rate of infection, is true? It would mean 11 patients were going to develop an infection *no matter which group they were randomized into*, and 9 patients would not develop an infection *no matter which group they were randomized into*. That is, if the vaccine did not affect the rate of infection, the difference in the infection rates was due to chance alone in how the patients were randomized.

Now consider the alternative model: infection rates were influenced by whether a patient re ceived the vaccine or not. If this was true, and especially if this influence was substantial, we would expect to see some difference in the infection rates of patients in the groups.

We choose between these two competing claims by assessing if the data conflict so much with H_0 that the independence model cannot be deemed reasonable. If this is the case, and the data support H_A, then we will reject the notion of independence and conclude there was discrimination.

2.3.2 Simulating the study

We're going to implement simulations, where we will pretend we know that the malaria vaccine being tested does *not* work. Ultimately, we want to understand if the large difference we observed is common in these simulations. If it is common, then maybe the difference we observed was purely due to chance. If it is very uncommon, then the possibility that the vaccine was helpful seems more plausible.

Figure 2.29 shows that 11 patients developed infections and 9 did not. For our simulation, we will suppose the infections were independent of the vaccine and we were able to *rewind* back to when the researchers randomized the patients in the study. If we happened to randomize the patients differently, we may get a different result in this hypothetical world where the vaccine doesn't influence the infection. Let's complete another randomization using a simulation.

In this simulation, we take 20 notecards to represent the 20 patients, where we write down "infection" on 11 cards and "no infection" on 9 cards. In this hypothetical world, we believe each patient that got an infection was going to get it regardless of which group they were in, so let's see what happens if we randomly assign the patients to the treatment and control groups again. We thoroughly shuffle the notecards and deal 14 into a vaccine pile and 6 into a placebo pile. Finally, we tabulate the results, which are shown in Figure 2.30.

		outcome		
		infection	no infection	Total
treatment	vaccine	7	7	14
(simulated)	placebo	4	2	6
	Total	11	9	20

Figure 2.30: Simulation results, where any difference in infection rates is purely

due to chance.

GUIDED PRACTICE 2.34

What is the difference in infection rates between the two simulated groups in Figure 2.30? How does this compare to the observed 64.3% difference in the actual data?[27]

2.3.3 Checking for independence

We computed one possible difference under the independence model in Guided Practice 2.34, which represents one difference due to chance. While in this first simulation, we physically dealt out notecards to represent the patients, it is more efficient to perform this simulation using a computer. Repeating the simulation on a computer, we get another difference due to chance:

9
6− 14= −0.310
2

And another:

3
8
6− 14= −0.071

And so on until we repeat the simulation enough times that we have a good idea of what represents the *distribution of differences from chance alone*. Figure 2.31 shows a stacked plot of the differences found from 100 simulations, where each dot represents a simulated difference between the infection rates (control rate minus treatment rate).

Note that the distribution of these simulated differences is centered around 0. We simulated these differences assuming that the independence model was true, and under this condition, we expect the difference to be near zero with some random fluctuation, where *near* is pretty generous in this case since the sample sizes are so small in this study.

EXAMPLE 2.35

How often would you observe a difference of at least 64.3% (0.643) according to Figure 2.31? Often, sometimes, rarely, or never?

It appears that a difference of at least 64.3% due to chance alone would only happen about 2% of the time according to Figure 2.31. Such a low probability indicates a rare event.

[27] 4/6 − 7/14 = 0.167 or about 16.7% in favor of the vaccine. This difference due to chance is much smaller than the difference observed in the actual groups.

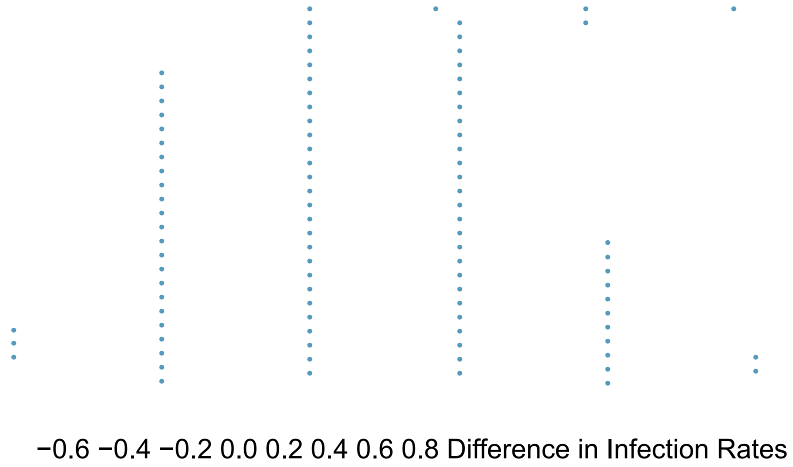


Figure 2.31: A stacked dot plot of differences from 100 simulations produced under the independence model, H_0, where in these simulations infections are unaffected by the vaccine. Two of the 100 simulations had a difference of at least 64.3%, the difference observed in the study.

The difference of 64.3% being a rare event suggests two possible interpretations of the results of the study:

H_0 Independence model. The vaccine has no effect on infection rate, and we just happened to observe a difference that would only occur on a rare occasion.

H_A Alternative model. The vaccine has an effect on infection rate, and the difference we observed was actually due to the vaccine being effective at combatting malaria, which explains the large difference of 64.3%.

Based on the simulations, we have two options. (1) We conclude that the study results do not provide strong evidence against the independence model. That is, we do not have sufficiently strong evidence to conclude the vaccine had an effect in this clinical setting. (2) We conclude the evidence is sufficiently strong to reject H_0 and assert that the vaccine was useful. When we conduct formal studies, usually we reject the notion that we just happened to observe a rare event.[28] So in this case, we reject the independence model in favor of the alternative. That is, we are concluding the data provide strong evidence that the vaccine provides some protection against malaria in this clinical setting.

One field of statistics, statistical inference, is built on evaluating whether such differences are due to chance. In statistical inference, data scientists evaluate which model is most reasonable given the data. Errors do occur, just like rare events, and we might choose the wrong model. While we do not always choose correctly, statistical inference gives us tools to control and evaluate how often these errors occur. In Chapter 5, we give a formal introduction to the problem of model selection. We spend the next two chapters building a foundation of probability and theory necessary to make that discussion rigorous.

[28] This reasoning does not generally extend to anecdotal observations. Each of us observes incredibly rare events every day, events we could not possibly hope to predict. However, in the non-rigorous setting of anecdotal evidence, almost anything may appear to be a rare event, so the idea of looking for rare events in day-to-day activities is treacherous. For example, we might look at the lottery: there was only a 1 in 292 million chance that the Powerball numbers for the largest jackpot in history (January 13th, 2016) would be (04, 08, 19, 27, 34) with a Powerball of (10), but nonetheless those numbers came up! However, no matter what numbers had turned up, they would have had the same incredibly rare odds. That is, *any set of numbers we could have observed would ultimately be incredibly rare*. This type of situation is typical of our daily lives: each possible event in itself seems incredibly rare, but if we consider every alternative, those outcomes are also incredibly rare. We should be cautious not to misinterpret such anecdotal evidence.

Exercises

2.25 Side effects of Avandia. Rosiglitazone is the active ingredient in the controversial type 2 diabetes medicine Avandia and has been linked to an increased risk of serious cardiovascular problems such as stroke, heart failure, and death. A common alternative treatment is pioglitazone, the active ingredient in a diabetes medicine called Actos. In a nationwide retrospective observational study of 227,571 Medicare beneficiaries aged 65 years or older, it was found that 2,593 of the 67,593 patients using rosiglitazone and 5,386 of the 159,978 using pioglitazone had serious cardiovascular problems. These data are summarized in the contingency table below.[29]

		Cardiovascular problems		
		Yes	No	Total
Treatment	Rosiglitazone	2,593	65,000	67,593
	Pioglitazone	5,386	154,592	159,978
	Total	7,979	219,592	227,571

(a) Determine if each of the following statements is true or false. If false, explain why. *Be careful:* The reasoning may be wrong even if the statement's conclusion is correct. In such cases, the statement should be considered false.

i. Since more patients on pioglitazone had cardiovascular problems (5,386 vs. 2,593), we can conclude that the rate of cardiovascular problems for those on a pioglitazone treatment is higher.

ii. The data suggest that diabetic patients who are taking rosiglitazone are more likely to have cardio vascular problems since the rate of incidence was (2,593 / 67,593 = 0.038) 3.8% for patients on this treatment, while it was only (5,386 / 159,978 = 0.034) 3.4% for patients on pioglitazone.

iii. The fact that the rate of incidence is higher for the rosiglitazone group proves that rosiglitazone causes serious cardiovascular problems.

iv. Based on the information provided so far, we cannot tell if the difference between the rates of incidences is due to a relationship between the two variables or due to chance.

(b) What proportion of all patients had cardiovascular problems?

(c) If the type of treatment and having cardiovascular problems were independent, about how many patients in the rosiglitazone group would we expect to have had cardiovascular problems?

(d) We can investigate the relationship between outcome and treatment in this study using a randomization technique. While in reality we would carry out the simulations required for randomization using statisti cal software, suppose we actually simulate using index cards. In order to simulate from the independence model, which states that the outcomes were independent of the treatment, we write whether or not each patient had a cardiovascular problem on cards, shuffled all the cards together, then deal them into two groups of size 67,593 and 159,978. We repeat this simulation 1,000 times and each time record the num ber of people in the rosiglitazone group who had cardiovascular problems. Use the relative frequency histogram of these counts to answer (i)-(iii).

0.2

i. What are the claims being tested?

ii. Compared to the number calculated in part (b), which would provide more support for the alterna tive hypothesis, *more* or *fewer* patients with car diovascular problems in the rosiglitazone group?

0

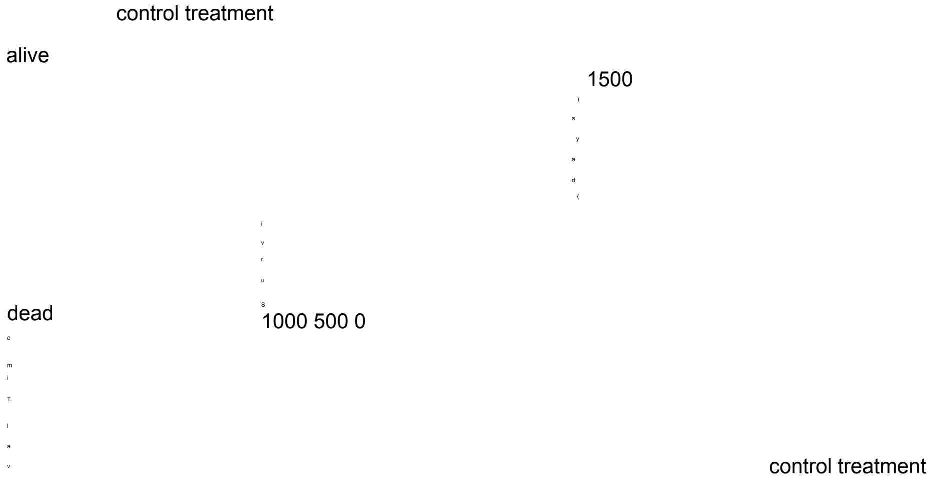
iii. What do the simulation results suggest about the relationship between taking rosiglitazone and hav ing cardiovascular problems in diabetic patients?

2250 2350 2450 Simulated rosiglitazone cardiovascular events

0.1

[29] D.J. Graham et al. "Risk of acute myocardial infarction, stroke, heart failure, and death in elderly Medicare patients treated with rosiglitazone or pioglitazone". In: *JAMA* 304.4 (2010), p. 411. issn: 0098-7484.

2.26 Heart transplants. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was designated an official heart transplant candidate, meaning that he was gravely ill and would most likely benefit from a new heart. Some patients got a transplant and some did not. The variable transplant indicates which group the patients were in; patients in the treatment group got a transplant and those in the control group did not. Of the 34 patients in the control group, 30 died. Of the 69 people in the treatment group, 45 died. Another variable called survived was used to indicate whether or not the patient was alive at the end of the study.[30]



(a) Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.

(b) What do the box plots below suggest about the efficacy (effectiveness) of the heart transplant treatment.

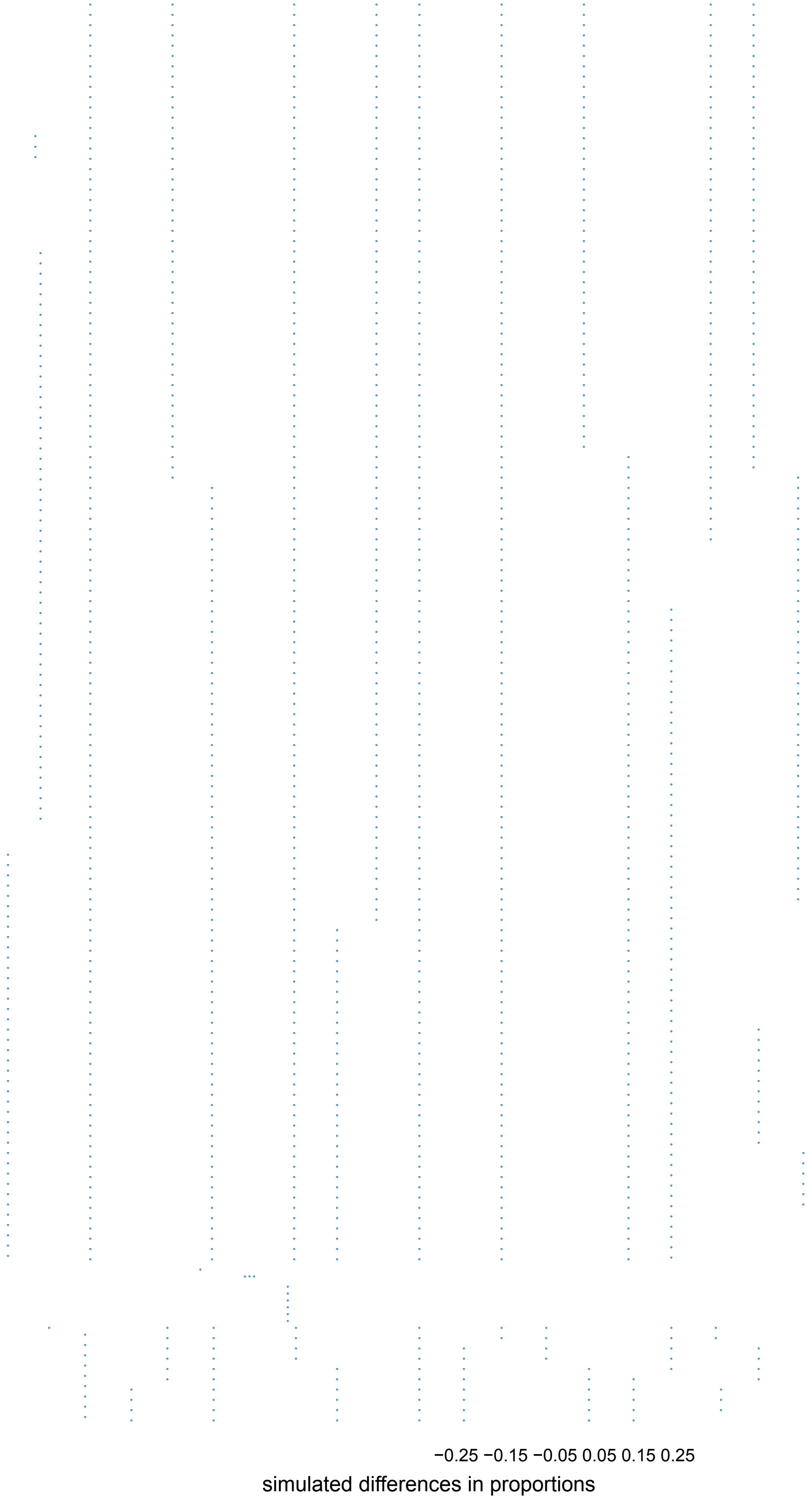
(c) What proportion of patients in the treatment group and what proportion of patients in the control group died?

(d) One approach for investigating whether or not the treatment is effective is to use a randomization technique.

i. What are the claims being tested?

ii. The paragraph below describes the set up for such approach, if we were to do it without using statistical software. Fill in the blanks with a number or phrase, whichever is appropriate. We write *alive* on cards representing patients who were alive at the end of the study, and *dead* on cards representing patients who were not. Then, we shuffle these cards and split them into two groups: one group of size representing treatment, and another group of size representing control. We calculate the difference between the proportion of *dead* cards in the treatment and control groups (treatment - control) and record this value. We repeat this 100 times to build a distribution centered at . Lastly, we calculate the fraction of simulations where the simulated differences in proportions are . If this fraction is low, we conclude that it is unlikely to have observed such an outcome by chance and that the null hypothesis should be rejected in favor of the alternative.

iii. What do the simulation results shown below suggest about the effectiveness of the transplant program?



[30] B. Turnbull et al. "Survivorship of Heart Transplant Data". In: *Journal of the American Statistical Association* 69 (1974), pp. 74–80.

Chapter exercises

2.27 Make-up exam. In a class of 25 students, 24 of them took an exam in class and 1 student took a make-up exam the following day. The professor graded the first batch of 24 exams and found an average score of 74 points with a standard deviation of 8.9 points. The student who took the make-up the following day scored 64 points on the exam.

(a) Does the new student's score increase or decrease the average score?

(b) What is the new average?

(c) Does the new student's score increase or decrease the standard deviation of the scores?

2.28 Infant mortality. The infant mortality rate is defined as the number of infant deaths per 1,000 live births. This rate is often used as an indicator of the level of health in a country. The relative frequency histogram below shows the distribution of estimated infant death rates for 224 countries for which such data were available in 2014.[31]

(a) Estimate Q1, the median, and Q3 from the histogram.

(b) Would you expect the mean of this data set to be smaller or larger than the median? Explain your reasoning.

0.4

0.3

0.2

0.1

0

0 20 40 60 80 100 120 Infant Mortality (per 1000 Live Births)

2.29 TV watchers. Students in an AP Statistics class were asked how many hours of television they watch per week (including online streaming). This sample yielded an average of 4.71 hours, with a standard deviation of 4.18 hours. Is the distribution of number of hours students watch television weekly symmetric? If not, what shape would you expect this distribution to have? Explain your reasoning.

2.30 A new statistic. The statistic $\frac{\bar{x}}{median}$ can be used as a measure of skewness. Suppose we have a distribution where all observations are greater than 0, $x_i > 0$. What is the expected shape of the distribution under the following conditions? Explain your reasoning.

(a) $\frac{\bar{x}}{median} = 1$

(b) $\frac{\bar{x}}{median} < 1$

(c) $\frac{\bar{x}}{median} > 1$

2.31 Oscar winners. The first Oscar awards for best actor and best actress were given out in 1929. The histograms below show the age distribution for all of the best actor and best actress winners from 1929 to 2018. Summary statistics for these distributions are also provided. Compare the distributions of ages of best actor and actress winners.[32]

Best actress

50

40

30

20

10

0

Best actor

50

[31] CIA Factbook, Country Comparisons, 2014.

[32] Oscar winners from 1929 – 2012, data up to 2009 from the Journal of Statistics Education data archive and more current data from wikipedia.org.

2.32 Exam scores. The average on a history exam (scored out of 100 points) was 85, with a standard deviation of 15. Is the distribution of the scores on this exam symmetric? If not, what shape would you expect this distribution to have? Explain your reasoning.

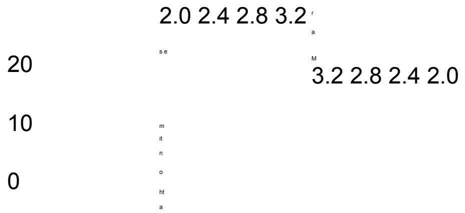
2.33 Stats scores. Below are the final exam scores of twenty introductory statistics students. 57,

66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

Min	Q1	Q2 (Median)	Q3	Max
57	72.5	78.5	82.5	94

2.34 Marathon winners. The histogram and box plots below show the distribution of finishing times for male and female winners of the New York Marathon between 1970 and 1999.



(a) What features of the distribution are apparent in the histogram and not the box plot? What features are apparent in the box plot but not in the histogram?

(b) What may be the reason for the bimodal distribution? Explain.

(c) Compare the distribution of marathon times for men and women based on the box plot shown below.

(d) The time series plot shown below is another way to look at these data. Describe what is visible in this plot but not in the others.

3.2
2.8
2.4 2.0
Women Men
1970 1975 1980 1985 1990 1995 2000

Chapter 3 Probability

Probability forms the foundation of statistics, and you’re probably already aware of many of the ideas presented in this chapter. However, formalization of probability concepts is likely new for most readers.

While this chapter provides a theoretical foundation for the ideas in later chapters and provides a path to a deeper understanding, mastery of the concepts introduced in this chapter is not required for applying the methods introduced in the rest of this book.

For videos, slides, and other resources, please visit
www.openintro.org/os