

Introductory Statistics for the Life and Biomedical Sciences

First Edition

Authors of the derivative work:

Francisco Javier Muñoz Almaraz

Senior Lecturer

CEU-Cardenal Herrera University

Authors of the Original work:

Julie Vu

Preceptor in Statistics

Harvard University

David Harrington

Professor of Biostatistics (Emeritus)

Harvard T.H. Chan School of Public Health

Dana-Farber Cancer Institute

Copyright © 2021. First Edition.
Version date: February 5th, 2021.

This textbook may be downloaded for free at
https://github.com/fjmalmaraz/oi_biostat_text_vet.

The original textbook and extra material may be downloaded for free at
openintro.org/book/biostat.

This textbook is a devivative of *Introductory Statistics for the Life and Biomedical Scienes* by Julie Vu and David Harrington, which is a derivative of *OpenIntro Statistics* 3rd Edition by Diez, Barr, and Çetinkaya-Rundel, and it is available under a Creative Commons Attribution-ShareAlike 3.0 Unported United States license. License details are available at the Creative Commons website: **creativecommons.org**.

Source files for this book may be found on Github at
github.com/OI-Biostat/oi_biostat_text.

Author wants to acknowledge the contribution of his colleages Mónica Alacreu García, Paloma Botella Rocamora y Miguel Ángel Martínez Beneito for their contributions.

Table of Contents

1	Introduction to data	6
1.1	Case study	8
1.2	Data basics	10
1.3	Categorical data	15
1.4	Numerical data	17
1.5	Comparing numerical data across groups	24
1.6	Exercises	26
2	Probability and Distributions of Random Variables	28
2.1	Defining probability	30
2.2	Random variables	43
2.3	Normal distribution	51
2.4	Exercises	59
3	Foundations for inference	62
3.1	Data collection principles	65
3.2	Variability in estimates	77
3.3	Exercises	82
4	Confidence Intervals	83
4.1	Confidence intervals	85
4.2	Single-sample inference with the t -distribution	92
4.3	Confidence interval for a single proportion	97
4.4	Exercises	102
5	One-sample hypothesis testing	104
5.1	Hypothesis testing	106
5.2	Hypothesis testing of the mean	111
5.3	Hypothesis testing of proportions	116
5.4	Understanding hypothesis testing	119
5.5	Notes	123
5.6	Exercises	124
6	Two-sample Hypothesis testing	127
6.1	Inference for the difference of two proportions	129
6.2	Two-sample test for paired data	136
6.3	Testing the equality of two variances	139
6.4	Two-sample test for independent data with identical variances	141

6.5	Two-sample test for independent data with non-identical variances	143
6.6	Notes	149
6.7	Exercises	150
7	Analysis of Variance	152
7.1	Comparing means with ANOVA	154
7.2	Multiple comparisons and controlling Type I Error rate	163
7.3	Exercises	167
8	The chi-square test	170
8.1	Inference for two or more groups	172
8.2	Examples of the chi-square test	178
8.3	Exercises	183
9	Simple linear regression	185
9.1	Summaries of two quantitative variables	188
9.2	Examining scatterplots	193
9.3	Estimating a regression line using least squares	195
9.4	Interpreting a linear model	198
9.5	Statistical inference with regression	207
9.6	Interval estimates with regression	211
9.7	Notes	213
A	Solutions	214
B	Distribution tables	219
	Index	225

Chapter 1

Introduction to data

1.1 Case study

1.2 Data basics

1.3 Categorical data

1.4 Numerical data

1.5 Comparing numerical data across groups

1.6 Exercises

Making observations and recording **data** form the backbone of empirical research, and represent the beginning of a systematic approach to investigating scientific questions. As a discipline, statistics focuses on addressing the following three questions in a rigorous and efficient manner:

- How can data best be collected?
- How should data be analyzed?
- What can be inferred from data?

It is helpful to put statistics in the context of a general process of investigation:

1. Formulation of the research problem.
2. Identification of key variables.
3. Statistical design of an experiment.
4. Collection of data.
5. Statistical analysis of the data.
6. Interpretation of the analytical results. Decision making. Statistics focuses on stage 4-5 and try to answer: How best can we collect data? How should it be analyzed? And what can we infer from the analysis? An Analysis of data which consists on summarizing the data information by means of numbers and graphical representation is called **descriptive statistics**. In other circumstances , we try to determine whether the apparent differences in a quantity are real or may be due to chance. In this case, it is called **inferential statistics**.

Biostatistics is the application of statistical techniques to the life and health science.

This chapter provides a brief discussion on the principles of data collection, and introduces basic methods for summarizing and exploring data.



For labs, slides, and other resources, please visit
www.openintro.org/book/biostat

1.1 Case study: preventing peanut allergies

The proportion of young children in Western countries with peanut allergies has doubled in the last 10 years. Previous research suggests that exposing infants to peanut-based foods, rather than excluding such foods from their diets, may be an effective strategy for preventing the development of peanut allergies. The "Learning Early about Peanut Allergy" (LEAP) study was conducted to investigate whether early exposure to peanut products reduces the probability that a child will develop peanut allergies.¹

The study team enrolled children in the United Kingdom between 2006 and 2009, selecting 640 infants with eczema, egg allergy, or both. Each child was randomly assigned to either the peanut consumption (treatment) group or the peanut avoidance (control) group. Children in the treatment group were fed at least 6 grams of peanut protein daily until 5 years of age, while children in the control group avoided consuming peanut protein until 5 years of age.

At 5 years of age, each child was tested for peanut allergy using an oral food challenge (OFC): 5 grams of peanut protein in a single dose. A child was recorded as passing the oral food challenge if no allergic reaction was detected, and failing the oral food challenge if an allergic reaction occurred. These children had previously been tested for peanut allergy through a skin test, conducted at the time of study entry; the main analysis presented in the paper was based on data from 530 children with an earlier negative skin test.²

Individual-level data from the study are shown in Figure 1.1 for 5 of the 530 children—each row represents a participant and shows the participant's study ID number, treatment group assignment, and OFC outcome.³

participant.ID	treatment.group	overall.V60.outcome
LEAP_100522	Peanut Consumption	PASS OFC
LEAP_103358	Peanut Consumption	PASS OFC
LEAP_105069	Peanut Avoidance	PASS OFC
LEAP_994047	Peanut Avoidance	PASS OFC
LEAP_997608	Peanut Consumption	PASS OFC

Figure 1.1: Individual-level LEAP results, for five children.

The data can be organized in the form of a two-way summary table; Figure 1.2 shows the results categorized by treatment group and OFC outcome.

	FAIL OFC	PASS OFC	Sum
Peanut Avoidance	36	227	263
Peanut Consumption	5	262	267
Sum	41	489	530

Figure 1.2: Summary of LEAP results, organized by treatment group (either peanut avoidance or consumption) and result of the oral food challenge at 5 years of age (either pass or fail).

¹Du Toit, George, et al. Randomized trial of peanut consumption in infants at risk for peanut allergy. *New England Journal of Medicine* 372.9 (2015): 803-813.

²Although a total of 542 children had an earlier negative skin test, data collection did not occur for 12 children.

³The data are available as LEAP in the R package oibioestat.

The summary table makes it easier to identify patterns in the data. Recall that the question of interest is whether children in the peanut consumption group are more or less likely to develop peanut allergies than those in the peanut avoidance group. In the avoidance group, the proportion of children failing the OFC is $36/263 = 0.137$ (13.7%); in the consumption group, the proportion of children failing the OFC is $5/267 = 0.019$ (1.9%). Figure 1.3 shows a graphical method of displaying the study results, using either the number of individuals per category from Figure 1.2 or the proportion of individuals with a specific OFC outcome in a group.

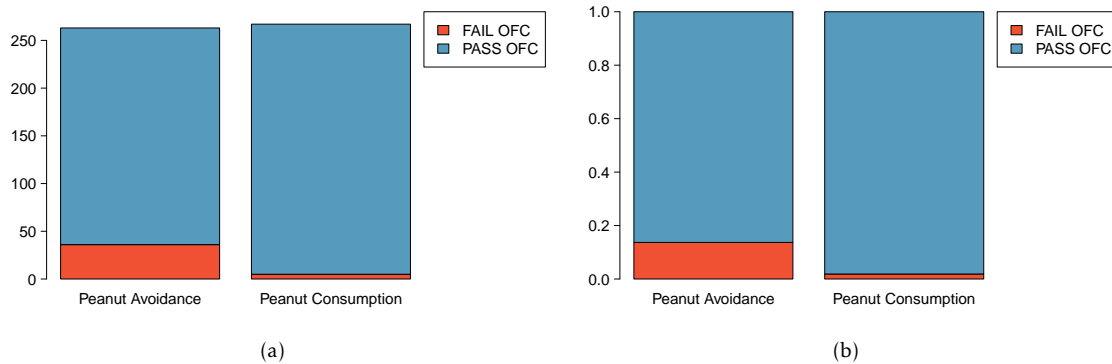


Figure 1.3: (a) A bar plot displaying the number of individuals who failed or passed the OFC in each treatment group. (b) A bar plot displaying the proportions of individuals in each group that failed or passed the OFC.

The proportion of participants failing the OFC is 11.8% higher in the peanut avoidance group than the peanut consumption group. Another way to summarize the data is to compute the ratio of the two proportions ($0.137/0.019 = 7.31$), and conclude that the proportion of participants failing the OFC in the avoidance group is more than 7 times as large as in the consumption group; i.e., the risk of failing the OFC was more than 7 times as great for participants in the avoidance group relative to the consumption group.

Based on the results of the study, it seems that early exposure to peanut products may be an effective strategy for reducing the chances of developing peanut allergies later in life. It is important to note that this study was conducted in the United Kingdom at a single site of pediatric care; it is not clear that these results can be generalized to other countries or cultures.

The results also raise an important statistical issue: does the study provide definitive evidence that peanut consumption is beneficial? In other words, is the 11.8% difference between the two groups larger than one would expect by chance variation alone? The material on inference in later chapters will provide the statistical tools to evaluate this question.

1.2 Data basics

Effective organization and description of data is a first step in most analyses. This section introduces a structure for organizing data and basic terminology used to describe data.

1.2.1 Observations, variables, and data matrices

In evolutionary biology, parental investment refers to the amount of time, energy, or other resources devoted towards raising offspring. This section introduces the frog dataset, which originates from a 2013 study about maternal investment in a frog species.⁴ Reproduction is a costly process for female frogs, necessitating a trade-off between individual egg size and total number of eggs produced. Researchers were interested in investigating how maternal investment varies with altitude and collected measurements on egg clutches found at breeding ponds across 11 study sites; for 5 sites, the body size of individual female frogs was also recorded.

An **Observational units**, which is also sometimes called a **unit of observation** or **case**, is any individual that provide information about the research question. Curiously, in Statistics any object can be an individual, not only humans. For instance, the individuals are egg clutches in the frog dataset.

	altitude	latitude	egg.size	clutch.size	clutch.volume	body.size
1	3,462.00	34.82	1.95	181.97	177.83	3.63
2	3,462.00	34.82	1.95	269.15	257.04	3.63
3	3,462.00	34.82	1.95	158.49	151.36	3.72
150	2,597.00	34.05	2.24	537.03	776.25	NA

Figure 1.4: Data matrix for the frog dataset.

Figure 1.4 displays rows 1, 2, 3, and 150 of the data from the 431 clutches observed as part of the study.⁵ Each row in the table corresponds to a single clutch, indicating where the clutch was collected (altitude and latitude), egg.size, clutch.size, clutch.volume, and body.size of the mother when available. "NA" corresponds to a missing value, indicating that information on an individual female was not collected for that particular clutch. The recorded characteristics are referred to as **variables**; in this table, each column represents a variable.

variable	description
altitude	Altitude of the study site in meters above sea level
latitude	Latitude of the study site measured in degrees
egg.size	Average diameter of an individual egg to the 0.01 mm
clutch.size	Estimated number of eggs in clutch
clutch.volume	Volume of egg clutch in mm ³
body.size	Length of mother frog in cm

Figure 1.5: Variables and their descriptions for the frog dataset.

It is important to check the definitions of variables, as they are not always obvious. For example, why has clutch.size not been recorded as whole numbers? For a given clutch, researchers counted approximately 5 grams' worth of eggs and then estimated the total number of eggs based on the mass of the entire clutch. Definitions of the variables are given in Figure 1.5.⁶

⁴Chen, W., et al. Maternal investment increases with altitude in a frog on the Tibetan Plateau. *Journal of evolutionary biology* 26.12 (2013): 2710-2715.

⁵The frog dataset is available in the R package oibistat.

⁶The data discussed here are in the original scale; in the published paper, some values have undergone a natural log

The data in Figure 1.4 are organized as a **data matrix**. Each row of a data matrix corresponds to an observational unit, and each column corresponds to a variable. A piece of the data matrix for the LEAP study introduced in Section 1.1 is shown in Figure 1.1; the rows are study participants and three variables are shown for each participant. Data matrices are a convenient way to record and store data. If the data are collected for another individual, another row can easily be added; similarly, another column can be added for a new variable.

1.2.2 Types of variables

A COMMON CLASSIFICATION IS:

Quantitative variable The value of the variable comes from measuring something. Hence, their values has to be numeric. For this reason, this type of variable are also called numerical.

Continuous variable Any value can be taken within a range. In general, these variable are measured with some instrument (for instance, a scale) and the value depend on how accurate is the instrument. Almost always, a continuous variable answers a “how much” question.

Discrete variable Not all the values can be taken within the range. Almost always, the values of a discrete variable are interger numbers and they are answering a “how many” question.

Qualitative or categorical These variables classify or identify the observational unit. Usually, they are answering a “what” question. All the possible values are called the variable’s **levels**. They can be divided in two subtypes:

Ordinal variable There exists a logical order for the variables.

Nominal variable A logical order for the values of the variable has not meaning for a stadistical study.

The Functional polymorphisms Associated with human Muscle Size and Strength study (FAMuSS) measured a variety of demographic, phenotypic, and genetic characteristics for about 1,300 participants.⁷ Data from the study have been used in a number of subsequent studies,⁸ such as one examining the relationship between muscle strength and genotype at a location on the ACTN3 gene.⁹

The famuss dataset is a subset of the data for 595 participants.¹⁰ Four rows of the famuss dataset are shown in Figure 1.6, and the variables are described in Figure 1.7.

The variables age, height, weight, and ndrm.ch are **numerical variables**. They take on numerical values, and it is reasonable to add, subtract, or take averages with these values. In contrast, a variable reporting telephone numbers would not be classified as numerical, since sums, differ-

transformation.

⁷Thompson PD, Moyna M, Seip, R, et al., 2004. Functional Polymorphisms Associated with Human Muscle Size and Strength. *Medicine and Science in Sports and Exercise* 36:1132 - 1139.

⁸Pescatello L, et al. Highlights from the functional single nucleotide polymorphisms associated with human muscle size and strength or FAMuSS study, *BioMed Research International* 2013.

⁹Clarkson P, et al., *Journal of Applied Physiology* 99: 154-163, 2005.

¹⁰The subset is from Foulkes, Andrea S. *Applied statistical genetics with R: for population-based association studies*. Springer Science & Business Media, 2009. The full version of the data is available at <http://people.umass.edu/foulkes/asg/data.html>.

	sex	age	race	height	weight	actn3.r577x	ndrm.ch
1	Female	27	Caucasian	65.0	199.0	CC	40.0
2	Male	36	Caucasian	71.7	189.0	CT	25.0
3	Female	24	Caucasian	65.0	134.0	CT	40.0
595	Female	30	Caucasian	64.0	134.0	CC	43.8

Figure 1.6: Four rows from the famuss data matrix.

variable	description
sex	Sex of the participant
age	Age in years
race	Race, recorded as African Am (African American), Caucasian, Asian, Hispanic or Other
height	Height in inches
weight	Weight in pounds
actn3.r577x	Genotype at the location r577x in the ACTN3 gene.
ndrm.ch	Percent change in strength in the non-dominant arm, comparing strength after to before training

Figure 1.7: Variables and their descriptions for the famuss dataset.

ences, and averages in this context have no meaning. Age measured in years is said to be **discrete**, since it can only take on numerical values with jumps; i.e., positive integer values. Percent change in strength in the non-dominant arm (ndrm.ch) is **continuous**, and can take on any value within a specified range.

The variables sex, race, and actn3.r577x are **categorical variables**, which take on values that are names or labels. The possible values of a categorical variable are called the variable's **levels**.¹¹ For example, the levels of actn3.r577x are the three possible genotypes at this particular locus: CC, CT, or TT. Categorical variables without a natural ordering are called **nominal categorical variables**; sex, race, and actn3.r577x are all nominal categorical variables. Categorical variables with levels that have a natural ordering are referred to as **ordinal categorical variables**. For example, age of the participants grouped into 5-year intervals (15-20, 21-25, 26-30, etc.) is an ordinal categorical variable.

¹¹Categorical variables are sometimes called **factor variables**.

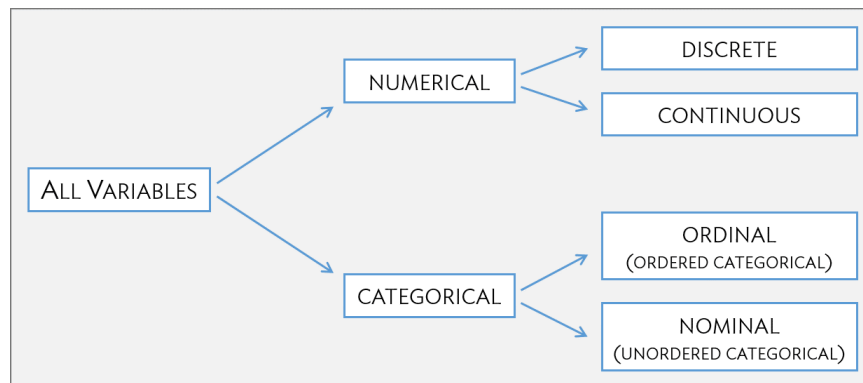


Figure 1.8: Breakdown of variables into their respective types.

EXAMPLE 1.1

Classify the variables in the frog dataset: `altitude`, `latitude`, `egg.size`, `clutch.size`, `clutch.volume`, and `body.size`.

E

The variables `egg.size`, `clutch.size`, `clutch.volume`, and `body.size` are continuous numerical variables, and can take on all positive values.

In the context of this study, the variables `altitude` and `latitude` are best described as categorical variables, since the numerical values of the variables correspond to the 11 specific study sites where data were collected. Researchers were interested in exploring the relationship between altitude and maternal investment; it would be reasonable to consider `altitude` an ordinal categorical variable.

GUIDED PRACTICE 1.2**G**

Characterize the variables `treatment.group` and `overall.V60.outcome` from the LEAP study (discussed in Section 1.1).¹²

GUIDED PRACTICE 1.3**G**

Suppose that on a given day, a research assistant collected data on the first 20 individuals visiting a walk-in clinic: age (measured as less than 21, 21 - 65, and greater than 65 years of age), sex, height, weight, and reason for the visit. Classify each of the variables.¹³

¹²These variables measure non-numerical quantities, and thus are categorical variables with two levels.

¹³Height and weight are continuous numerical variables. Age as measured by the research assistant is ordinal categorical. Sex and the reason for the visit are nominal categorical variables.

1.2.3 Relationships between variables

Many studies are motivated by a researcher examining how two or more variables are related. For example, do the values of one variable increase as the values of another decrease? Do the values of one variable tend to differ by the levels of another variable?

One study used the famuss data to investigate whether ACTN3 genotype at a particular location (residue 577) is associated with change in muscle strength. The ACTN3 gene codes for a protein involved in muscle function. A common mutation in the gene at a specific location changes the cytosine (C) nucleotide to a thymine (T) nucleotide; individuals with the TT genotype are unable to produce any ACTN3 protein.

Researchers hypothesized that genotype at this location might influence muscle function. As a measure of muscle function, they recorded the percent change in non-dominant arm strength after strength training; this variable, `ndrm.ch`, is the **response variable** in the study. A response variable is defined by the particular research question a study seeks to address, and measures the outcome of interest in the study. A study will typically examine whether the values of a response variable differ as values of an **explanatory variable** change, and if so, how the two variables are related. A given study may examine several explanatory variables for a single response variable.¹⁴ The explanatory variable examined in relation to `ndrm.ch` in the study is `actn3.r557x`, ACTN3 genotype at location 577.

EXAMPLE 1.4

In the maternal investment study conducted on frogs, researchers collected measurements on egg clutches and female frogs at 11 study sites, located at differing altitudes, in order to investigate how maternal investment varies with altitude. Identify the response and explanatory variables in the study.

E The variables `egg.size`, `clutch.size`, and `clutch.volume` are response variables indicative of maternal investment.

The explanatory variable examined in the study is altitude.

While latitude is an environmental factor that might potentially influence features of the egg clutches, it is not a variable of interest in this particular study.

Female body size (`body.size`) is neither an explanatory nor response variable.

GUIDED PRACTICE 1.5

G Refer to the variables from the famuss dataset described in Figure 1.7 to formulate a question about the relationships between these variables, and identify the response and explanatory variables in the context of the question.¹⁵

¹⁴Response variables are sometimes called dependent variables and explanatory variables are often called independent variables or predictors.

¹⁵Two sample questions: (1) Does change in participant arm strength after training seem associated with race? The response variable is `ndrm.ch` and the explanatory variable is `race`. (2) Do male participants appear to respond differently to strength training than females? The response variable is `ndrm.ch` and the explanatory variable is `sex`.

1.3 Categorical data

This section introduces tables and plots for summarizing categorical data, using the famuss dataset introduced in Section 1.2.2.

1.3.1 Frequency

We call (absolute) **frequency** the number of occurrences in the sample of a given value for a variable. We call **relative frequency** the proportion of occurrences in the sample of a given value for a variable with respect to the overall number of observational units.

A table for a single variable is called a **frequency table**. Figure 1.9 is a frequency table for the `actn3.r577x` variable, showing the distribution of genotype at location r577x on the ACTN3 gene for the FAMuSS study participants.

In a **relative frequency table** like Figure 1.10, the proportions per each category are shown instead of the counts.

	CC	CT	TT	Sum
Counts	173	261	161	595

Figure 1.9: A frequency table for the `actn3.r577x` variable.

	CC	CT	TT	Sum
Proportions	0.291	0.439	0.271	1.000

Figure 1.10: A relative frequency table for the `actn3.r577x` variable.

The **mode** of a categorical variable is the value (or category) with the highest frequency. The category CT is the mode of the `actn3.r577x` variable for FAMuSS study.

1.3.2 Bar plots

A bar plot is a common way to display a single categorical variable. The left panel of Figure 1.11 shows a **bar plot** of the counts per genotype for the `actn3.r577x` variable. The plot in the right panel shows the proportion of observations that are in each level (i.e. in each genotype).

1.3.3 Pie charts

While pie charts are well known, they are not typically as useful as other charts in a data analysis. A **pie chart** is shown in Figure 1.12 on the following page alongside a bar plot. It is generally more difficult to compare group sizes in a pie chart than in a bar plot, especially when categories have nearly identical counts or proportions. In the case of the CC and bTT categories, the difference is so slight you may be unable to distinguish any difference in group sizes for either plot!

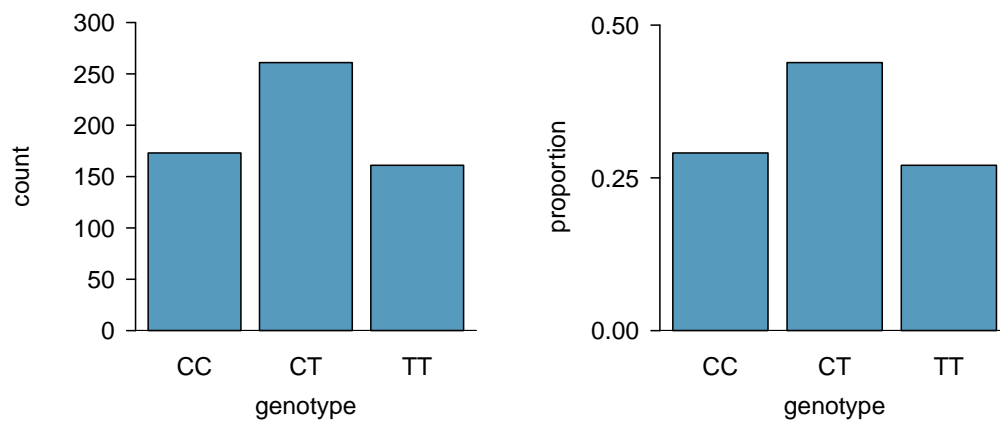


Figure 1.11: Two bar plots of `actn3.r577x`. The left panel shows the counts, and the right panel shows the proportions for each genotype.

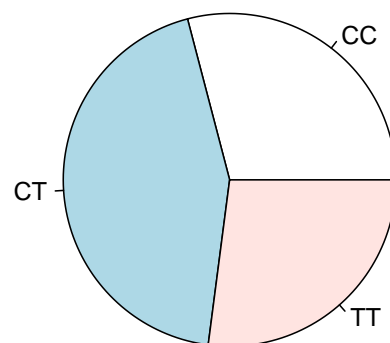


Figure 1.12: A pie chart of the variable `actn3.r577x`.

1.4 Numerical data

This section discusses techniques for exploring and summarizing numerical variables, using the frog data from the parental investment study introduced in Section 1.2.

1.4.1 Measures of center: mean and median

The **mean**, sometimes called the average, is a measure of center for a **distribution** of data. To find the average clutch volume for the observed egg clutches, add all the clutch volumes and divide by the total number of clutches.¹⁶

$$\bar{x} = \frac{177.8 + 257.0 + \cdots + 933.3}{431} = 882.5 \text{ mm}^3.$$

The sample mean is often labeled \bar{x} , to distinguish it from μ , the mean of the entire population from which the sample is drawn. The letter x is being used as a generic placeholder for the variable of interest, `clutch.volume`.

\bar{x}
sample
mean
 μ
population
mean

MEAN

The sample mean of a numerical variable is the sum of the values of all observations divided by the number of observations:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}, \quad (1.6)$$

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

The **median** is another measure of center; it is the middle number in a distribution after the values have been ordered from smallest to largest. If the distribution contains an even number of observations, the median is the average of the middle two observations. There are 431 clutches in the dataset, so the median is the clutch volume of the 216th observation in the sorted values of `clutch.volume`: 831.8 mm³.

1.4.2 Measure of location: percentile

A 80th **percentile** is a value such that a 80 percent of the values are lower than this value. Consequently, the median is the 50th percentile. There are several ways to calculate percentile, we are going to use the procedure chosen by R.

¹⁶For computational convenience, the volumes are rounded to the first decimal.

HOW TO CALCULATE A PERCENTILE

Given $(0 \leq p \leq 1)$ and a quantitative variable which data are x_1, x_2, \dots, x_n .

The ordered sequence (depending on the value x_i) is $x_{(1)}, x_{(2)}, \dots, x_{(n)}$.

The p -quantile or $100p^{\text{th}}$ percentile is calculated by the formula

$$(1 - \gamma) \cdot x_{(k)} + \gamma \cdot x_{(k+1)}$$

where:

1. k is the greatest integer less or equal than $(n - 1) \cdot p + 1$.
2. γ is the decimal part of $(n - 1) \cdot p + 1$, i.e. $\gamma = (n - 1) \cdot p + 1 - k$

EXAMPLE 1.7

Calculate the 80^{th} percentile of a variable takes the following values for a sample

Its sample size is $n = 9$. The sequence is ordered

$$\begin{array}{ccccc} x_{(1)} = -0.96 & x_{(2)} = -0.94 & x_{(3)} = -0.85 & x_{(4)} = -0.63 & x_{(5)} = -0.61 \\ x_{(6)} = 0.47 & x_{(7)} = 0.95 & x_{(8)} = 1.2 & x_{(9)} = 1.22 & \end{array}$$

The value of $(n - 1) \cdot p + 1 = 8 \cdot 0.8 + 1 = 7.4$. Therefore the integer part is $k = 7$ and the decimal part is $\gamma = 0.4$. The 0.8 -quantile is

$$(1 - \gamma) \cdot x_{(k)} + \gamma \cdot x_{(k+1)} = 0.6 \cdot 0.95 + 0.4 \cdot 1.2 = 1.05$$

1.4.3 Measures of spread: standard deviation and interquartile range

The spread of a distribution refers to how similar or varied the values in the distribution are to each other; i.e., whether the values are tightly clustered or spread over a wide range.

The standard deviation for a set of data describes the typical distance between an observation and the mean. The distance of a single observation from the mean is its **deviation**. Below are the deviations for the 1^{st} , 2^{nd} , 3^{rd} , and 431^{st} observations in the clutch.volume variable.

$$\begin{aligned} x_1 - \bar{x} &= 177.8 - 882.5 = -704.7 \\ x_2 - \bar{x} &= 257.0 - 882.5 = -625.5 \\ x_3 - \bar{x} &= 151.4 - 882.5 = -731.1 \\ &\vdots \\ x_{431} - \bar{x} &= 933.2 - 882.5 = 50.7 \end{aligned}$$

The sample **variance**, the average of the squares of these deviations, is denoted by s^2 :

s^2
sample
variance

$$\begin{aligned}
s^2 &= \frac{(-704.7)^2 + (-625.5)^2 + (-731.1)^2 + \cdots + (50.7)^2}{431 - 1} \\
&= \frac{496,602.09 + 391,250.25 + 534,507.21 + \cdots + 2570.49}{430} \\
&= 143,680.9.
\end{aligned}$$

The denominator is $n - 1$ rather than n ; this mathematical nuance accounts for the fact that sample mean has been used to estimate the population mean in the calculation. Details on the statistical theory can be found in more advanced texts.

The sample **standard deviation** s is the square root of the variance:

$$s = \sqrt{143,680.9} = 379.05 \text{ mm}^3.$$

Like the mean, the population values for variance and standard deviation are denoted by Greek letters: σ^2 for the variance and σ for the standard deviation.

s
sample
standard
deviation
 σ^2
population
variance
 σ
population
standard
deviation

STANDARD DEVIATION

The sample standard deviation of a numerical variable is computed as the square root of the variance, which is the sum of squared deviations divided by the number of observations minus 1.

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}}, \quad (1.8)$$

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

Variability can also be measured using the **interquartile range** (IQR). The IQR for a distribution is the difference between the first and third quartiles: $Q_3 - Q_1$. The first quartile (Q_1) is equivalent to the 25th percentile; i.e., 25% of the data fall below this value. The third quartile (Q_3) is equivalent to the 75th percentile. By definition, the median represents the second quartile, with half the values falling below it and half falling above. The IQR for clutch.volume is $1096.0 - 609.6 = 486.4 \text{ mm}^3$.

Measures of center and spread are ways to summarize a distribution numerically. Using numerical summaries allows for a distribution to be efficiently described with only a few numbers.¹⁷ For example, the calculations for clutch.volume indicate that the typical egg clutch has volume of about 880 mm^3 , while the middle 50% of egg clutches have volumes between approximately 600 mm^3 and 1100.0 mm^3 .

1.4.4 Robust estimates

Figure 1.13 shows the values of clutch.volume as points on a single axis. There are a few values that seem extreme relative to the other observations: the four largest values, which appear distinct from the rest of the distribution. How do these extreme values affect the value of the numerical summaries?

Figure 1.14 shows the summary statistics calculated under two scenarios, one with and one without the four largest observations. For these data, the median does not change, while the IQR

¹⁷Numerical summaries are also known as summary statistics.

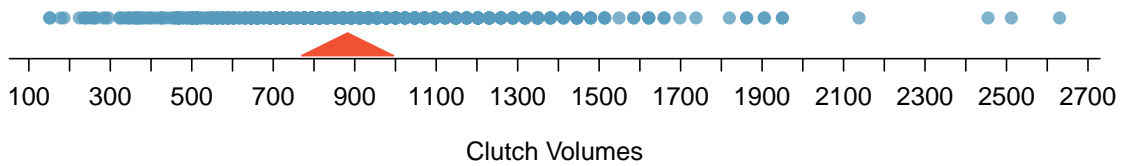


Figure 1.13: **Dot plot** of clutch volumes from the frog data.

differs by only about 6 mm^3 . In contrast, the mean and standard deviation are much more affected, particularly the standard deviation.

scenario	robust		not robust	
	median	IQR	\bar{x}	s
original data (with extreme observations)	831.8	486.9	882.5	379.1
data without four largest observations	831.8	493.9	867.9	349.2

Figure 1.14: A comparison of how the median, IQR, mean (\bar{x}), and standard deviation (s) change when extreme observations are present.

The median and IQR are referred to as **robust estimates** because extreme observations have little effect on their values. For distributions that contain extreme values, the median and IQR will provide a more accurate sense of the center and spread than the mean and standard deviation.

1.4.5 Visualizing distributions of data: histograms and boxplots

Graphs show important features of a distribution that are not evident from numerical summaries, such as asymmetry or extreme values. While dot plots show the exact value of each observation, histograms and boxplots graphically summarize distributions.

In a **histogram**, observations are grouped into bins and plotted as bars. Figure 1.15 shows the number of clutches with volume between 0 and 200 mm^3 , 200 and 400 mm^3 , etc. up until 2,600 and $2,800 \text{ mm}^3$.¹⁸ These binned counts are plotted in Figure 1.16.

Clutch volumes	0-200	200-400	400-600	600-800	...	2400-2600	2600-2800
Count	4	29	69	99	...	2	1

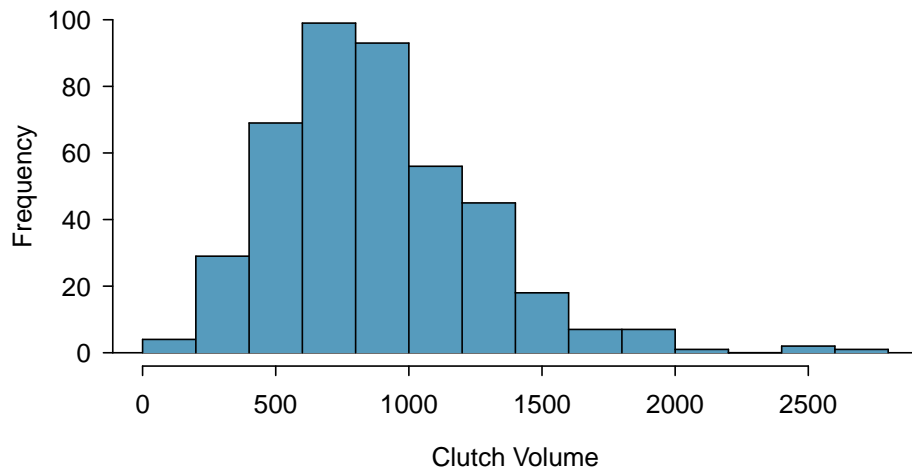
Figure 1.15: The counts for the binned clutch volume data.

Histograms provide a view of the **data density**. Higher bars indicate more frequent observations, while lower bars represent relatively rare observations. Figure 1.16 shows that most of the egg clutches have volumes between $500\text{--}1,000 \text{ mm}^3$, and there are many more clutches with volumes smaller than $1,000 \text{ mm}^3$ than clutches with larger volumes.

Histograms show the **shape** of a distribution. The tails of a **symmetric** distribution are roughly equal, with data trailing off from the center roughly equally in both directions. Asymmetry arises when one tail of the distribution is longer than the other. A distribution is said to be **right skewed** when data trail off to the right, and **left skewed** when data trail off to the left.¹⁹ Figure 1.16 shows that the distribution of clutch volume is right skewed; most clutches have relatively small volumes, and only a few clutches have high volumes.

¹⁸By default in R, the bins are left-open and right-closed; i.e., the intervals are of the form $(a, b]$. Thus, an observation with value 200 would fall into the 0-200 bin instead of the 200-400 bin.

¹⁹Other ways to describe data that are skewed to the right/left: **skewed to the right/left** or **skewed to the positive/negative end**.

Figure 1.16: A histogram of `clutch.volume`.

A **mode** is represented by a prominent peak in the distribution.²⁰ Figure 1.17 shows histograms that have one, two, or three major peaks. Such distributions are called **unimodal**, **bimodal**, and **multimodal**, respectively. Any distribution with more than two prominent peaks is called multimodal. Note that the less prominent peak in the unimodal distribution was not counted since it only differs from its neighboring bins by a few observations. Prominent is a subjective term, but it is usually clear in a histogram where the major peaks are.

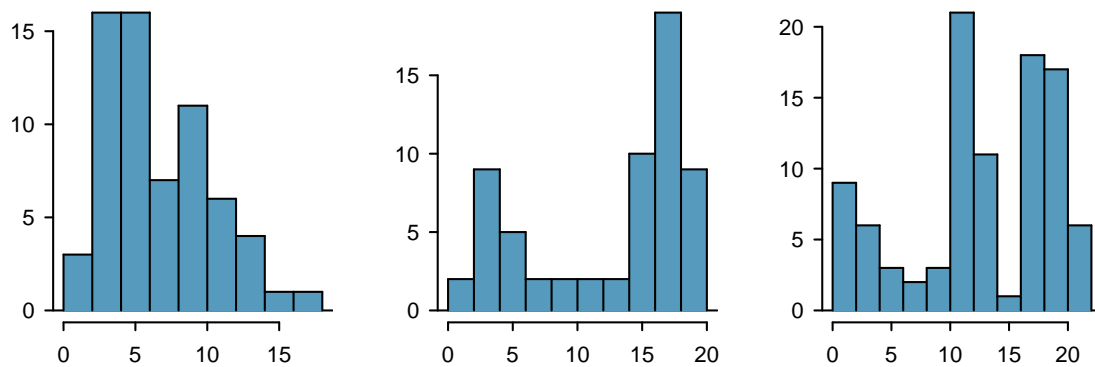


Figure 1.17: From left to right: unimodal, bimodal, and multimodal distributions.

A **boxplot** indicates the positions of the first, second, and third quartiles of a distribution in addition to extreme observations.²¹ Figure 1.18 shows a boxplot of `clutch.volume` alongside a vertical dot plot.

²⁰Another definition of mode, which is not typically used in statistics, is the value with the most occurrences. It is common that a dataset contains *no* observations with the same value, which makes this other definition impractical for many datasets.

²¹Boxplots are also known as box-and-whisker plots.

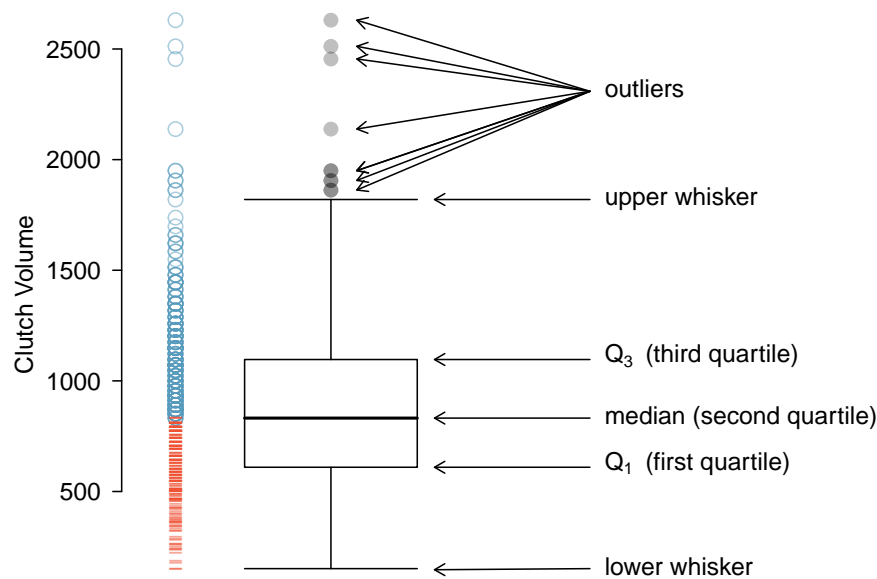


Figure 1.18: A boxplot and dot plot of `clutch.volume`. The horizontal dashes indicate the bottom 50% of the data and the open circles represent the top 50%.

In a boxplot, the interquartile range is represented by a rectangle extending from the first quartile to the third quartile, and the rectangle is split by the median (second quartile). Extending outwards from the box, the **whiskers** capture the data that fall between $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$. The whiskers must end at data points; the values given by adding or subtracting $1.5 \times IQR$ define the maximum reach of the whiskers. For example, with the `clutch.volume` variable, $Q_3 + 1.5 \times IQR = 1,096.5 + 1.5 \times 486.4 = 1,826.1 \text{ mm}^3$. However, there was no clutch with volume $1,826.1 \text{ mm}^3$; thus, the upper whisker extends to $1,819.7 \text{ mm}^3$, the largest observation that is smaller than $Q_3 + 1.5 \times IQR$.

Any observation that lies beyond the whiskers is shown with a dot; these observations are called outliers. An **outlier** is a value that appears extreme relative to the rest of the data. For the `clutch.volume` variable, there are several large outliers and no small outliers, indicating the presence of some unusually large egg clutches.

The high outliers in Figure 1.18 reflect the right-skewed nature of the data. The right skew is also observable from the position of the median relative to the first and third quartiles; the median is slightly closer to the first quartile. In a symmetric distribution, the median will be halfway between the first and third quartiles.

GUIDED PRACTICE 1.9



Use the histogram and boxplot in Figure 1.19 to describe the distribution of height in the `famuss` data, where height is measured in inches.²²

²²The data are roughly symmetric (the left tail is slightly longer than the right tail), and the distribution is unimodal with one prominent peak at about 67 inches. The middle 50% of individuals are between 5.5 feet and just under 6 feet tall. There is one low outlier and one high outlier, representing individuals that are unusually short/tall relative to the other individuals.

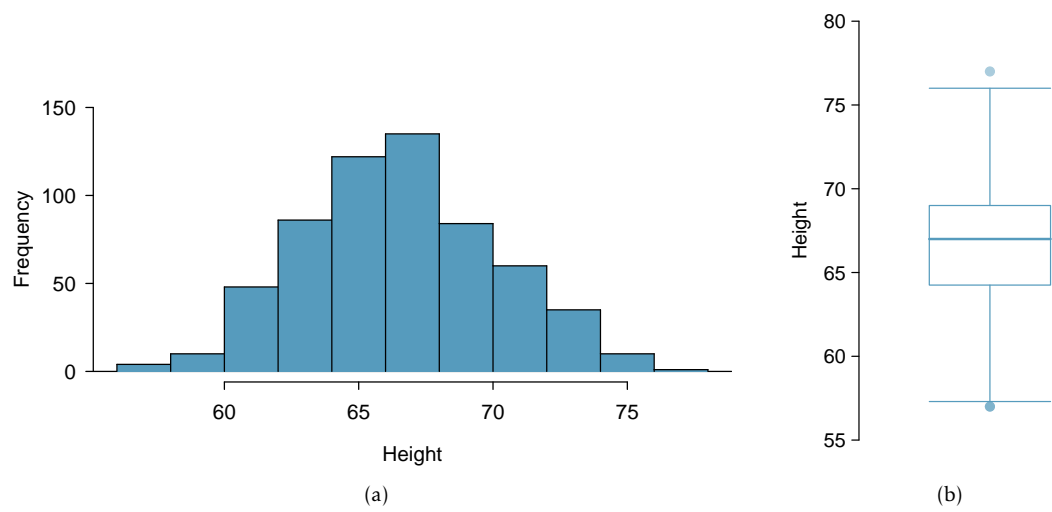


Figure 1.19: A histogram and boxplot of height in the famuss data.

1.5 Comparing numerical data across groups

Methods for comparing numerical data across groups are based on the approaches introduced in Section 1.4. **Side-by-side boxplots** and **hollow histograms** are useful for directly comparing how the distribution of a numerical variable differs by category.

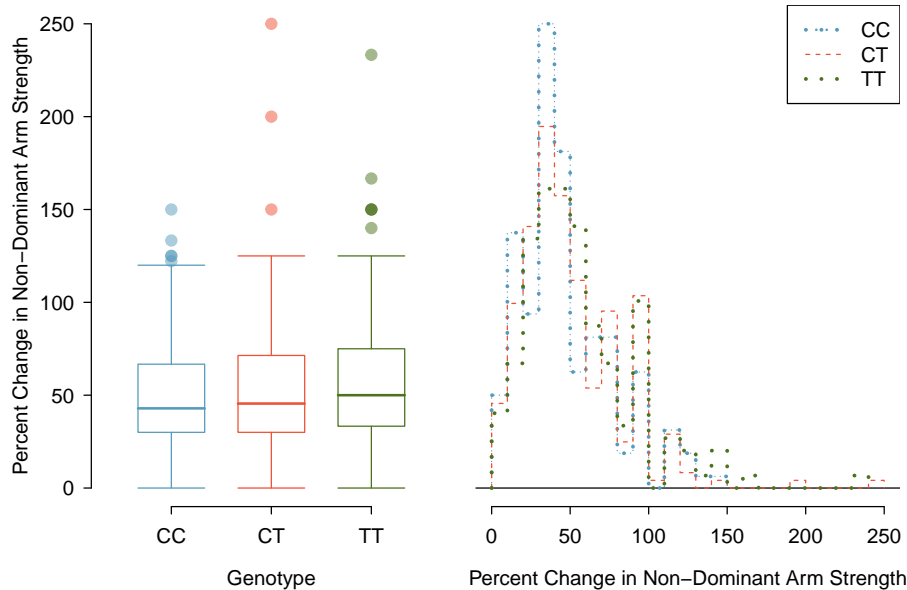


Figure 1.20: Side-by-side boxplot and hollow histograms for `ndrm.ch`, split by levels of `actn3.r577x`.

Recall the question introduced in Section 1.2.3: is ACTN3 genotype associated with variation in muscle function? Figure 1.20 visually shows the relationship between muscle function (measured as percent change in non-dominant arm strength) and ACTN3 genotype in the `famuss` data with side-by-side boxplots and hollow histograms. The hollow histograms highlight how the shapes of the distributions of `ndrm.ch` for each genotype are essentially similar, although the distribution for the CC genotype has less right skewing. The side-by-side boxplots are especially useful for comparing center and spread, and reveal that the T allele appears to be associated with greater muscle function; median percent change in non-dominant arm strength increases across the levels from CC to TT.



GUIDED PRACTICE 1.10

Using Figure 1.21, assess how maternal investment varies with altitude.²³

²³As a general rule, clutches found at higher altitudes have greater volume; median clutch volume tends to increase as altitude increases. This suggests that increased altitude is associated with a higher level of maternal investment.

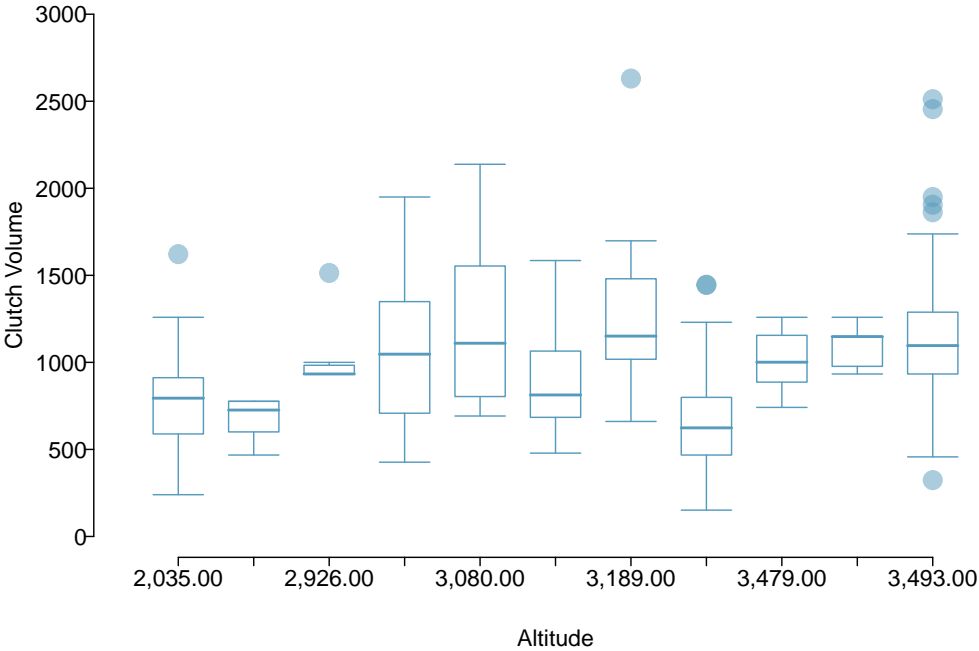


Figure 1.21: Side-by-side boxplot comparing the distribution of clutch.volume for different altitudes.

1.6 Exercises

Exercise. 1.1

Classify the following variables depending on their types (nominal, ordinal, discrete, continuous)

- | | |
|---|---|
| a. Size of a T-Shirt (XS,S,M,L,XL,XXL); | k. The amount of money in a bank account; |
| b. Shoes size; | l. The height of a building; |
| c. Credit card number; | m. Animal weight; |
| d. Patient temperature; | n. Horse breed; |
| e. Number of children in a household; | o. Body temperature; |
| f. Last seen film; | p. Day of the week when an analysis is performed; |
| g. Phone number; | q. Number of puppies born in a delivery; |
| h. Number of passed subjects of the first semester; | r. Brand of animal feed; |
| i. Cell phone company; | s. Whether or not a dog had a previous bone fracture. |
| j. Passport number; | |

Exercise. 1.2

The number of animals in a veterinary clinic during 32 days are:

53 54 42 43 41 43 49 48 37 57 63 44 52 50 43 45
59 40 41 54 63 57 46 51 68 46 42 42 44 48 52 53

- Calculate 10–percentile, 35–percentile and 62–percentile.
- Calculate all the quartiles.
- If there are outlying values, locate it.
- Draw a box plot.

Exercise. 1.3

In a questionnaire about overweighted cats, the owner is asked for the number of times that the pet weight has been measured for the last 6 months. The answers were:

3 5 2 0 2 1 6 2 0 6 2 0 4 3 3 5 2 0 0 1
5 3 6 6 4 6 0 3 1 1 0 5 6 4 4 6 2 3 3 6

- Classify the type of the variable.
- Summarize the previous information with a frequency table.
- Find the mode of the number of controls.

Exercise. 1.4

A veterinary physician is interested in knowing when emergencies occur, so he has collected data from the last week. He has classified them depending on the time in four groups: Morning, Afternoon, Night and Nonworking days (Bank holidays and Sundays).

Morning	Morning	Night	Nonworking	Night	Afternoon	Night
Morning	Morning	Nonworking	Night	Afternoon	Afternoon	Morning
Morning	Morning	Afternoon	Morning	Night	Afternoon	Afternoon
Morning	Afternoon	Nonworking	Morning	Night	Nonworking	Morning
Afternoon	Nonworking	Afternoon	Night			

- Classify the type of the variable.
- Can you calculate the median of the variable?
- Calculate absolute and relative frequencies. Summarize this information with a frequency table.
- Sketch a bar plot.

Exercise. 1.5

A research group has done a study of the copper levels in urine with a sample of 40 dogs between 1 and 5 years old and the following values were measured:

0.10	0.30	0.34	0.36	0.42	0.42	0.45	0.48	0.50	0.52
0.55	0.58	0.62	0.63	0.64	0.65	0.65	0.66	0.69	0.70
0.72	0.73	0.74	0.74	0.75	0.76	0.77	0.78	0.81	0.83
0.85	0.86	0.88	0.90	0.94	0.98	1.04	1.12	1.16	1.24

- What is the variable of study? Classify this variable.
- Calculate the median and the range.
- Calculate the lower and upper quartile.
- Calculate the 10-percentile and the 95-percentile.
- Is there any outlying value?
- Sketch a histogram and a box plot.

Exercise. 1.6

In the records of a zoo, the weight (in grams) of 16 gorillas one month after they were born are shown in the table

4123	4336	4160	4165	4422	3853	3281	3990
4096	4166	3596	4127	4017	3769	4240	4194

- Classify the variable.
- Calculate the following statistics: minimum and maximum, 10 and 90 percentiles, quartiles, median, mean, mode, range, variance, standard deviation.

Chapter 2

Probability and Distributions of Random Variables

2.1 Defining probability

2.2 Random variables

2.3 Normal distribution

2.4 Exercises

What are the chances that a woman with an abnormal mammogram has breast cancer? What is the probability that a woman with an abnormal mammogram has breast cancer, given that she is in her 40's? What is the likelihood that out of 100 women who undergo a mammogram and test positive for breast cancer, at least one of the women has received a false positive result?

These questions use the language of probability to express statements about outcomes that may or may not occur. More specifically, probability is used to quantify the level of uncertainty about each outcome. Like all mathematical tools, probability becomes easier to understand and work with once important concepts and terminology have been formalized.

We use probability to build tools to describe and understand apparent randomness and uncertainty. We often frame probability in terms of **random process** given rise to an **outcome** also known as an **event**.

This chapter introduces that formalization, using two types of examples. One set of examples uses settings familiar to most people – rolling dice or picking cards from a deck. The other set of examples draws from medicine, biology, and public health, reflecting the contexts and language specific to those fields. The approaches to solving these two types of problems are surprisingly similar, and in both cases, seemingly difficult problems can be solved in a series of reliable steps.



For labs, slides, and other resources, please visit
www.openintro.org/book/biostat

2.1 Defining probability

2.1.1 Some examples

The rules of probability can easily be modeled with classic scenarios, such as flipping coins or rolling dice. When a coin is flipped, there are only two possible outcomes, heads or tails. With a fair coin, each outcome is equally likely; thus, the chance of flipping heads is $1/2$, and likewise for tails. The following examples deal with rolling a die or multiple dice; a die is a cube with six faces numbered 1, 2, 3, 4, 5, and 6.

EXAMPLE 2.1

What is the chance of getting 1 when rolling a die?

E

If the die is fair, then there must be an equal chance of rolling a 1 as any other possible number. Since there are six outcomes, the chance must be 1-in-6 or, equivalently, $1/6$.

EXAMPLE 2.2

What is the chance of not rolling a 2?

E

Not rolling a 2 is the same as getting a 1, 3, 4, 5, or 6, which makes up five of the six equally likely outcomes and has probability $5/6$.

EXAMPLE 2.3

Consider rolling two fair dice. What is the chance of getting two 1s?

E

If $1/6^{th}$ of the time the first die is a 1 and $1/6^{th}$ of *those* times the second die is also a 1, then the chance that both dice are 1 is $(1/6)(1/6)$ or $1/36$.

Probability can also be used to model less artificial contexts, such as to predict the inheritance of genetic disease. Cystic fibrosis (CF) is a life-threatening genetic disorder caused by mutations in the *CFTR* gene located on chromosome 7. Defective copies of *CFTR* can result in the reduced quantity and function of the CFTR protein, which leads to the buildup of thick mucus in the lungs and pancreas.¹ CF is an autosomal recessive disorder; an individual only develops CF if they have inherited two affected copies of *CFTR*. Individuals with one normal (wild-type) copy and one defective (mutated) copy are known as carriers; they do not develop CF, but may pass the disease-causing mutation onto their offspring.

¹The CFTR protein is responsible for transporting sodium and chloride ions across cell membranes.

EXAMPLE 2.4

Suppose that both members of a couple are CF carriers. What is the probability that a child of this couple will be affected by CF? Assume that a parent has an equal chance of passing either gene copy (i.e., allele) to a child.

*Solution 1: Enumerate all of the possible outcomes and exploit the fact that the outcomes are equally likely, as in Example 2.1. Figure 2.1 shows the four possible genotypes for a child of these parents. The paternal chromosome is in blue and the maternal chromosome in green, while chromosomes with the wild-type and mutated versions of *CFTR* are marked with + and –, respectively. The child is only affected if they have genotype (–/–), with two mutated copies of *CFTR*. Each of the four outcomes occurs with equal likelihood, so the child will be affected with probability 1-in-4, or 1/4. It is important to recognize that the child being an unaffected carrier (+/–) consists of two distinct outcomes, not one.*

Solution 2: Calculate the proportion of outcomes that produce an affected child, as in Example 2.3. During reproduction, one parent will pass along an affected copy half of the time. When the child receives an affected allele from one parent, half of the those times, they will also receive an affected allele from the other parent. Thus, the proportion of times the child will have two affected copies is $(1/2) \times (1/2) = 1/4$.

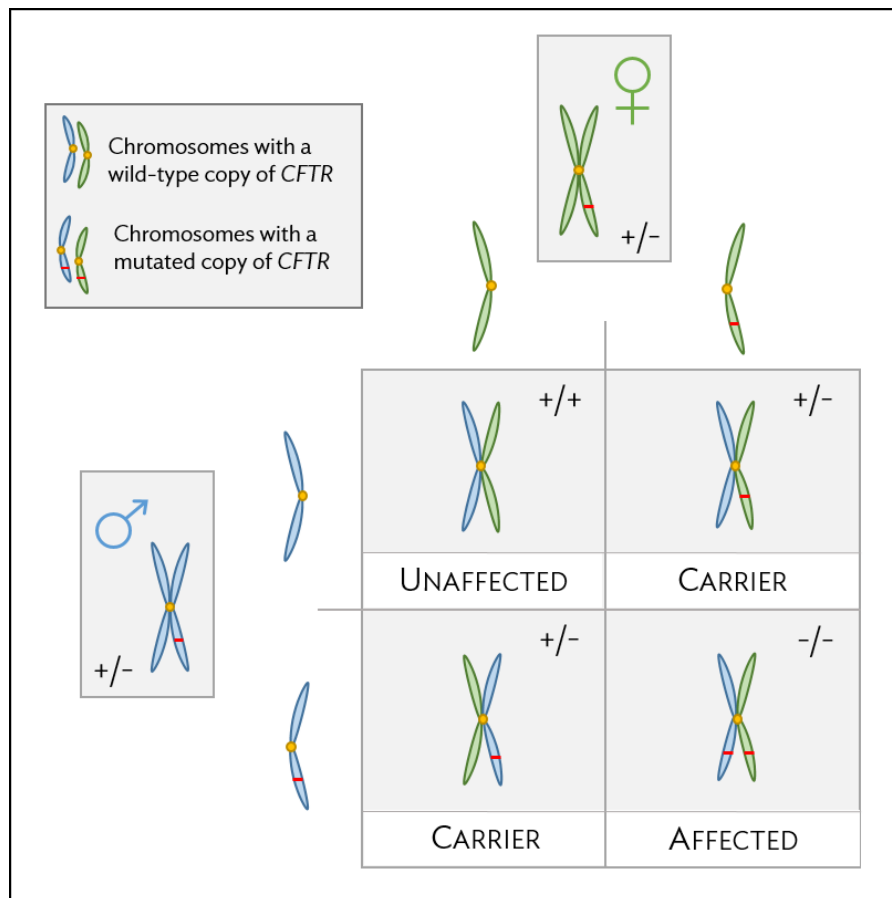


Figure 2.1: Pattern of CF inheritance for a child of two unaffected carriers

GUIDED PRACTICE 2.5



Suppose the father has CF and the mother is an unaffected carrier. What is the probability that their child will be affected by the disease?²

2.1.2 Probability

Probability is used to assign a level of uncertainty to the outcomes of phenomena that either happen randomly (e.g. rolling dice, inheriting of disease alleles), or appear random because of a lack of understanding about exactly how the phenomenon occurs (e.g. a woman in her 40's developing breast cancer). Modeling these complex phenomena as random can be useful, and in either case, the interpretation of probability is the same: the chance that some event will occur.

Mathematicians and philosophers have struggled for centuries to arrive at a clear statement of how probability is defined, or what it means. The most common definition is used in this text.

PROBABILITY

The **probability** of an outcome is the proportion of times the outcome would occur if the random phenomenon could be observed an infinite number of times.

This definition of probability can be illustrated by simulation. Suppose a die is rolled many times. Let \hat{p}_n be the proportion of outcomes that are 1 after the first n rolls. As the number of rolls increases, \hat{p}_n will converge to the probability of rolling a 1, $p = 1/6$. Figure 2.2 shows this convergence for 100,000 die rolls. The tendency of \hat{p}_n to stabilize around p is described by the **Law of Large Numbers**. The behavior shown in Figure 2.2 matches most people's intuition about probability, but proving mathematically that the behavior is always true is surprisingly difficult and beyond the level of this text.

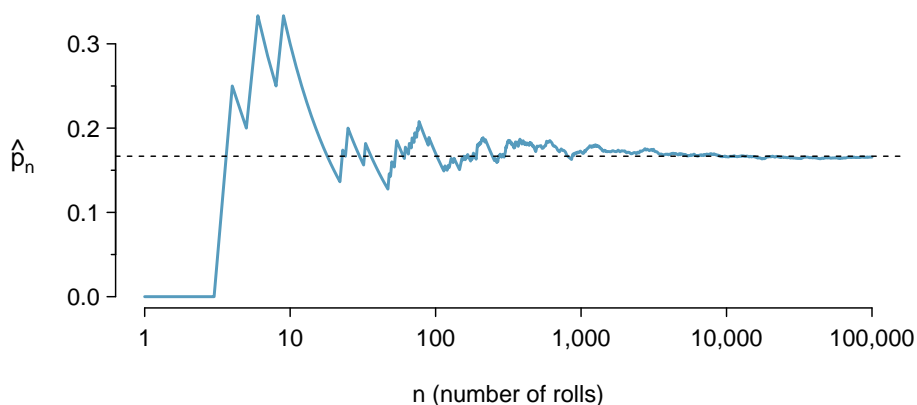


Figure 2.2: The fraction of die rolls that are 1 at each stage in a simulation. The proportion tends to get closer to the probability $1/6 \approx 0.167$ as the number of rolls increases.

Occasionally the proportion veers off from the probability and appear to defy the Law of Large Numbers, as \hat{p}_n does many times in Figure 2.2. However, the likelihood of these large deviations becomes smaller as the number of rolls increases.

²Since the father has CF, he must have two affected copies; he will always pass along a defective copy of the gene. Since the mother will pass along a defective copy half of the time, the child will be affected half of the time, or with probability $(1) \times (1/2) = 1/2$.

LAW OF LARGE NUMBERS

As more observations are collected, the proportion \hat{p}_n of occurrences with a particular outcome converges to the probability p of that outcome.

Probability is defined as a proportion, and it always takes values between 0 and 1 (inclusively). It may also be expressed as a percentage between 0% and 100%. The probability of rolling a 1, p , can also be written as $P(\text{rolling a 1})$.

This notation can be further abbreviated. For instance, if it is clear that the process is “rolling a die”, $P(\text{rolling a 1})$ can be written as $P(1)$. There also exists a notation for an event itself; the event A of rolling a 1 can be written as $A = \{\text{rolling a 1}\}$, with associated probability $P(A)$.

$P(A)$
Probability of
outcome A

2.1.3 Disjoint or mutually exclusive outcomes

Two outcomes are **disjoint** or **mutually exclusive** if they cannot both happen at the same time. When rolling a die, the outcomes 1 and 2 are disjoint since they cannot both occur. However, the outcomes 1 and “rolling an odd number” are not disjoint since both occur if the outcome of the roll is a 1.³

What is the probability of rolling a 1 or a 2? When rolling a die, the outcomes 1 and 2 are disjoint. The probability that one of these outcomes will occur is computed by adding their separate probabilities:

$$P(1 \text{ or } 2) = P(1) + P(2) = 1/6 + 1/6 = 1/3.$$

What about the probability of rolling a 1, 2, 3, 4, 5, or 6? Here again, all of the outcomes are disjoint, so add the individual probabilities:

$$\begin{aligned} P(1 \text{ or } 2 \text{ or } 3 \text{ or } 4 \text{ or } 5 \text{ or } 6) \\ &= P(1) + P(2) + P(3) + P(4) + P(5) + P(6) \\ &= 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1. \end{aligned}$$

ADDITION RULE OF DISJOINT OUTCOMES

If A_1 and A_2 represent two disjoint outcomes, then the probability that either one of them occurs is given by

$$P(A_1 \text{ or } A_2) = P(A_1) + P(A_2).$$

If there are k disjoint outcomes A_1, \dots, A_k , then the probability that either one of these outcomes will occur is

$$P(A_1) + P(A_2) + \dots + P(A_k). \quad (2.6)$$

³The terms *disjoint* and *mutually exclusive* are equivalent and interchangeable.

GUIDED PRACTICE 2.7

G

Consider the CF example. Is the event that two carriers of CF have a child that is also a carrier represented by mutually exclusive outcomes? Calculate the probability of this event.⁴

Probability problems often deal with *sets* or *collections* of outcomes. Let A represent the event in which a die roll results in 1 or 2 and B represent the event that the die roll is a 4 or a 6. We write A as the set of outcomes $\{1, 2\}$ and $B = \{4, 6\}$. These sets are commonly called **events**. Because A and B have no elements in common, they are disjoint events. A and B are represented in Figure 2.3.

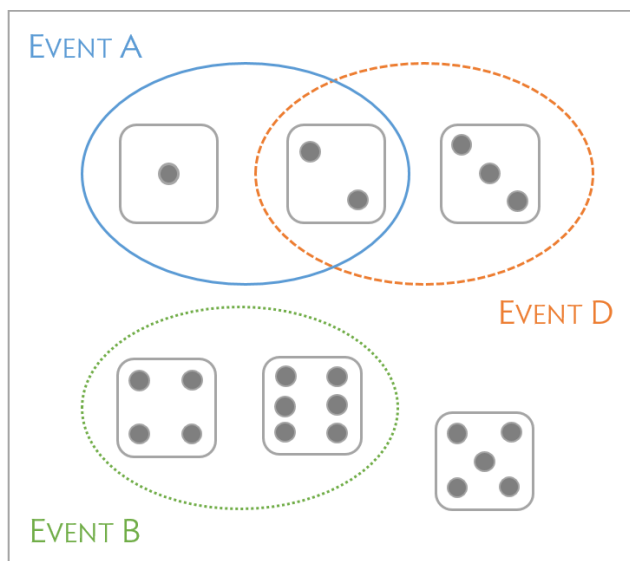


Figure 2.3: Three events, A , B , and D , consist of outcomes from rolling a die. A and B are disjoint since they do not have any outcomes in common.

The Addition Rule applies to both disjoint outcomes and disjoint events. The probability that one of the disjoint events A or B occurs is the sum of the separate probabilities:

$$P(A \text{ or } B) = P(A) + P(B) = 1/3 + 1/3 = 2/3.$$

GUIDED PRACTICE 2.8

G

(a) Verify the probability of event A , $P(A)$, is $1/3$ using the Addition Rule. (b) Do the same for event B .⁵

GUIDED PRACTICE 2.9

G

(a) Using Figure 2.3 as a reference, which outcomes are represented by event D ? (b) Are events B and D disjoint? (c) Are events A and D disjoint?⁶

⁴Yes, there are two mutually exclusive outcomes for which a child of two carriers can also be a carrier - a child can either receive an affected copy of *CFTR* from the mother and a normal copy from the father, or vice versa (since each parent can only contribute one allele). Thus, the probability that a child will be a carrier is $1/4 + 1/4 = 1/2$.

⁵(a) $P(A) = P(1 \text{ or } 2) = P(1) + P(2) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$. (b) Similarly, $P(B) = 1/3$.

⁶(a) Outcomes 2 and 3. (b) Yes, events B and D are disjoint because they share no outcomes. (c) The events A and D share an outcome in common, 2, and so are not disjoint.

GUIDED PRACTICE 2.10

G

In Guided Practice 2.9, you confirmed B and D from Figure 2.3 are disjoint. Compute the probability that event B or event D occurs.⁷

2.1.4 Probabilities when events are not disjoint

Venn diagrams are useful when outcomes can be categorized as “in” or “out” for two or three variables, attributes, or random processes. The Venn diagram in Figure 2.5 uses one oval to represent diamonds and another to represent face cards (the cards labeled jacks, queens, and kings); if a card is both a diamond and a face card, it falls into the intersection of the ovals.

2♣	3♣	4♣	5♣	6♣	7♣	8♣	9♣	10♣	J♣	Q♣	K♣	A♣
2♦	3♦	4♦	5♦	6♦	7♦	8♦	9♦	10♦	J♦	Q♦	K♦	A♦
2♥	3♥	4♥	5♥	6♥	7♥	8♥	9♥	10♥	J♥	Q♥	K♥	A♥
2♠	3♠	4♠	5♠	6♠	7♠	8♠	9♠	10♠	J♠	Q♠	K♠	A♠

Figure 2.4: A regular deck of 52 cards is split into four suits: ♣ (club), ♦ (diamond), ♥ (heart), ♠ (spade). Each suit has 13 labeled cards: 2, 3, ..., 10, J (jack), Q (queen), K (king), and A (ace). Thus, each card is a unique combination of a suit and a label, e.g. 4♥ and J♣.

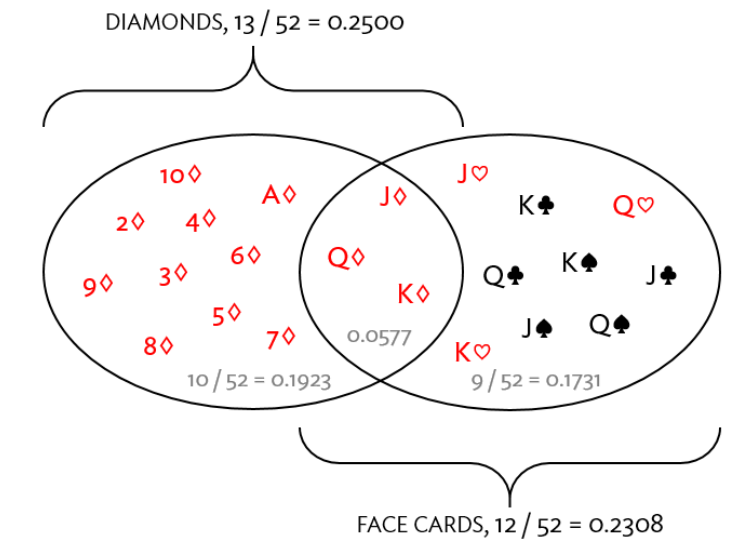


Figure 2.5: A Venn diagram for diamonds and face cards.

GUIDED PRACTICE 2.11

G

(a) What is the probability that a randomly selected card is a diamond? (b) What is the probability that a randomly selected card is a face card?⁸

⁷Since B and D are disjoint events, use the Addition Rule: $P(B \text{ or } D) = P(B) + P(D) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$.

⁸(a) There are 52 cards and 13 diamonds. If the cards are thoroughly shuffled, each card has an equal chance of being drawn, so the probability that a randomly selected card is a diamond is $P(\diamond) = \frac{13}{52} = 0.250$. (b) Likewise, there are 12 face cards, so $P(\text{face card}) = \frac{12}{52} = \frac{3}{13} = 0.231$.

Let A represent the event that a randomly selected card is a diamond and B represent the event that it is a face card. Events A and B are not disjoint – the cards $J\heartsuit$, $Q\heartsuit$, and $K\heartsuit$ fall into both categories.

As a result, adding the probabilities of the two events together is not sufficient to calculate $P(A \text{ or } B)$:

$$P(A) + P(B) = P(\heartsuit) + P(\text{face card}) = 12/52 + 13/52.$$

Instead, a small modification is necessary. The three cards that are in both events were counted twice. To correct the double counting, subtract the probability that both events occur:

$$\begin{aligned} P(A \text{ or } B) &= P(\text{face card or } \heartsuit) \\ &= P(\text{face card}) + P(\heartsuit) - P(\text{face card and } \heartsuit) \\ &= 13/52 + 12/52 - 3/52 \\ &= 22/52 = 11/26. \end{aligned} \tag{2.12}$$

Equation (2.12) is an example of the **General Addition Rule**.

GENERAL ADDITION RULE

If A and B are any two events, disjoint or not, then the probability that at least one of them will occur is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B), \tag{2.13}$$

where $P(A \text{ and } B)$ is the probability that both events occur.

Note that in the language of statistics, "or" is inclusive such that A or B occurs means A , B , or both A and B occur.

GUIDED PRACTICE 2.14

G

(a) If A and B are disjoint, describe why this implies $P(A \text{ and } B) = 0$. (b) Using part (a), verify that the General Addition Rule simplifies to the Addition Rule for disjoint events if A and B are disjoint.⁹

GUIDED PRACTICE 2.15

G

Human immunodeficiency virus (HIV) and tuberculosis (TB) affect substantial proportions of the population in certain areas of the developing world. Individuals sometimes are co-infected (i.e., have both diseases). Children of HIV-infected mothers may have HIV and TB can spread from one family member to another. In a mother-child pair, let $A = \{\text{the mother has HIV}\}$, $B = \{\text{the mother has TB}\}$, $C = \{\text{the child has HIV}\}$, $D = \{\text{the child has TB}\}$. Write out the definitions of the events $A \text{ or } B$, $A \text{ and } B$, $A \text{ and } C$, $A \text{ or } D$.¹⁰

⁹(a) If A and B are disjoint, A and B can never occur simultaneously. (b) If A and B are disjoint, then the last term of Equation (2.13) is 0 (see part (a)) and we are left with the Addition Rule for disjoint events.

¹⁰Events $A \text{ or } B$: the mother has HIV, the mother has TB, or the mother has both HIV and TB. Events $A \text{ and } B$: the mother has both HIV and TB. Events $A \text{ and } C$: The mother has HIV and the child has HIV. $A \text{ or } D$: The mother has HIV, the child has TB, or the mother has HIV and the child has TB.

2.1.5 Complement of an event

Rolling a die produces a value in the set $\{1, 2, 3, 4, 5, 6\}$. This set of all possible outcomes is called the **sample space** (S) for rolling a die.

Let $D = \{2, 3\}$ represent the event that the outcome of a die roll is 2 or 3. The **complement** of D represents all outcomes in the sample space that are not in D , which is denoted by $D^c = \{1, 4, 5, 6\}$. That is, D^c is the set of all possible outcomes not already included in D . Figure 2.6 shows the relationship between D , D^c , and the sample space S .

S
Sample space
 A^c
Complement
of outcome A

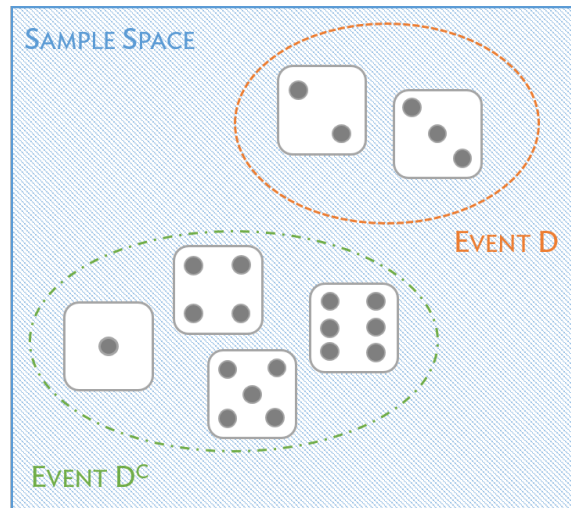


Figure 2.6: Event $D = \{2, 3\}$ and its complement, $D^c = \{1, 4, 5, 6\}$. S represents the sample space, which is the set of all possible events.

GUIDED PRACTICE 2.16

G

(a) Compute $P(D^c) = P(\text{rolling a 1, 4, 5, or 6})$. (b) What is $P(D) + P(D^c)$?¹¹

GUIDED PRACTICE 2.17

G

Events $A = \{1, 2\}$ and $B = \{4, 6\}$ are shown in Figure 2.3 on page 34. (a) Write out what A^c and B^c represent. (b) Compute $P(A^c)$ and $P(B^c)$. (c) Compute $P(A) + P(A^c)$ and $P(B) + P(B^c)$.¹²

A complement of an event A is constructed to have two very important properties: every possible outcome not in A is in A^c , and A and A^c are disjoint. If every possible outcome not in A is in A^c , this implies that

$$P(A \text{ or } A^c) = 1. \quad (2.18)$$

Then, by Addition Rule for disjoint events,

$$P(A \text{ or } A^c) = P(A) + P(A^c). \quad (2.19)$$

Combining Equations (2.18) and (2.19) yields a useful relationship between the probability of an event and its complement.

¹¹(a) The outcomes are disjoint and each has probability $1/6$, so the total probability is $4/6 = 2/3$. (b) We can also see that $P(D) = \frac{1}{6} + \frac{1}{6} = 1/3$. Since D and D^c are disjoint, $P(D) + P(D^c) = 1$.

¹²Brief solutions: (a) $A^c = \{3, 4, 5, 6\}$ and $B^c = \{1, 2, 3, 5\}$. (b) Noting that each outcome is disjoint, add the individual outcome probabilities to get $P(A^c) = 2/3$ and $P(B^c) = 2/3$. (c) A and A^c are disjoint, and the same is true of B and B^c . Therefore, $P(A) + P(A^c) = 1$ and $P(B) + P(B^c) = 1$.

COMPLEMENT

The complement of event A is denoted A^c , and A^c represents all outcomes not in A . A and A^c are mathematically related:

$$P(A) + P(A^c) = 1, \quad \text{i.e.} \quad P(A) = 1 - P(A^c). \quad (2.20)$$

In simple examples, computing either A or A^c is feasible in a few steps. However, as problems grow in complexity, using the relationship between an event and its complement can be a useful strategy.

GUIDED PRACTICE 2.21

G

Let A represent the event of selecting an adult from the US population with height between 180 and 185 cm, as calculated in Example 2.32. What is $P(A^c)$?¹³

GUIDED PRACTICE 2.22

G

Let A represent the event in which two dice are rolled and their total is less than 12. (a) What does the event A^c represent? (b) Determine $P(A^c)$ from Figure 2.11 on page 44. (c) Determine $P(A)$.¹⁴

GUIDED PRACTICE 2.23

G

Consider again the probabilities from Figure 2.11 and rolling two dice. Find the following probabilities: (a) The sum of the dice is *not* 6. (b) The sum is at least 4. That is, determine the probability of the event $B = \{4, 5, \dots, 12\}$. (c) The sum is no more than 10. That is, determine the probability of the event $D = \{2, 3, \dots, 10\}$.¹⁵

2.1.6 Independence

Just as variables and observations can be independent, random phenomena can also be independent. Two processes are **independent** if knowing the outcome of one provides no information about the outcome of the other. For instance, flipping a coin and rolling a die are two independent processes – knowing that the coin lands heads up does not help determine the outcome of the die roll. On the other hand, stock prices usually move up or down together, so they are not independent.

¹³ $P(A^c) = 1 - P(A) = 1 - 0.1157 = 0.8843$.

¹⁴(a) The complement of A : when the total is equal to 12. (b) $P(A^c) = 1/36$. (c) Use the probability of the complement from part (b), $P(A^c) = 1/36$, and Equation (2.20): $P(\text{less than } 12) = 1 - P(12) = 1 - 1/36 = 35/36$.

¹⁵(a) First find $P(6) = 5/36$, then use the complement: $P(\text{not } 6) = 1 - P(6) = 31/36$.

(b) First find the complement, which requires much less effort: $P(2 \text{ or } 3) = 1/36 + 2/36 = 1/12$. Then calculate $P(B) = 1 - P(B^c) = 1 - 1/12 = 11/12$.

(c) As before, finding the complement is the more direct way to determine $P(D)$. First find $P(D^c) = P(11 \text{ or } 12) = 2/36 + 1/36 = 1/12$. Then calculate $P(D) = 1 - P(D^c) = 11/12$.

Example 2.3 provides a basic example of two independent processes: rolling two dice. What is the probability that both will be 1? Suppose one of the dice is blue and the other green. If the outcome of the blue die is a 1, it provides no information about the outcome of the green die. This question was first encountered in Example 2.3: $1/6^{\text{th}}$ of the time the blue die is a 1, and $1/6^{\text{th}}$ of *those* times the green die will also be 1. This is illustrated in Figure 2.7. Because the rolls are independent, the probabilities of the corresponding outcomes can be multiplied to obtain the final answer: $(1/6)(1/6) = 1/36$. This can be generalized to many independent processes.

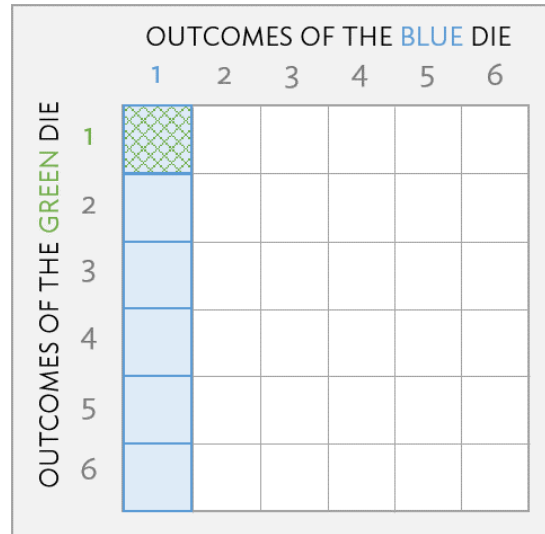


Figure 2.7: $1/6^{\text{th}}$ of the time, the first roll is a 1. Then $1/6^{\text{th}}$ of *those* times, the second roll will also be a 1.

Complicated probability problems, such as those that arise in biology or medicine, are often solved with the simple ideas used in the dice example. For instance, independence was used implicitly in the second solution to Example 2.4, when calculating the probability that two carriers will have an affected child with cystic fibrosis. Genes are typically passed along from the mother and father independently. This allows for the assumption that, on average, half of the offspring who receive a mutated gene copy from the mother will also receive a mutated copy from the father.

GUIDED PRACTICE 2.24

- (G) What if there were also a red die independent of the other two? What is the probability of rolling the three dice and getting all 1s?¹⁶

GUIDED PRACTICE 2.25

- (G) Three US adults are randomly selected. The probability the height of a single adult is between 180 and 185 cm is 0.1157.¹⁷

- What is the probability that all three are between 180 and 185 cm tall?
- What is the probability that none are between 180 and 185 cm tall?

¹⁶The same logic applies from Example 2.3. If $1/36^{\text{th}}$ of the time the blue and green dice are both 1, then $1/6^{\text{th}}$ of *those* times the red die will also be 1, so multiply:

$$\begin{aligned} P(\text{blue} = 1 \text{ and } \text{green} = 1 \text{ and } \text{red} = 1) &= P(\text{blue} = 1)P(\text{green} = 1)P(\text{red} = 1) \\ &= (1/6)(1/6)(1/6) = 1/216. \end{aligned}$$

¹⁷Brief answers: (a) $0.1157 \times 0.1157 \times 0.1157 = 0.0015$. (b) $(1 - 0.1157)^3 = 0.692$.

MULTIPLICATION RULE FOR INDEPENDENT PROCESSES

If A and B represent events from two different and independent processes, then the probability that both A and B occur is given by:

$$P(A \text{ and } B) = P(A)P(B). \quad (2.26)$$

Similarly, if there are k events A_1, \dots, A_k from k independent processes, then the probability they all occur is

$$P(A_1)P(A_2)\cdots P(A_k).$$

EXAMPLE 2.27

Mandatory drug testing. Mandatory drug testing in the workplace is common practice for certain professions, such as air traffic controllers and transportation workers. A false positive in a drug screening test occurs when the test incorrectly indicates that a screened person is an illegal drug user. Suppose a mandatory drug test has a false positive rate of 1.2% (i.e., has probability 0.012 of indicating that an employee is using illegal drugs when that is not the case). Given 150 employees who are in reality drug free, what is the probability that at least one will (falsely) test positive? Assume that the outcome of one drug test has no effect on the others.

First, note that the complement of at least 1 person testing positive is that no one tests positive (i.e., all employees test negative). The multiplication rule can then be used to calculate the probability of 150 negative tests.

$$\begin{aligned} P(\text{At least 1 "+"}) &= P(1 \text{ or } 2 \text{ or } 3 \dots \text{or } 150 \text{ are "+"}) \\ &= 1 - P(\text{None are "+"}) \\ &= 1 - P(150 \text{ are "-"}) \\ &= 1 - P("-")^{150} \\ &= 1 - (0.988)^{150} = 1 - 0.16 = 0.84. \end{aligned}$$

Even when using a test with a small probability of a false positive, the company is more than 80% likely to incorrectly claim at least one employee is an illegal drug user!

GUIDED PRACTICE 2.28

Because of the high likelihood of at least one false positive in company wide drug screening programs, an individual with a positive test is almost always re-tested with a different screening test: one that is more expensive than the first, but has a lower false positive probability. Suppose the second test has a false positive rate of 0.8%. What is the probability that an employee who is not using illegal drugs will test positive on both tests?¹⁸

¹⁸The outcomes of the two tests are independent of one another; $P(A \text{ and } B) = P(A) \times P(B)$, where events A and B are the results of the two tests. The probability of a false positive with the first test is 0.012 and 0.008 with the second. Thus, the probability of an employee who is not using illegal drugs testing positive on both tests is $0.012 \times 0.008 = 9.6 \times 10^{-5}$

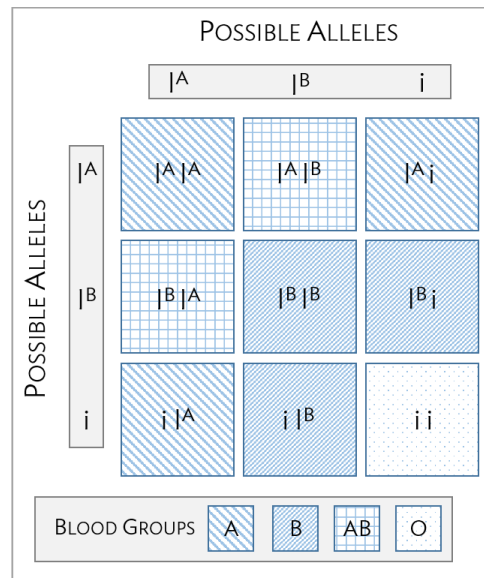


Figure 2.8: Inheritance of ABO blood groups.

EXAMPLE 2.29

ABO blood groups. There are four different common blood types (A, B, AB, and O), which are determined by the presence of certain antigens located on cell surfaces. Antigens are substances used by the immune system to recognize self versus non-self; if the immune system encounters antigens not normally found on the body's own cells, it will attack the foreign cells. When patients receive blood transfusions, it is critical that the antigens of transfused cells match those of the patient's, or else an immune system response will be triggered.

The ABO blood group system consists of four different blood groups, which describe whether an individual's red blood cells carry the A antigen, B antigen, both, or neither. The ABO gene has three alleles: I^A , I^B , and i . The i allele is recessive to both I^A and I^B , and does not produce antigens; thus, an individual with genotype $I^A i$ is blood group A and an individual with genotype $I^B i$ is blood group B. The I^A and I^B alleles are codominant, such that individuals of $I^A I^B$ genotype are AB. Individuals homozygous for the i allele are known as blood group O, with neither A nor B antigens.

Suppose that both members of a couple have Group AB blood.

- What is the probability that a child of this couple will have Group A blood?
- What is the probability that they have two children with Group A blood?

a) An individual with Group AB blood is genotype $I^A I^B$. Two $I^A I^B$ parents can produce children with genotypes $I^A I^B$, $I^A I^A$, or $I^B I^B$. Of these possibilities, only children with genotype $I^A I^A$ have Group A blood. Each parent has 0.5 probability of passing down their I^A allele. Thus, the probability that a child of this couple will have Group A blood is $P(\text{parent 1 passes down } I^A \text{ allele}) \times P(\text{parent 2 passes down } I^A \text{ allele}) = 0.5 \times 0.5 = 0.25$.

b) Inheritance of alleles is independent between children. Thus, the probability of two children having Group A blood equals $P(\text{child 1 has Group A blood}) \times P(\text{child 2 has group A blood})$. The probability of a child of this couple having Group A blood was previously calculated as 0.25. The answer is given by $0.25 \times 0.25 = 0.0625$.

The previous examples in this section have used independence to solve probability problems. The definition of independence can also be used to check whether two events are independent – two events A and B are independent if they satisfy Equation (2.26).

EXAMPLE 2.30

Is the event of drawing a heart from a deck of cards independent of drawing an ace?

The probability the card is a heart is $1/4$ ($13/52 = 1/4$) and the probability that it is an ace is $1/13$ ($4/52 = 1/13$). The probability that the card is the ace of hearts ($A \heartsuit$) is $1/52$. Check whether Equation 2.26 is satisfied:

$$P(\heartsuit)P(A) = \left(\frac{1}{4}\right)\left(\frac{1}{13}\right) = \frac{1}{52} = P(\heartsuit \text{ and } A).$$

Since the equation holds, the event that the card is a heart and the event that the card is an ace are independent events.

EXAMPLE 2.31

In the general population, about 15% of adults between 25 and 40 years of age are hypertensive. Suppose that among males of this age, hypertension occurs about 18% of the time. Is hypertension independent of sex?

Assume that the population is 50% male, 50% female; it is given in the problem that hypertension occurs about 15% of the time in adults between ages 25 and 40.

$$P(\text{hypertension}) \times P(\text{male}) = (0.15)(0.50) = 0.075 \neq 0.18.$$

Equation 2.26 is not satisfied, therefore hypertension is not independent of sex. In other words, knowing whether an individual is male or female is informative as to whether they are hypertensive. If hypertension and sex were independent, then we would expect hypertension to occur at an equal rate in males as in females.

2.2 Random variables

2.2.1 Distributions of random variables

Formally, a **random variable** assigns numerical values to the outcome of a random phenomenon, and is usually written with a capital letter such as X , Y , or Z .

If a coin is tossed three times, the outcome is the sequence of observed heads and tails. One such outcome might be TTH: tails on the first two tosses, heads on the third. If the random variable X is the number of heads for the three tosses, $X = 1$; if Y is the number of tails, then $Y = 2$. For the sequence THT, only the order has changed, but the values of X and Y remain the same. For the sequence HHH, however, $X = 3$ and $Y = 0$. Even in this simple setting, is possible to define other random variables; for example, if Z is the toss when the first H occurs, then $Z = 3$ for the first set of tosses (TTH) and 1 for the third set (HHH).


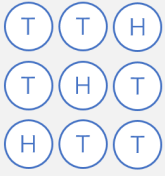
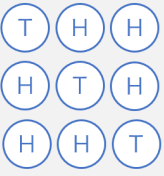

			
$X = 0$	$X = 1$	$X = 2$	$X = 3$

Figure 2.9: Possible outcomes for number of heads in three tosses of a coin.

If probabilities can be assigned to the outcomes in a random phenomenon or study, then those can be used to assign probabilities to values of a random variable. Using independence, $P(\text{HHH}) = (1/2)^3 = 1/8$. Since X in the above example can only be three if the three tosses are all heads, $P(X = 3) = 1/8$. The distribution of a random variable is the collection of probabilities for all of the variable's unique values. Figure 2.9 shows the eight possible outcomes when a coin is tossed three times: TTT, HTT, THT, TTH, HHT, HTH, THH, HHH. For the first set of tosses, $X = 0$; for the next three, $X = 1$, then $X = 2$ for the following three tosses and $X = 3$ for the last set (HHH).

Using independence again, each of the 8 outcomes have probability $1/8$, so $P(X = 0) = P(X = 3) = 1/8$ and $P(X = 1) = P(X = 2) = 3/8$. Figure 2.10 shows the probability distribution for X . Probability distributions for random variables follow the rules for probability; for instance, the sum of the probabilities must be 1.00. The possible outcomes of X are labeled with a corresponding lower case letter x and subscripts. The values of X are $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, and $x_4 = 3$; these occur with probabilities $1/8$, $3/8$, $3/8$ and $1/8$.

i	1	2	3	4	Total
x_i	0	1	2	3	–
$P(X = x_i)$	1/8	3/8	3/8	1/8	8/8 = 1.00

Figure 2.10: Tabular form for the distribution of the number of heads in three coin tosses.

Another example of probability distribution consists of all disjoint outcomes and their associated probabilities. Figure 2.11 shows the probability distribution for the sum of two dice.

Dice sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Figure 2.11: Probability distribution for the sum of two dice.

RULES FOR A PROBABILITY DISTRIBUTION

A probability distribution is a list of all possible outcomes and their associated probabilities that satisfies three rules:

- 1. The outcomes listed must be disjoint.
- 2. Each probability must be between 0 and 1.
- 3. The probabilities must total to 1.

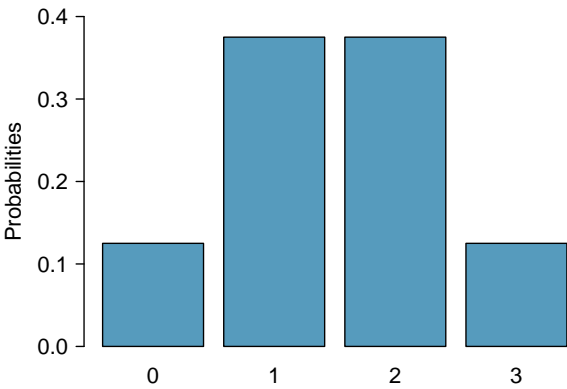


Figure 2.12: Bar plot of the distribution of the number of heads in three coin tosses.

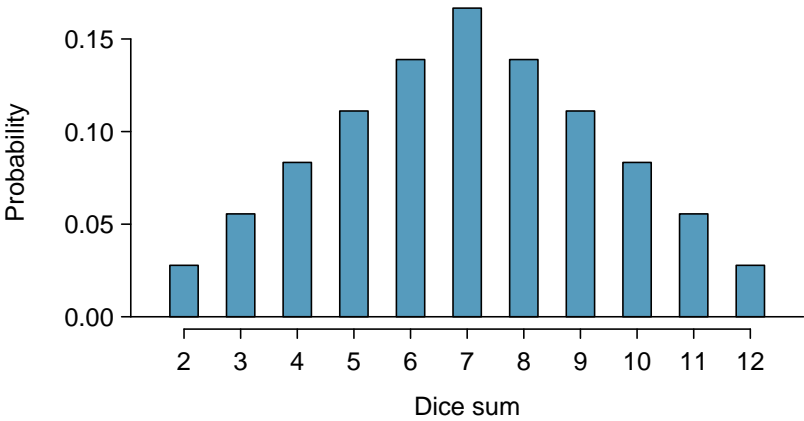


Figure 2.13: The probability distribution of the sum of two dice.

Bar graphs can be used to show the distribution of a random variable. Figure 2.12 is a bar graph of the distribution of X in the coin tossing example and Figure 2.13 is a bar graph of the distribution in the sum of two dice. When bar graphs are used to show the distribution of a dataset, the heights of the bars show the frequency of observations; in contrast, bar heights for a probability distribution show the probabilities of possible values of a random variable.

X is an example of a **discrete random variable** since it takes on a finite number of values.¹⁹

A **continuous random variable** can take on any real value in an interval.

Consider how the probability distribution for adult heights in the US might best be represented. Unlike the sum of two dice rolls, height can occupy any value over a continuous range. Thus, height is a random variable which has a continuous probability distribution, which is specified by a **probability density function** rather than a table; Figure 2.14 shows a histogram of the height for 3 million US adults from the mid-1990's, with an overlaid density curve.²⁰

Just as in the discrete case, the probabilities of all possible outcomes must still sum to 1; the total area under a probability density function equals 1.

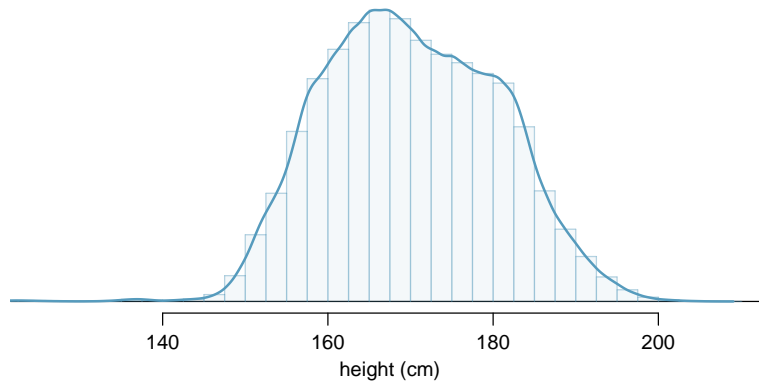


Figure 2.14: The continuous probability distribution of heights for US adults.

EXAMPLE 2.32

Estimate the probability that a randomly selected adult from the US population has height between 180 and 185 centimeters. In Figure 2.15(a), the two bins between 180 and 185 centimeters have counts of 195,307 and 156,239 people.

Find the proportion of the histogram's area that falls in the range 180 cm and 185: add the heights of the bins in the range and divide by the sample size:

$$\frac{195,307 + 156,239}{3,000,000} = 0.1172.$$

The probability can be calculated precisely with the use of computing software, by finding the area of the shaded region under the curve between 180 and 185:

$$P(\text{height between 180 and 185}) = \text{area between 180 and 185} = 0.1157.$$

¹⁹Some discrete random variables have an infinite number of possible values, such as all the non-negative integers.

²⁰This sample can be considered a simple random sample from the US population. It relies on the USDA Food Commodity Intake Database.

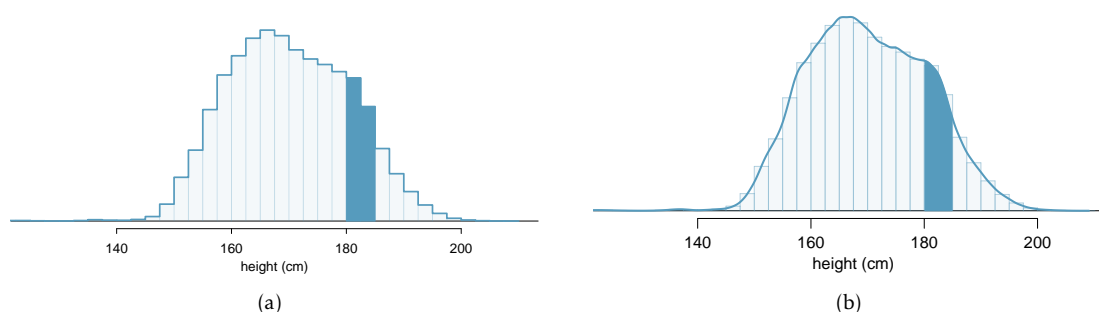


Figure 2.15: (a) A histogram with bin sizes of 2.5 cm, with bars between 180 and 185 cm shaded. (b) Density for heights in the US adult population with the area between 180 and 185 cm shaded.

EXAMPLE 2.33

What is the probability that a randomly selected person is **exactly** 180 cm? Assume that height can be measured perfectly.

E

This probability is zero. A person might be close to 180 cm, but not exactly 180 cm tall. This also coheres with the definition of probability as an area under the density curve; there is no area captured between 180 cm and 180 cm.

GUIDED PRACTICE 2.34

G

Suppose a person's height is rounded to the nearest centimeter. Is there a chance that a random person's **measured** height will be 180 cm?²¹

SOME PROPERTIES OF CONTINUOUS DISTRIBUTIONS

Let X be a continuous random variable, a and b real numbers.

- | | |
|-----------------------------------|---|
| 1. $P(X = a) = 0$ | 4. $P(X > a) = 1 - P(X < a)$ |
| 2. $P(X \leq a) = P(X < a)$ | |
| 3. $P(-\infty < X < +\infty) = 1$ | 5. $P(a < X < b) = P(X < b) - P(X < a)$ |

2.2.2 Expectation

Just like distributions of data, distributions of random variables also have means, variances, standard deviations, medians, etc.; these characteristics are computed a bit differently for random variables. The mean of a random variable is called its **expected value** and written $E(X)$. To calculate the mean of a random variable, multiply each possible value by its corresponding probability and add these products.

²¹This has positive probability. Anyone between 179.5 cm and 180.5 cm will have a *measured* height of 180 cm. This a more realistic scenario to encounter in practice versus Example 2.33.

EXPECTED VALUE OF A DISCRETE RANDOM VARIABLE

If X takes on outcomes x_1, \dots, x_k with probabilities $P(X = x_1), \dots, P(X = x_k)$, the expected value of X is the sum of each outcome multiplied by its corresponding probability:

$$\begin{aligned} E(X) &= x_1 P(X = x_1) + \dots + x_k P(X = x_k) \\ &= \sum_{i=1}^k x_i P(X = x_i). \end{aligned} \quad (2.35)$$

The Greek letter μ may be used in place of the notation $E(X)$.

EXAMPLE 2.36

Calculate the expected value of X , where X represents the number of heads in three tosses of a fair coin.

X can take on values 0, 1, 2, and 3. The probability of each x_k is given in Figure 2.10.

E

$$\begin{aligned} E(X) &= x_1 P(X = x_1) + \dots + x_k P(X = x_k) \\ &= (0)(P(X = 0)) + (1)(P(X = 1)) + (2)(P(X = 2)) + (3)(P(X = 3)) \\ &= (0)(1/8) + (1)(3/8) + (2)(3/8) + (3)(1/8) = 12/8 \\ &= 1.5. \end{aligned}$$

$E(X)$
Expected Value
of X

The expected value of X is 1.5.

The expected value for a random variable represents the average outcome. For example, $E(X) = 1.5$ represents the average number of heads in three tosses of a coin, if the three tosses were repeated many times.²² It often happens with discrete random variables that the expected value is not precisely one of the possible outcomes of the variable.

GUIDED PRACTICE 2.37**G**

Calculate the expected value of Y , where Y represents the number of heads in three tosses of an unfair coin, where the probability of heads is 0.70.²³

It is also possible to compute the expected value of a continuous random variable. However, it requires a little calculus and this is beyond the scope of this course.²⁴

In Physics, the expectation holds the same meaning as the center of gravity. The distribution can be represented by a series of weights at each outcome, and the mean represents the balancing point. Figure 2.16 shows a continuous probability distribution balanced atop a wedge placed at the mean.

²²The expected value $E(X)$ can also be expressed as μ , e.g. $\mu = 1.5$

²³First, calculate the probability distribution. $P(Y = 0) = (1 - 0.70)^3 = 0.027$ and $P(Y = 3) = (0.70)^3 = 0.343$. Note that there are three ways to obtain 1 head (HTT, THT, TTH), thus, $P(Y = 1) = (3)(0.70)(1 - 0.70)^2 = 0.189$. By the same logic, $P(Y = 2) = (3)(0.70)^2(1 - 0.70) = 0.441$. Thus, $E(Y) = (0)(0.027) + (1)(0.189) + (2)(0.441) + (3)(0.343) = 2.1$. The expected value of Y is 2.1.

²⁴ $\mu = \int xf(x)dx$ where $f(x)$ represents the probability density function of the random variable.

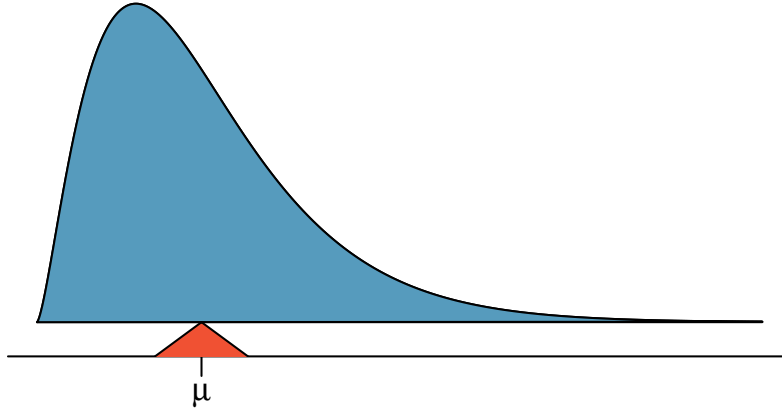


Figure 2.16: A continuous distribution can also be balanced at its mean.

2.2.3 Variability of random variables

The variability of a random variable can be described with variance and standard deviation. For data, the variance is computed by squaring deviations from the mean ($x_i - \mu$) and then averaging over the number of values in the dataset (Section 1.4.3).

In the case of a random variable, the squared deviations from the mean of the random variable are used instead, and their sum is weighted by the corresponding probabilities. This weighted sum of squared deviations equals the variance; the standard deviation is the square root of the variance.

VARIANCE OF A DISCRETE RANDOM VARIABLE

If X takes on outcomes x_1, \dots, x_k with probabilities $P(X = x_1), \dots, P(X = x_k)$ and expected value $\mu = E(X)$, then the variance of X , denoted by $\text{Var}(X)$ or σ^2 , is

$$\begin{aligned} \text{Var}(X) &= (x_1 - \mu)^2 P(X = x_1) + \dots + (x_k - \mu)^2 P(X = x_k) \\ &= \sum_{i=1}^k (x_i - \mu)^2 P(X = x_i). \end{aligned} \quad (2.38)$$

The standard deviation of X , labeled $SD(X)$ or σ , is the square root of the variance.

The variance of a random variable can be interpreted as the expectation of the terms $(x_i - \mu)^2$; i.e., $\sigma^2 = E(X - \mu)^2$. While this compact form is not useful for direct computation, it can be helpful for understanding the concept of variability in the context of a random variable; variance is simply the average of the deviations from the mean.

$\text{Var}(X)$
Variance
of X

EXAMPLE 2.39

Compute the variance and standard deviation of X , the number of heads in three tosses of a fair coin.

In the formula for the variance, $k = 4$ and $\mu_X = E(X) = 1.5$.

$$\begin{aligned}\sigma_X^2 &= (x_1 - \mu_X)^2 P(X = x_1) + \cdots + (x_4 - \mu)^2 P(X = x_4) \\ &= (0 - 1.5)^2(1/8) + (1 - 1.5)^2(3/8) + (2 - 1.5)^2(3/8) + (3 - 1.5)^2(1/8) \\ &= 3/4.\end{aligned}$$

The variance is $3/4 = 0.75$ and the standard deviation is $\sqrt{3/4} = 0.866$.

The coin tossing scenario provides a simple illustration of the mean and variance of a random variable.

2.2.4 Linear combinations of random variables

Sums of random variables arise naturally in many problems. In a health insurance, the amount spent by the employee during her next five years of employment can be represented as $X_1 + X_2 + X_3 + X_4 + X_5$, where X_1 is the cost of the first year, X_2 the second year, etc. If the employee's domestic partner has health insurance with another employer, the total annual cost to the couple would be the sum of the costs for the employee (X) and for her partner (Y), or $X + Y$. In each of these examples, it is intuitively clear that the average cost would be the sum of the average of each term.

Sums of random variables represent a special case of linear combinations of variables.

LINEAR COMBINATIONS OF RANDOM VARIABLES AND THEIR EXPECTED VALUES

If X and Y are random variables, then a linear combination of the random variables is given by

$$aX + bY,$$

where a and b are constants. The mean of a linear combination of random variables is

$$E(aX + bY) = aE(X) + bE(Y) = a\mu_X + b\mu_Y.$$

The formula easily generalizes to a sum of any number of random variables. For example, the average health care cost for 5 years, given that the cost for services remains the same, is

$$E(X_1 + X_2 + X_3 + X_4 + X_5) = E(5X_1) = 5E(X_1) = (5)(1010) = \$5,050.$$

The formula implies that for a random variable Z , $E(a + Z) = a + E(Z)$. This could have been used when calculating the average health costs for the employee by defining a as the fixed cost of the premium ($a = \$948$) and Z as the cost of the physician visits. Thus, the total annual cost for a year could be calculated as: $E(a + Z) = a + E(Z) = \$948 + E(Z) = \$948 + .30(1 \times \$20) + .40(3 \times \$20) + .20(4 \times \$20) + 0.10(8 \times \$20) = \$1,010.00$.

GUIDED PRACTICE 2.40**G**

Suppose the employee will begin a domestic partnership in the next year. Although she and her companion will begin living together and sharing expenses, they will each keep their existing health insurance plans; both, in fact, have the same plan from the same employer. In the last five years, her partner visited a physician only once in four of the ten years, and twice in the other six years. Calculate the expected total cost of health insurance to the couple in the next year.²⁵

Calculating the variance and standard deviation of a linear combination of random variables requires more care. The formula given here requires that the random variables in the linear combination be independent, such that an observation on one of the variables provides no information about the value of the other variable.

VARIABILITY OF LINEAR COMBINATIONS OF RANDOM VARIABLES

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y).$$

This equation is valid only if the random variables are independent of each other.

For the transformation $a + bZ$, the variance is $b^2\text{Var}(Z)$, since a constant a has variance 0. When $b = 1$, variance of $a + Z$ is $\text{Var}(Z)$ —adding a constant to a random variable has no effect on the variability of the random variable.

EXAMPLE 2.41

Calculate the variance and standard deviation for the combined cost of next year's health care for the two partners, assuming that the costs for each person are independent.

Let X represent the sum of costs for the employee and Y the sum of costs for her partner.

E

First, calculate the variance of health care costs for the partner. The partner's costs are the sum of the annual fixed cost and the variable annual costs, so the variance will simply be the variance of the variable costs. If Z represents the component of the variable costs, $E(Z) = 0.4(1 \times \$20) + 0.6(2 \times \$20) = \$8 + \$24 = \$32$. Thus, the variance of Z equals

$$\text{Var}(Z) = 0.4(20 - 32)^2 + 0.6(40 - 32)^2 = 96.$$

Under the assumption of independence, $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) = 1556 + 96 = 1652$, and the standard deviation is $\sqrt{1652} = \$40.64$.

²⁵Let X represent the costs for the employee and Y represent the costs for her partner. $E(X) = \$1,010.00$, as previously calculated. $E(Y) = 948 + 0.4(1 \times \$20) + 0.6(2 \times \$20) = \980.00 . Thus, $E(X + Y) = E(X) + E(Y) = \$1,010.00 + \$980.00 = \$1,990.00$.

2.3 Normal distribution

Among the many distributions seen in practice, one is by far the most common: the **normal distribution**, which has the shape of a symmetric, unimodal bell curve. Many variables are nearly normal, which makes the normal distribution useful for a variety of problems. For example, characteristics such as human height closely follow the normal distribution.

2.3.1 Normal distribution model

The normal distribution model always describes a symmetric, unimodal, bell-shaped curve. However, the curves can differ in center and spread; the model can be adjusted using mean and standard deviation. Changing the mean shifts the bell curve to the left or the right, while changing the standard deviation stretches or constricts the curve. Figure 2.17 shows the normal distribution with mean 0 and standard deviation 1 in the left panel and the normal distribution with mean 19 and standard deviation 4 in the right panel. Figure 2.18 shows these distributions on the same axis.

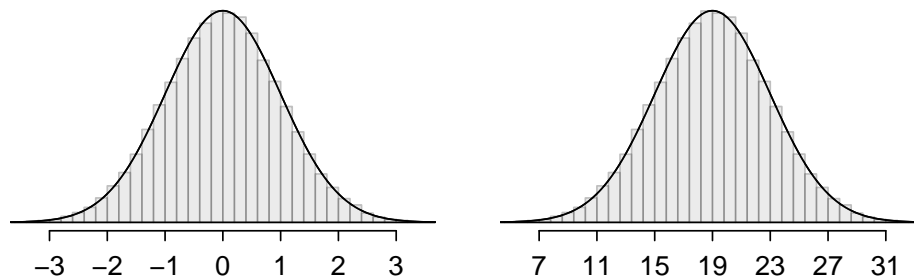


Figure 2.17: Both curves represent the normal distribution; however, they differ in their center and spread. The normal distribution with mean 0 and standard deviation 1 is called the **standard normal distribution**.

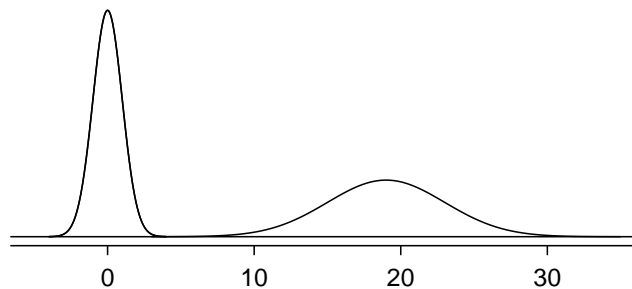


Figure 2.18: The normal models shown in Figure 2.17 but plotted together and on the same scale.

For any given normal distribution with mean μ and standard deviation σ , the distribution can be written as $N(\mu, \sigma)$; μ and σ are the parameters of the normal distribution. For example, $N(0, 1)$ refers to the standard normal distribution, as shown in Figure 2.17.

$N(\mu, \sigma)$
Normal dist.
with mean μ
& st. dev. σ

2.3.2 Standardizing with Z-scores

Z

Z-score, the
standardized
observation

The **Z-score** of an observation quantifies how far the observation is from the mean, in units of standard deviation(s). If x is an observation from a distribution $N(\mu, \sigma)$, the Z-score is mathematically defined as:

$$Z = \frac{x - \mu}{\sigma}.$$

An observation equal to the mean has a Z-score of 0. Observations above the mean have positive Z-scores, while observations below the mean have negative Z-scores. For example, if an observation is one standard deviation above the mean, it has a Z-score of 1; if it is 1.5 standard deviations below the mean, its Z-score is -1.5.

Z-scores can be used to identify which observations are more extreme than others, and are especially useful when comparing observations from different normal distributions. One observation x_1 is said to be more unusual than another observation x_2 if the absolute value of its Z-score is larger than the absolute value of the other observation's Z-score: $|Z_1| > |Z_2|$. In other words, the further an observation is from the mean in either direction, the more extreme it is.

EXAMPLE 2.42

The SAT and the ACT are two standardized tests commonly used for college admissions in the United States. The distribution of test scores are both nearly normal. For the SAT, $N(1500, 300)$; for the ACT, $N(21, 5)$. While some colleges request that students submit scores from both tests, others allow students the choice of either the ACT or the SAT. Suppose that one student scores an 1800 on the SAT (Student A) and another scores a 24 on the ACT (Student B). A college admissions officer would like to compare the scores of the two students to determine which student performed better.

Calculate a Z-score for each student; i.e., convert x to Z .

Using $\mu_{SAT} = 1500$, $\sigma_{SAT} = 300$, and $x_A = 1800$, find Student A's Z-score:

$$Z_A = \frac{x_A - \mu_{SAT}}{\sigma_{SAT}} = \frac{1800 - 1500}{300} = 1.$$

For Student B:

$$Z_B = \frac{x_B - \mu_{ACT}}{\sigma_{ACT}} = \frac{24 - 21}{5} = 0.6.$$

Student A's score is 1 standard deviation above average on the SAT, while Student B's score is 0.6 standard deviations above the mean on the ACT. As illustrated in Figure 2.19, Student A's score is more extreme, indicating that Student A has scored higher with respect to other scores than Student B.

THE Z-SCORE

The Z-score of an observation quantifies how far the observation is from the mean, in units of standard deviation(s). The Z-score for an observation x that follows a distribution with mean μ and standard deviation σ can be calculated using

$$Z = \frac{x - \mu}{\sigma}.$$

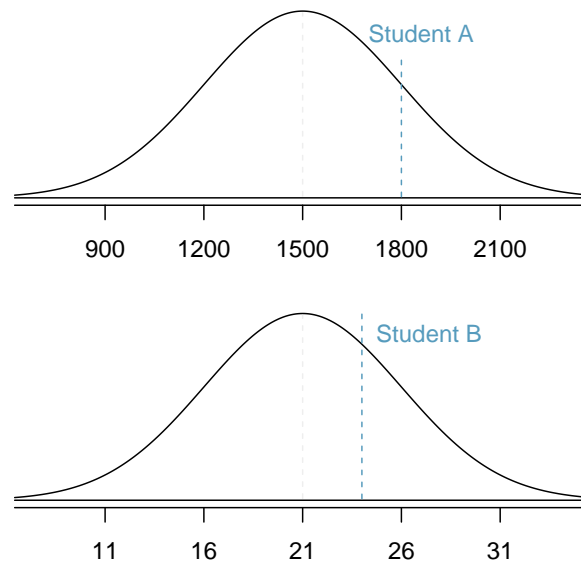


Figure 2.19: Scores of Students A and B plotted on the distributions of SAT and ACT scores.

EXAMPLE 2.43

How high would a student need to score on the ACT to have a score equivalent to Student A's score of 1800 on the SAT?

As shown in Example 2.19, a score of 1800 on the SAT is 1 standard deviation above the mean. ACT scores are normally distributed with mean 21 and standard deviation 5. To convert a value from the standard normal curve (Z) to one on a normal distribution $N(\mu, \sigma)$:

$$x = \mu + Z\sigma.$$

Thus, a student would need a score of $21 + 1(5) = 26$ on the ACT to have a score equivalent to 1800 on the SAT.

GUIDED PRACTICE 2.44

Systolic blood pressure (SBP) for adults in the United States aged 18-39 follow an approximate normal distribution, $N(115, 17.5)$. As age increases, systolic blood pressure also tends to increase. Mean systolic blood pressure for adults 60 years of age and older is 136 mm Hg, with standard deviation 40 mm Hg. Systolic blood pressure of 140 mm Hg or higher is indicative of hypertension (high blood pressure). (a) How many standard deviations away from the mean is a 30-year-old with systolic blood pressure of 125 mm Hg? (b) Compare how unusual a systolic blood pressure of 140 mm Hg is for a 65-year-old, versus a 30-year-old.²⁶

²⁶(a) Calculate the Z -score: $\frac{\bar{x} - \mu}{\sigma} = \frac{125 - 115}{17.5} = 0.571$. A 30-year-old with systolic blood pressure of 125 mm Hg is about 0.6 standard deviations above the mean. (b) For $x_1 = 140$ mm Hg: $Z_1 = \frac{x_1 - \mu}{\sigma} = \frac{140 - 115}{17.5} = 1.43$. For $x_2 = 140$ mm Hg: $Z_2 = \frac{x_2 - \mu}{\sigma} = \frac{140 - 137}{40} = 0.1$. While an SBP of 140 mm Hg is almost 1.5 standard deviations above the mean for a 30-year-old, it is only 0.1 standard deviations above the mean for a 65-year-old.

2.3.3 The empirical rule

The empirical rule (also known as the 68-95-99.7 rule) states that for a normal distribution, almost all observations will fall within three standard deviations of the mean. Specifically, 68% of observations are within one standard deviation of the mean, 95% are within two SD's, and 99.7% are within three SD's.

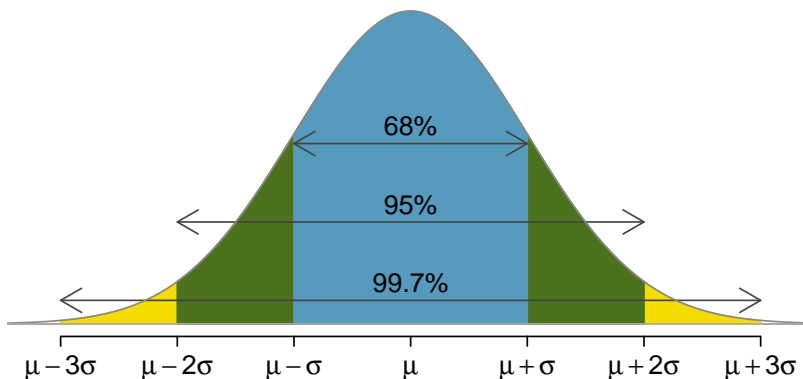


Figure 2.20: Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.

While it is possible for a normal random variable to take on values 4, 5, or even more standard deviations from the mean, these occurrences are extremely rare if the data are nearly normal. For example, the probability of being further than 4 standard deviations from the mean is about 1-in-30,000.

2.3.4 Calculating normal probabilities

The normal distribution is a continuous probability distribution. Recall from Section 2.2 that the total area under the density curve is always equal to 1, and the probability that a variable has a value within a specified interval is the area under the curve over that interval. By using either statistical software or normal probability tables, the normal model can be used to identify a probability or percentile based on the corresponding Z-score (and vice versa).



Figure 2.21: The area to the left of Z represents the percentile of the observation.

A **normal probability table** is given in Appendix B.1 on page 220 and abbreviated in Figure 2.22. This table can be used to identify the **percentile** corresponding to any particular Z-score; for instance, the percentile of $Z = 0.43$ is shown in row 0.4 and column 0.03 in Figure 2.22: 0.6664, or the 66.64th percentile. First, find the proper row in the normal probability table up through the first decimal, and then determine the column representing the second decimal value. The intersection of this row and column is the percentile of the observation. This value also rep-

Z	Second decimal place of Z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 2.22: A section of the normal probability table. The percentile for a normal random variable with $Z = 0.43$ has been **highlighted**, and the percentile closest to 0.8000 has also been **highlighted**.

resents the probability that the standard normal variable Z takes on a value of 0.43 or less; i.e. $P(Z \leq 0.43) = 0.6664$.

The table can also be used to find the Z -score associated with a percentile. For example, to identify Z for the 80th percentile, look for the value closest to 0.8000 in the middle portion of the table: 0.7995. The Z -score for the 80th percentile is given by combining the row and column Z values: 0.84.

EXAMPLE 2.45

Student A from Example 2.42 earned a score of 1800 on the SAT, which corresponds to $Z = 1$. What percentile is this score associated with?

E

In this context, the **percentile** is the percentage of people who earned a lower SAT score than Student A. From the normal table, Z of 1.00 is 0.8413. Thus, the student is in the 84th percentile of test takers. This area is shaded in Figure 2.23.

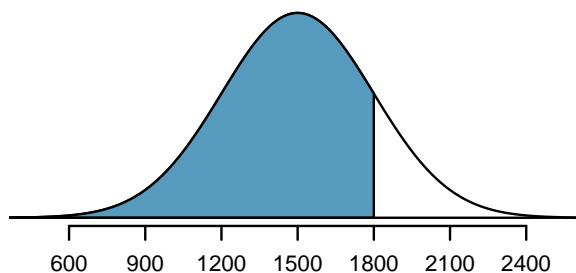


Figure 2.23: The normal model for SAT scores, with shaded area representing scores below 1800.

G

GUIDED PRACTICE 2.46

Determine the proportion of SAT test takers who scored better than Student A on the SAT.²⁷

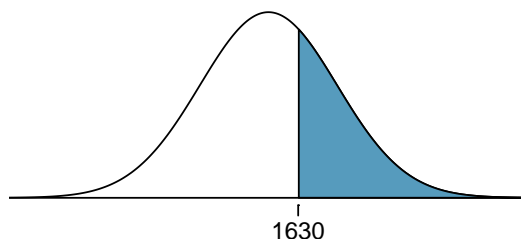
2.3.5 Normal probability examples

There are two main types of problems that involve the normal distribution: calculating probabilities from a given value (whether X or Z), or identifying the observation that corresponds to a particular probability.

EXAMPLE 2.47

Cumulative SAT scores are well-approximated by a normal model, $N(1500, 300)$. What is the probability that a randomly selected test taker scores at least 1630 on the SAT?

For any normal probability problem, it can be helpful to start out by drawing the normal curve and shading the area of interest.

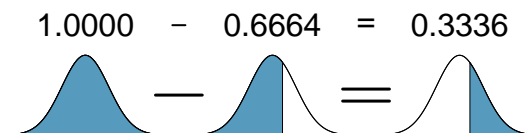


E

To find the shaded area under the curve, convert 1630 to a Z -score:

$$Z = \frac{x - \mu}{\sigma} = \frac{1630 - 1500}{300} = \frac{130}{300} = 0.43.$$

Look up the percentile of $Z = 0.43$ in the normal probability table shown in Figure 2.22 or in Appendix B.1 on page 220: 0.6664. However, note that the percentile describes those who had a Z -score *lower* than 0.43, or in other words, the area *below* 0.43. To find the area *above* $Z = 0.43$, subtract the area of the lower tail from the total area under the curve, 1:



The probability that a student scores at least 1630 on the SAT is 0.3336.

DISCRETE VERSUS CONTINUOUS PROBABILITIES

Recall that the probability of a continuous random variable equaling some exact value is always 0. As a result, for a continuous random variable X , $P(X \leq x) = P(X < x)$ and $P(X \geq x) = P(X > x)$. It is valid to state that $P(X \geq x) = 1 - P(X \leq x) = 1 - P(X < x)$.

This is *not* the case for discrete random variables. For example, for a discrete random variable Y , $P(Y \geq 2) = 1 - P(Y < 2) = 1 - P(Y \leq 1)$. It would be incorrect to claim that $P(Y \geq 2) = 1 - P(Y \leq 2)$.

²⁷If 84% had lower scores than Student A, the number of people who had better scores must be 16%.

GUIDED PRACTICE 2.48

What is the probability of a student scoring at most 1630 on the SAT?²⁸

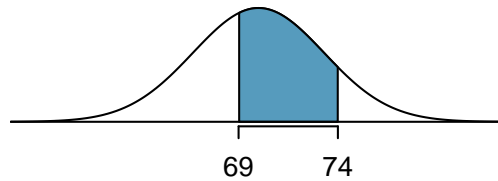
GUIDED PRACTICE 2.49

Systolic blood pressure for adults 60 years of age and older in the United States is approximately normally distributed: $N(136, 40)$. What is the probability of an adult in this age group having systolic blood pressure of 140 mm Hg or greater?²⁹

EXAMPLE 2.50

The height of adult males in the United States between the ages of 20 and 62 is nearly normal, with mean 70 inches and standard deviation 3.3 inches.³⁰ What is the probability that a random adult male is between 5'9" and 6'2"?

These heights correspond to 69 inches and 74 inches. First, draw the figure. The area of interest is an interval, rather than a tail area.

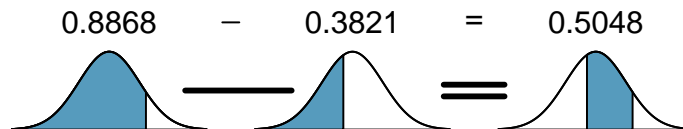


To find the middle area, find the area to the left of 74; from that area, subtract the area to the left of 69.

First, convert to Z-scores:

$$Z_{74} = \frac{x - \mu}{\sigma} = \frac{74 - 70}{3.3} = 1.21, \quad Z_{69} = \frac{x - \mu}{\sigma} = \frac{69 - 70}{3.3} = -0.30.$$

From the normal probability table, the areas are respectively, 0.8868 and 0.3821. The middle area is $0.8868 - 0.3821 = 0.5048$. The probability of being between heights 5'9" and 6'2" is 0.5048.

**GUIDED PRACTICE 2.51**

What percentage of adults in the United States ages 60 and older have blood pressure between 145 and 130 mm Hg?³¹

²⁸This probability was calculated as part of Example 2.47: 0.6664. A picture for this exercise is represented by the shaded area below "0.6664" in Example 2.47.

²⁹The Z-score for this observation was calculated in Exercise 2.44 as 0.1. From the table, this corresponds to 0.54.

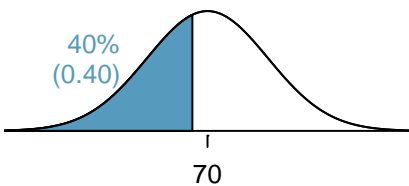
³⁰As based on a sample of 100 men, from the USDA Food Commodity Intake Database.

³¹First calculate Z-scores, then find the percent below 145 mm Hg and below 130 mm Hg: $Z_{145} = 0.23 \rightarrow 0.5890$, $Z_{130} = -0.15 \rightarrow 0.4404$ (area above). Final answer: $0.5890 - 0.4404 = 0.1486$.

EXAMPLE 2.52

How tall is a man with height in the 40th percentile?

First, draw a picture. The lower tail probability is 0.40, so the shaded area must start before the mean.



E

Determine the Z-score associated with the 40th percentile. Because the percentile is below 50%, Z will be negative. Look for the probability inside the negative part of table that is closest to 0.40: 0.40 falls in row -0.2 and between columns 0.05 and 0.06. Since it falls closer to 0.05, choose $Z = -0.25$.

Convert the Z-score to X, where $X \sim N(70, 3.3)$.

$$X = \mu + \sigma Z = 70 + (-0.25)(3.3) = 69.18.$$

A man with height in the 40th percentile is 69.18 inches tall, or about 5' 9".

GUIDED PRACTICE 2.53

G

(a) What is the 95th percentile for SAT scores? (b) What is the 97.5th percentile of the male heights?³²

³²(a) Look for 0.95 in the probability portion (middle part) of the normal probability table: row 1.6 and (about) column 0.05, i.e. $Z_{95} = 1.65$. Knowing $Z_{95} = 1.65$, $\mu = 1500$, and $\sigma = 300$, convert Z to x: $1500 + (1.65)(300) = 1995$. (b) Similarly, find $Z_{97.5} = 1.96$, and convert to x: $x_{97.5} = 76.5$ inches.

2.4 Exercises

Exercise. 2.1

Let E_1 be the events that a horse will have strongyles parasites and E_2 a horse will have piroplasmosis. Suppose $\Pr(E_1) = 0.34$ and $\Pr(E_2) = 0.54$. If $\Pr(E_1 \cap E_2) = 0.25$, what is the chance of a horse will have at least one of these diseases?

Exercise. 2.2

A team of vet professionals visited farms in the Southwest. Among the samples of 568 farms they visited, 28 of them had at least two animals with botulism. What is the estimated probability that a farm from this area has at most one animal with botulism?

Exercise. 2.3

Let B be the event that at least one rabbit in a farm with 200 rabbits as having buphthalmia, and D be the event that the diet in the farm has a Vitamin A deficit. Let $\Pr(B) = 0.15$ and $\Pr(D) = 0.24$. If B and D are statistically independent, what is the $\Pr(B \cup D)$?

Exercise. 2.4

Suppose that A is the event of a mother rabbit having a splay leg and that B is the event of a father having a splay leg. If it is known that $\Pr(A) = 0.10$, $\Pr(B) = 0.20$, and $\Pr(A \cap B) = 0.02$, are the events A and B statistically independent?

Exercise. 2.5

If Z is a standard normal random variable, find each of the following probabilities:

- | | | |
|----------------------------|-----------------------------|-----------------------------|
| i. $\Pr(0 < Z < 1.5)$ | iv. $\Pr(Z > 0.02)$ | vii. $\Pr(Z > 2)$ |
| ii. $\Pr(0.39 < Z < 2.67)$ | v. $\Pr(Z > 1)$ | |
| iii. $\Pr(Z > -2.24)$ | vi. $\Pr(-1.23 < Z < -0.3)$ | viii. $\Pr(-0.4 < Z < 1.5)$ |

Exercise. 2.6

Assume that Z is a standard normal variable. Find z_0 such that the following statement is true.

- | | |
|--------------------------------|-------------------------------------|
| i. $\Pr(Z < z_0) = 0.6368307$ | iii. $\Pr(0 < Z < z_0) = 0.4846137$ |
| ii. $\Pr(Z < z_0) = 0.9382198$ | iv. $\Pr(z_0 < Z) = 0.8925123$ |

Exercise. 2.7

If X is normally distributed variable with $\mu = 16$ and $\sigma = 4$, find each of the following probabilities:

- | | | |
|------------------|-------------------|-----------------------------|
| i. $\Pr(X < 27)$ | ii. $\Pr(X < 10)$ | iii. $\Pr(10.8 < X < 12.4)$ |
|------------------|-------------------|-----------------------------|

Exercise. 2.8

The historical data indicate that the scores of the National Board Exam Part I taken during the second year in vet school are normally distributed with the mean score of 73.1 and the variance of 16.2. In order to pass the exam one must score at least 75. Compute the percent of students that is expected to pass this year's exam.

Exercise. 2.9

The scores on a national achievement test are normally distributed with mean 500 and standard deviation 100. What percentage of those who took the test had a score between 300 and 700?³³

Exercise. 2.10

In a test given to a large group of peoples the score were normally distributed with mean 65 and standard deviation 10. What is the least whole-number score that a person could get and yet score in about the top 20%? ³⁴

Exercise. 2.11

The heights (in inches) of adults in a large population are normally distributed with $\mu = 68$ and $\sigma = 3$. What percentage of the group is under 6 feet tall? ³⁵

Exercise. 2.12

The IQs of a large population of children are normally distributed with mean 100.4 and standard deviation 11.6.

- i. What percentage of the children have IQs greater than 125?
- ii. About 90% of the children have IQs greater than what value? ³⁶

Exercise. 2.13

The weight of a one-year-old gorilla is approximately normal distributed with mean $\mu = 7$ kg and standard deviation $\sigma = 2$ kg.

- i. Calculate an interval whose midpoint is the mean and that contains 95% of the weights of one-year-old gorillas.
- ii. Calculate an interval whose midpoint is the mean and that contains 98% of the weights of one-year-old gorillas.
- iii. Calculate an interval that contains 95% of the weights of the heaviest one-year-old gorillas.
- iv. What is the percentage of one-year-old gorillas whose weight is less than 3.5 kg?
- v. What is the percentage of one-year-old gorillas whose weight is greater than 4.5 kg?
- vi. What is the percentage of one-year-old gorillas whose weight is between 6 and 8 kg?
- vii. What is the percentage of one-year-old gorillas whose weight is between 4 and 5 kg?
- viii. What is the percentile of one-year-old gorilla that weights 10.5 kg?

Exercise. 2.14

The height of an adult Appaloosa horse is approximately normal distributed with mean $\mu = 150$ cm and standard deviation $\sigma = 12$ cm.

³³E.F. Haeussler et al. *Introductory Mathematical Analysis for Business, Economics and the Life and Social Sciences + Student's Solutions Manual*. Pearson College Division, 2007. ISBN: 9780136008996. URL: <https://books.google.es/books?id=2-3xtgAACAAJ>.

³⁴Ibidem

³⁵Ibidem

³⁶Ibidem

- i. Calculate an interval whose midpoint is the mean and that contains 95% of the heights of adult Appaloosa horses.
- ii. Calculate an interval whose central point is the mean and that contains 90% of the heights of adult Appaloosa horses.
- iii. Calculate an interval that contains 95% of the heights of the shortest adult Appaloosa horses.
- iv. What is the percentage of adult Appaloosa horses whose height is less than 158 cm?
- v. What is the percentage of adult Appaloosa horses with height is greater than 141 cm?
- vi. What is the percentage of adult Appaloosa horses with height between 140 and 170 cm?
- vii. What is the percentile of adult Appaloosa horses that is 175 cm?

Chapter 3

Foundations for inference

3.1 Data collection principles

3.2 Variability in estimates

3.3 Exercises

Not surprisingly, many studies are now demonstrating the adverse effect of obesity on health outcomes. A 2017 study conducted by the consortium studying the global burden of disease estimates that high body mass index (a measure of body fat that adjusts for height and weight) may account for as many as 4.0 million deaths globally.¹ In addition to the physiologic effects of being overweight, other studies have shown that perceived weight status (feeling that one is overweight or underweight) may have a significant effect on self-esteem.^{2,3}

As stated in its mission statement, the United States Centers for Disease Control and Prevention (US CDC) "serves as the national focus for developing and applying disease prevention and control, environmental health, and health promotion and health education activities designed to improve the health of the people of the United States".⁴ Since it is not feasible to measure the health status and outcome of every single US resident, the CDC estimates features of health from samples taken from the population, via large surveys that are repeated periodically. These surveys include the National Health Interview Survey (NHIS), the National Health and Nutrition Examination Survey (NHANES), the Youth Risk Behavior Surveillance System (YRBSS) and the Behavior Risk Factor Surveillance System (BRFSS). In the language of statistics, the average weight of all US adults is a **population parameter**; the mean weight in a sample or survey is an **estimate** of population average weight. The principles of statistical inference provide not only estimates of population parameters, but also measures of uncertainty that account for the fact that different random samples will produce different estimates because of the variability of random sampling; i.e., two different random samples will not include exactly the same people.

The BRFSS was established in 1984 in 15 states to collect data using telephone interviews about health-related risk behaviors, chronic health conditions, and the use of preventive services. It now collects data in all 50 states and the District of Columbia from more than 400,000 interviews conducted each year. The data set cdc contains a small number of variables from a random sample of 20,000 responses from the 264,684 interviews from the BRFSS conducted in the year 2000. Part of this dataset is shown in Figure 3.1, with the variables described in Figure 3.2.⁵

Few studies are as large as the original BRFSS dataset (more than 250,000

¹DOI: 10.1056/NEJMoa1614362

²J Ment Health Policy Econ. 2010 Jun;13(2):53-63

³DOI: 10.1186/1471-2458-7-80

⁴<https://www.cdc.gov/maso/pdf/cdcmiss.pdf>

⁵With small modifications (character strings re-coded as factors), the data appears in this text as it does in an *OpenIntro* lab. https://www.openintro.org/go?id=statlab_r_core_intro_to_data

	case	age	gender	weight	wt Desire	height	genhlth	
	1	1	77	m	175	175	70	good
	2	2	33	f	125	115	64	good
	3	3	49	f	105	105	60	good
20000	20000	83	m	170	165	69	good	

Figure 3.1: Four cases from the cdc dataset.

Variable	Variable definition.
case	Case number in the dataset, ranging from 1 to 20,000.
age	Age in years.
gender	A factor variable, with levels m for male, f for female.
weight	Weight in pounds.
wt Desire	Weight that the respondent wishes to be, in pounds.
height	Height in inches.
genhlth	A factor variable describing general health status, with levels excellent, very good, good, fair, poor.

Figure 3.2: Some variables and their descriptions for the cdc dataset.

cases); in fact, few are as large as the 20,000 cases in the dataset cdc. The dataset cdc is large enough that estimates calculated from cdc can be thought of as essentially equivalent to the population characteristics of the entire US adult population. This chapter uses a random sample of 60 cases from cdc, stored as `cdc.samp`, to illustrate the effect of sampling variability and the ideas behind inference. In other words, suppose that cdc represents the population, and that `cdc.samp` is a sample from the population; the goal is to estimate characteristics of the population of 20,000 using only the data from the 60 individuals in the sample.



For labs, slides, and other resources, please visit
www.openintro.org/book/biostat

3.1 Data collection principles

The first step in research is to identify questions to investigate. A clearly articulated research question is essential for selecting subjects to be studied, identifying relevant variables, and determining how data should be collected.

3.1.1 Populations and samples

Consider the following research questions:

1. Do bluefin tuna from the Atlantic Ocean have particularly high levels of mercury, such that they are unsafe for human consumption?
2. For infants predisposed to developing a peanut allergy, is there evidence that introducing peanut products early in life is an effective strategy for reducing the risk of developing a peanut allergy?
3. Does a recently developed drug designed to treat glioblastoma, a form of brain cancer, appear more effective at inducing tumor shrinkage than the drug currently on the market?

Each of these questions refers to a specific target **population**. For example, in the first question, the target population consists of all bluefin tuna from the Atlantic Ocean; each individual bluefin tuna represents a case. It is almost always either too expensive or logistically impossible to collect data for every case in a population. As a result, nearly all research is based on information obtained about a sample from the population. A **sample** represents a small fraction of the population. Researchers interested in evaluating the mercury content of bluefin tuna from the Atlantic Ocean could collect a sample of 500 bluefin tuna (or some other quantity), measure the mercury content, and use the observed information to formulate an answer to the research question.



GUIDED PRACTICE 3.1

Identify the target populations for the remaining two research questions.⁶

⁶In Question 2, the target population consists of infants predisposed to developing a peanut allergy. In Question 3, the target population consists of patients with glioblastoma.

3.1.2 Anecdotal evidence

Anecdotal evidence typically refers to unusual observations that are easily recalled because of their striking characteristics. Physicians may be more likely to remember the characteristics of a single patient with an unusually good response to a drug instead of the many patients who did not respond. The dangers of drawing general conclusions from anecdotal information are obvious; no single observation should be used to draw conclusions about a population.

While it is incorrect to generalize from individual observations, unusual observations can sometimes be valuable. E.C. Heyde was a general practitioner from Vancouver who noticed that a few of his elderly patients with aortic-valve stenosis (an abnormal narrowing) caused by an accumulation of calcium had also suffered massive gastrointestinal bleeding. In 1958, he published his observation.⁷ Further research led to the identification of the underlying cause of the association, now called Heyde's Syndrome.⁸

An anecdotal observation can never be the basis for a conclusion, but may well inspire the design of a more systematic study that could be definitive.

⁷Heyde EC. Gastrointestinal bleeding in aortic stenosis. *N Engl J Med* 1958;259:196.

⁸Greenstein RJ, McElhinney AJ, Reuben D, Greenstein AJ. Co-ionic vascular ectasias and aortic stenosis: coincidence or causal relationship? *Am J Surg* 1986;151:347-51.

3.1.3 Sampling from a population

Sampling from a population, when done correctly, provides reliable information about the characteristics of a large population. The US Centers for Disease Control (US CDC) conducts several surveys to obtain information about the US population, including the Behavior Risk Factor Surveillance System (BRFSS).⁹ The BRFSS was established in 1984 to collect data about health-related risk behaviors, and now collects data from more than 400,000 telephone interviews conducted each year. Data from a recent BRFSS survey are used in Chapter 3. The CDC conducts similar surveys for diabetes, health care access, and immunization. Likewise, the World Health Organization (WHO) conducts the World Health Survey in partnership with approximately 70 countries to learn about the health of adult populations and the health systems in those countries.¹⁰

The general principle of sampling is straightforward: a sample from a population is useful for learning about a population only when the sample is **representative** of the population. In other words, the characteristics of the sample should correspond to the characteristics of the population.

Suppose that the quality improvement team at an integrated health care system, such as Harvard Pilgrim Health Care, is interested in learning about how members of the health plan perceive the quality of the services offered under the plan. A common pitfall in conducting a survey is to use a **convenience sample**, in which individuals who are easily accessible are more likely to be included in the sample than other individuals. If a sample were collected by approaching plan members visiting an outpatient clinic during a particular week, the sample would fail to enroll generally healthy members who typically do not use outpatient services or schedule routine physical examinations; this method would produce an unrepresentative sample (Figure 3.3).

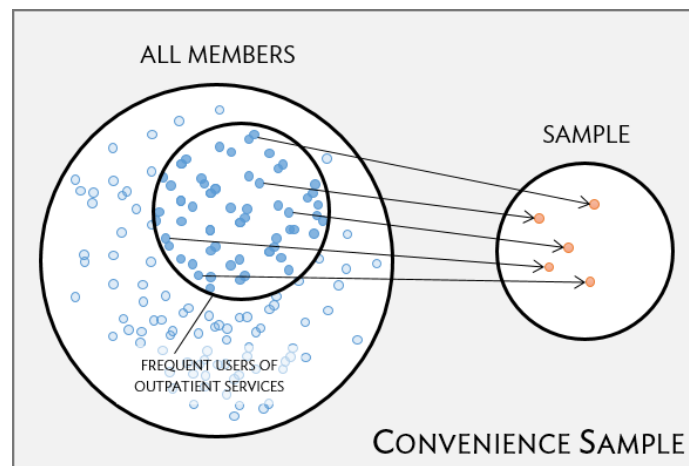


Figure 3.3: Instead of sampling from all members equally, approaching members visiting a clinic during a particular week disproportionately selects members who frequently use outpatient services.

Random sampling is the best way to ensure that a sample reflects a population. In a **simple random sample**, each member of a population has the same chance of being sampled. One way to achieve a simple random sample of the health plan members is to randomly select a certain number of names from the complete membership roster, and contact those individuals for an interview (Figure 3.4).

⁹<https://www.cdc.gov/brfss/index.html>

¹⁰<http://www.who.int/healthinfo/survey/en/>

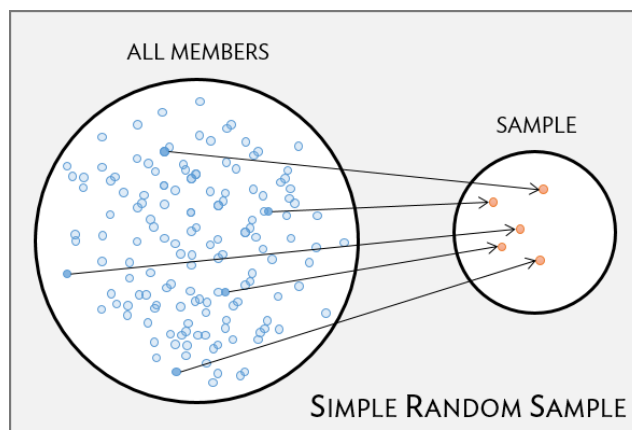


Figure 3.4: Five members are randomly selected from the population to be interviewed.

Even when a simple random sample is taken, it is not guaranteed that the sample is representative of the population. If the **non-response** rate for a survey is high, that may be indicative of a biased sample. Perhaps a majority of participants did not respond to the survey because only a certain group within the population is being reached; for example, if questions assume that participants are fluent in English, then a high non-response rate would be expected if the population largely consists of individuals who are not fluent in English (Figure 3.5). Such **non-response bias** can skew results; generalizing from an unrepresentative sample may likely lead to incorrect conclusions about a population.

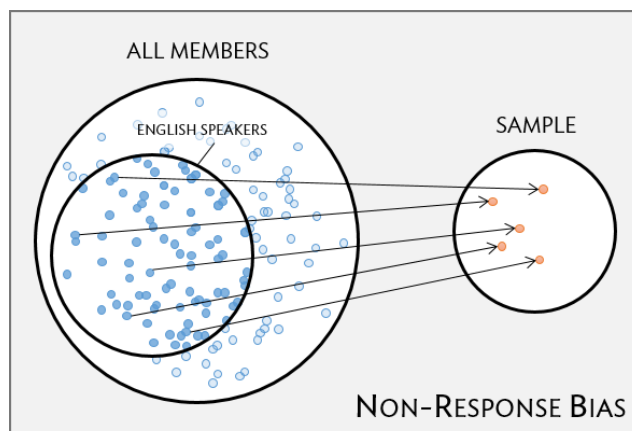


Figure 3.5: Surveys may only reach a certain group within the population, which leads to non-response bias. For example, a survey written in English may only result in responses from health plan members fluent in English.

GUIDED PRACTICE 3.2



It is increasingly common for health care facilities to follow-up a patient visit with an email providing a link to a website where patients can rate their experience. Typically, less than 50% of patients visit the website. If half of those who respond indicate a negative experience, do you think that this implies that at least 25% of patient visits are unsatisfactory?¹¹

¹¹It is unlikely that the patients who respond constitute a representative sample from the larger population of patients. This is not a random sample, because individuals are selecting themselves into a group, and it is unclear that each person has an equal chance of answering the survey. If our experience is any guide, dissatisfied people are more likely to respond to these informal surveys than satisfied patients.

3.1.4 Sampling methods

Almost all statistical methods are based on the notion of implied randomness. If data are not sampled from a population at random, these statistical methods – calculating estimates and errors associated with estimates – are not reliable. Four random sampling methods are discussed in this section: simple, stratified, cluster, and multistage sampling.

In a **simple random sample**, each case in the population has an equal chance of being included in the sample (Figure 3.6). Under simple random sampling, each case is sampled independently of the other cases; i.e., knowing that a certain case is included in the sample provides no information about which other cases have also been sampled.

In **stratified sampling**, the population is first divided into groups called **strata** before cases are selected within each stratum (typically through simple random sampling) (Figure 3.6). The strata are chosen such that similar cases are grouped together. Stratified sampling is especially useful when the cases in each stratum are very similar with respect to the outcome of interest, but cases between strata might be quite different.

Suppose that the health care provider has facilities in different cities. If the range of services offered differ by city, but all locations in a given city will offer similar services, it would be effective for the quality improvement team to use stratified sampling to identify participants for their study, where each city represents a stratum and plan members are randomly sampled from each city.

In a **cluster sample**, the population is first divided into many groups, called **clusters**. Then, a fixed number of clusters is sampled and all observations from each of those clusters are included in the sample (Figure 3.7). A **multistage sample** is similar to a cluster sample, but rather than keeping all observations in each cluster, a random sample is collected within each selected cluster (Figure 3.7).

Unlike with stratified sampling, cluster and multistage sampling are most helpful when there is high case-to-case variability within a cluster, but the clusters themselves are similar to one another. For example, if neighborhoods in a city represent clusters, cluster and multistage sampling work best when the population within each neighborhood is very diverse, but neighborhoods are relatively similar.

Applying stratified, cluster, or multistage sampling can often be more economical than only drawing random samples. However, analysis of data collected using such methods is more complicated than when using data from a simple random sample; this text will only discuss analysis methods for simple random samples.

EXAMPLE 3.3

Suppose researchers are interested in estimating the malaria rate in a densely tropical portion of rural Indonesia. There are 30 villages in the area, each more or less similar to the others. The goal is to test 150 individuals for malaria. Evaluate which sampling method should be employed.

E

A simple random sample would likely draw individuals from all 30 villages, which could make data collection extremely expensive. Stratified sampling is not advisable, since there is not enough information to determine how strata of similar individuals could be built. However, cluster sampling or multistage sampling are both reasonable options. For example, with multistage sampling, half of the villages could be randomly selected, and then 10 people selected from each village. This strategy is more efficient than a simple random sample, and can still provide a sample representative of the population of interest.

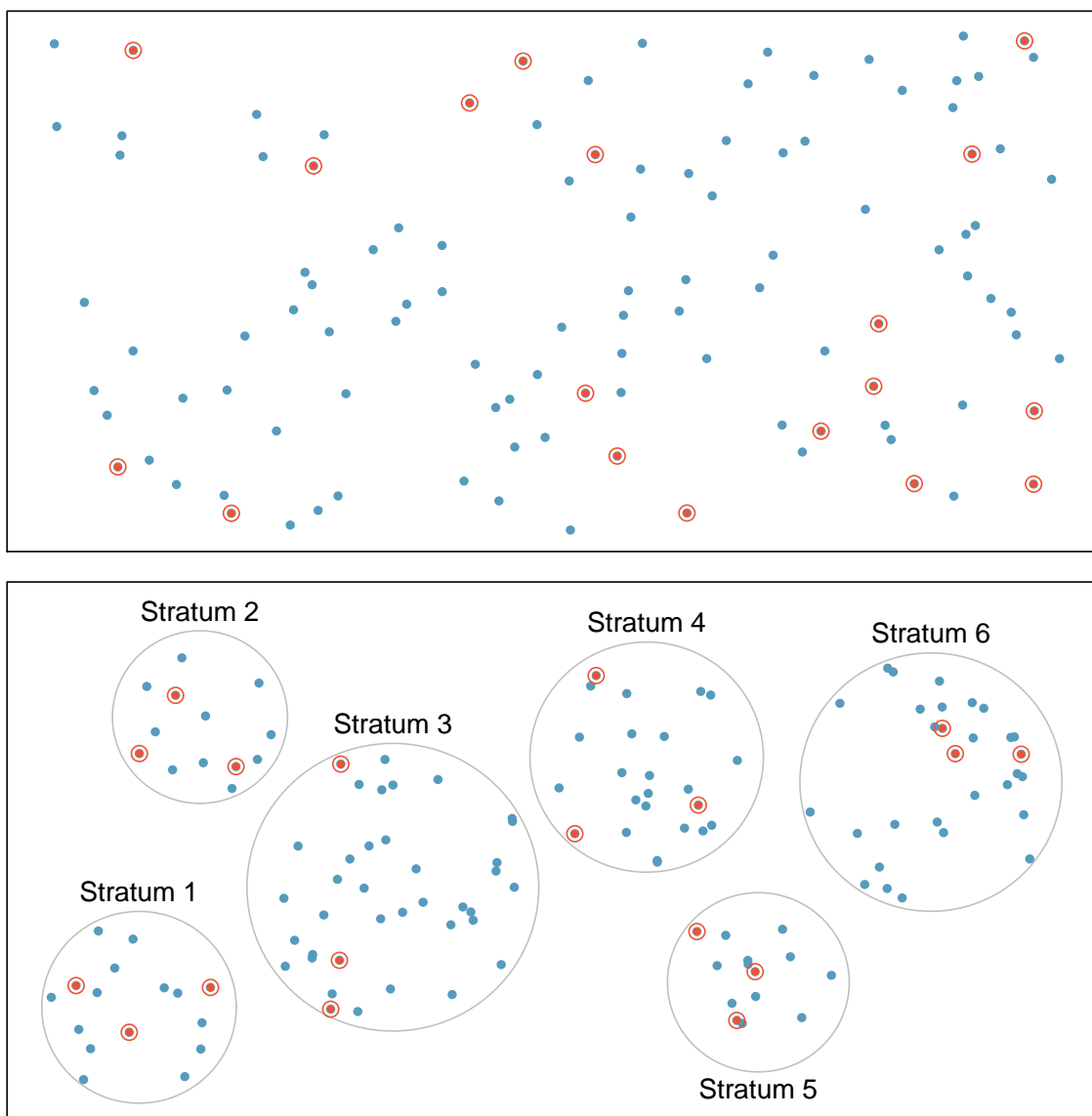


Figure 3.6: Examples of simple random and stratified sampling. In the top panel, simple random sampling is used to randomly select 18 cases (circled orange dots) out of the total population (all dots). The bottom panel illustrates stratified sampling: cases are grouped into six strata, then simple random sampling is employed within each stratum.

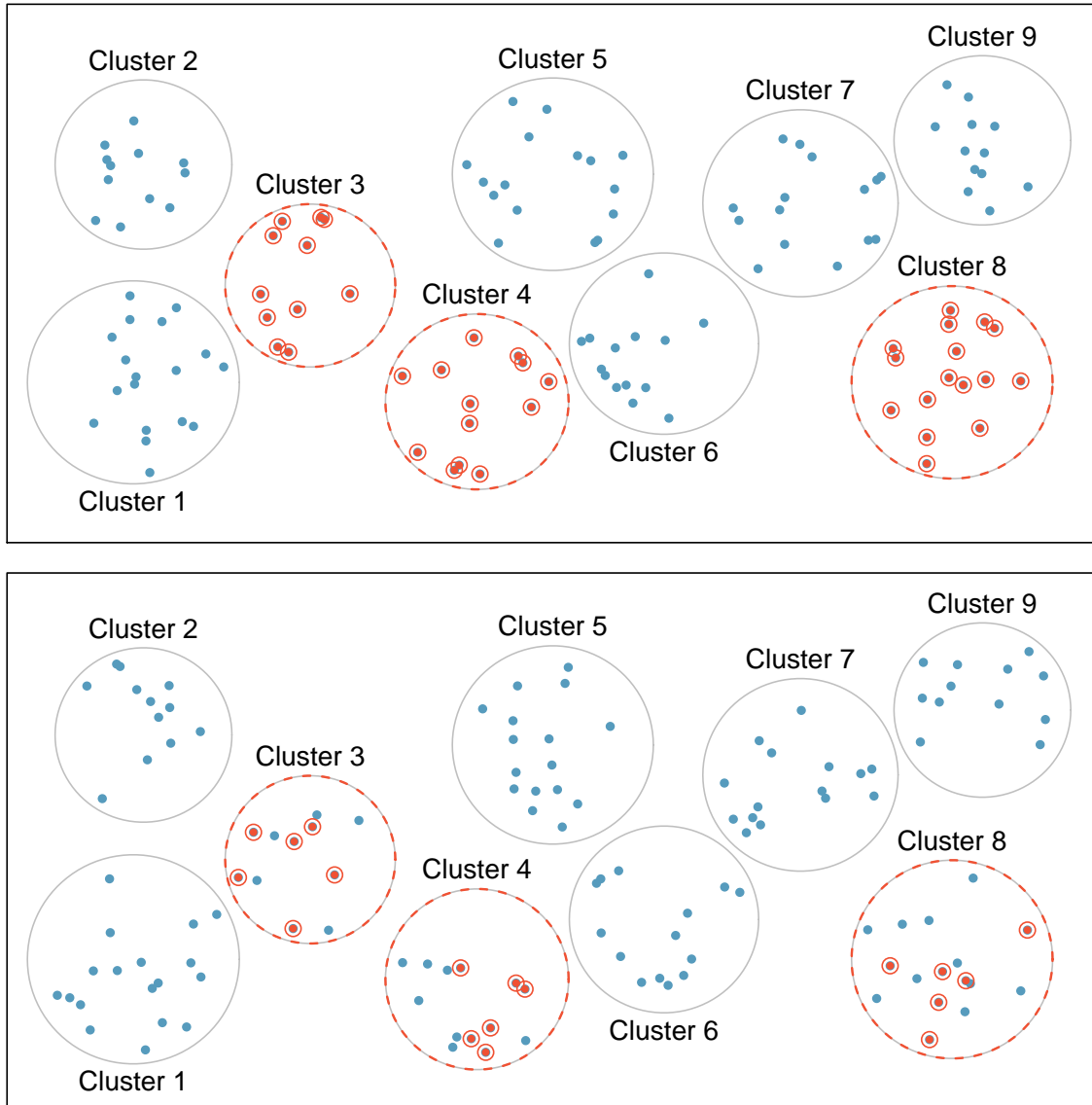


Figure 3.7: Examples of cluster and multistage sampling. The top panel illustrates cluster sampling: data are binned into nine clusters, three of which are sampled, and all observations within these clusters are sampled. The bottom panel illustrates multistage sampling, which differs from cluster sampling in that only a subset from each of the three selected clusters are sampled.

3.1.5 Introducing experiments and observational studies

The two primary types of study designs used to collect data are experiments and observational studies.

In an **experiment**, researchers directly influence how data arise, such as by assigning groups of individuals to different treatments and assessing how the outcome varies across treatment groups. The LEAP study is an example of an experiment with two groups, an experimental group that received the intervention (peanut consumption) and a control group that received a standard approach (peanut avoidance). In studies assessing effectiveness of a new drug, individuals in the control group typically receive a **placebo**, an inert substance with the appearance of the experimental intervention. The study is designed such that on average, the only difference between the individuals in the treatment groups is whether or not they consumed peanut protein. This allows for observed differences in experimental outcome to be directly attributed to the intervention and constitute evidence of a causal relationship between intervention and outcome.

In an **observational study**, researchers merely observe and record data, without interfering with how the data arise. For example, to investigate why certain diseases develop, researchers might collect data by conducting surveys, reviewing medical records, or following a **cohort** of many similar individuals. Observational studies can provide evidence of an association between variables, but cannot by themselves show a causal connection. However, there are many instances where randomized experiments are unethical, such as to explore whether lead exposure in young children is associated with cognitive impairment.

3.1.6 Experiments

Experimental design is based on three principles: control, randomization, and replication.

Control. When selecting participants for a study, researchers work to **control** for extraneous variables and choose a sample of participants that is representative of the population of interest. For example, participation in a study might be restricted to individuals who have a condition that suggests they may benefit from the intervention being tested. Infants enrolled in the LEAP study were required to be between 4 and 11 months of age, with severe eczema and/or allergies to eggs.

Randomization. Randomly assigning patients to treatment groups ensures that groups are balanced with respect to both variables that can and cannot be controlled. For example, randomization in the LEAP study ensures that the proportion of males to females is approximately the same in both groups. Additionally, perhaps some infants were more susceptible to peanut allergy because of an undetected genetic condition; under randomization, it is reasonable to assume that such infants were present in equal numbers in both groups. Randomization allows differences in outcome between the groups to be reasonably attributed to the treatment rather than inherent variability in patient characteristics, since the treatment represents the only systematic difference between the two groups.

In situations where researchers suspect that variables other than the intervention may influence the response, individuals can be first grouped into **blocks** according to a certain attribute and then randomized to treatment group within each block; this technique is referred to as **blocking** or **stratification**. The team behind the LEAP study stratified infants into two cohorts based on whether or not the child developed a red, swollen mark (a wheal) after a skin test at the time of enrollment; afterwards, infants were randomized between peanut consumption and avoidance groups. Figure 3.8 illustrates the blocking scheme used in the study.

Replication. The results of a study conducted on a larger number of cases are generally more reliable than smaller studies; observations made from a large sample are more likely to be representative of the population of interest. In a single study, **replication** is accomplished by collecting a sufficiently large sample. The LEAP study randomized a total of 640 infants.

Randomized experiments are an essential tool in research. The US Food and Drug Administration typically requires that a new drug can only be marketed after two independently conducted randomized trials confirm its safety and efficacy; the European Medicines Agency has a similar policy. Large randomized experiments in medicine have provided the basis for major public health initiatives. In 1954, approximately 750,000 children participated in a randomized study comparing polio vaccine with a placebo.¹² In the United States, the results of the study quickly led to the widespread and successful use of the vaccine for polio prevention.

¹²Meier, Paul. "The biggest public health experiment ever: the 1954 field trial of the Salk poliomyelitis vaccine." *Statistics: a guide to the unknown*. San Francisco: Holden-Day (1972): 2-13.

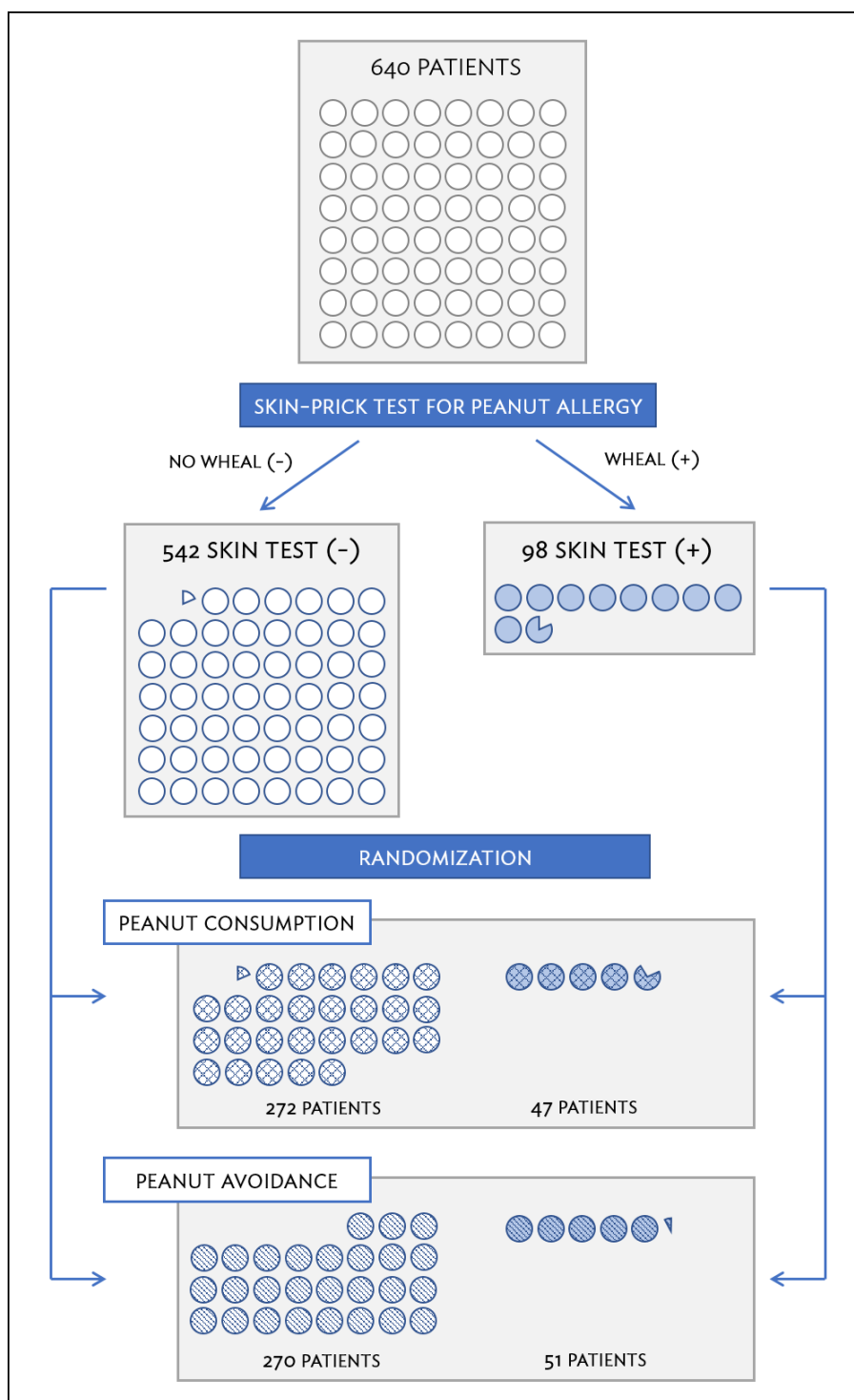
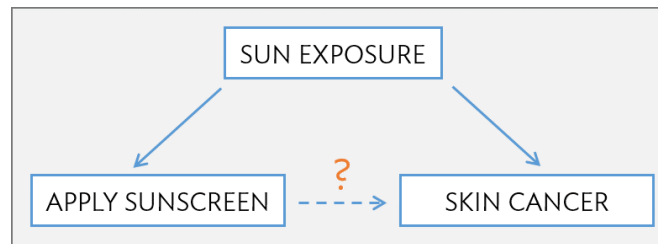


Figure 3.8: A simplified schematic of the blocking scheme used in the LEAP study, depicting 640 patients that underwent randomization. Patients are first divided into blocks based on response to the initial skin test, then each block is randomized between the avoidance and consumption groups. This strategy ensures an even representation of patients in each group who had positive and negative skin tests.

3.1.7 Observational studies

In observational studies, researchers simply observe selected potential explanatory and response variables. Participants who differ in important explanatory variables may also differ in other ways that influence response; as a result, it is not advisable to make causal conclusions about the relationship between explanatory and response variables based on observational data. For example, while observational studies of obesity have shown that obese individuals tend to die sooner than individuals with normal weight, it would be misleading to conclude that obesity causes shorter life expectancy. Instead, underlying factors are probably involved; obese individuals typically exhibit other health behaviors that influence life expectancy, such as reduced exercise or unhealthy diet.

Suppose that an observational study tracked sunscreen use and incidence of skin cancer, and found that the more sunscreen a person uses, the more likely they are to have skin cancer. These results do not mean that sunscreen causes skin cancer. One important piece of missing information is sun exposure – if someone is often exposed to sun, they are both more likely to use sunscreen and to contract skin cancer. Sun exposure is a **confounding variable**: a variable associated with both the explanatory and response variables.¹³ There is no guarantee that all confounding variables can be examined or measured; as a result, it is not advisable to draw causal conclusions from observational studies.



Confounding is not limited to observational studies. For example, consider a randomized study comparing two treatments (varenicline and bupropion) against a placebo as therapies for aiding smoking cessation.¹⁴ At the beginning of the study, participants were randomized into groups: 352 to varenicline, 329 to bupropion, and 344 to placebo. Not all participants successfully completed the assigned therapy: 259, 225, and 215 patients in each group did so, respectively. If an analysis were based only on the participants who completed therapy, this could introduce confounding; it is possible that there are underlying differences between individuals who complete the therapy and those who do not. Including all randomized participants in the final analysis maintains the original randomization scheme and controls for differences between the groups.¹⁵

GUIDED PRACTICE 3.4

As stated in Example 1.4, female body size (body.size) in the parental investment study is neither an explanatory nor a response variable. Previous research has shown that larger females tend to produce larger eggs and egg clutches; however, large body size can be costly at high altitudes. Discuss a possible reason for why the study team chose to measure female body size when it is not directly related to their main research question.¹⁶

¹³Also called a **lurking variable**, **confounding factor**, or a **confounder**.

¹⁴Jorenby, Douglas E., et al. "Efficacy of varenicline, an $\alpha 4\beta 2$ nicotinic acetylcholine receptor partial agonist, vs placebo or sustained-release bupropion for smoking cessation: a randomized controlled trial." JAMA 296.1 (2006): 56-63.

¹⁵This strategy, commonly used for analyzing clinical trial data, is referred to as an intention-to-treat analysis.

¹⁶Female body size is a potential confounding variable, since it may be associated with both the explanatory variable (altitude) and response variables (measures of maternal investment). If the study team observes, for example, that clutch size tends to decrease at higher altitudes, they should check whether the apparent association is not simply due to frogs at higher altitudes having smaller body size and thus, laying smaller clutches.

Observational studies may reveal interesting patterns or associations that can be further investigated with follow-up experiments. Several observational studies based on dietary data from different countries showed a strong association between dietary fat and breast cancer in women. These observations led to the launch of the Women's Health Initiative (WHI), a large randomized trial sponsored by the US National Institutes of Health (NIH). In the WHI, women were randomized to standard versus low fat diets, and the previously observed association was not confirmed.

Observational studies can be either prospective or retrospective. A **prospective study** identifies participants and collects information at scheduled times or as events unfold. For example, in the Nurses' Health Study, researchers recruited registered nurses beginning in 1976 and collected data through administering biennial surveys; data from the study have been used to investigate risk factors for major chronic diseases in women.¹⁷ **Retrospective studies** collect data after events have taken place, such as from medical records. Some datasets may contain both retrospectively- and prospectively-collected variables. The Cancer Care Outcomes Research and Surveillance Consortium (CanCORS) enrolled participants with lung or colorectal cancer, collected information about diagnosis, treatment, and previous health behavior, but also maintained contact with participants to gather data about long-term outcomes.¹⁸

¹⁷www.channing.harvard.edu/nhs

¹⁸Ayanian, John Z., et al. "Understanding cancer treatment and outcomes: the cancer care outcomes research and surveillance consortium." *Journal of Clinical Oncology* 22.15 (2004): 2992-2996

3.2 Variability in estimates

A natural way to estimate features of the population, such as the population mean weight, is to use the corresponding summary statistic calculated from the sample.¹⁹ The mean weight in the sample of 60 adults in `cdc.samp` is $\bar{x}_{\text{weight}} = 173.3$ lbs; this sample mean is a **point estimate** of the population mean, μ_{weight} . If a different random sample of 60 individuals were taken from `cdc`, the new sample mean would likely be different as a result of **sampling variation**. While estimates generally vary from one sample to another, the population mean is a fixed value.

GUIDED PRACTICE 3.5



How would one estimate the difference in average weight between men and women? Given that $\bar{x}_{\text{men}} = 185.1$ lbs and $\bar{x}_{\text{women}} = 162.3$ lbs, what is a good point estimate for the population difference?²⁰

Point estimates become more accurate with increasing sample size. Figure 3.9 shows the sample mean weight calculated for random samples drawn from `cdc`, where sample size increases by 1 for each draw until sample size equals 500. The red dashed horizontal line in the figure is drawn at the average weight of all adults in `cdc`, 169.7 lbs, which represents the population mean weight.²¹

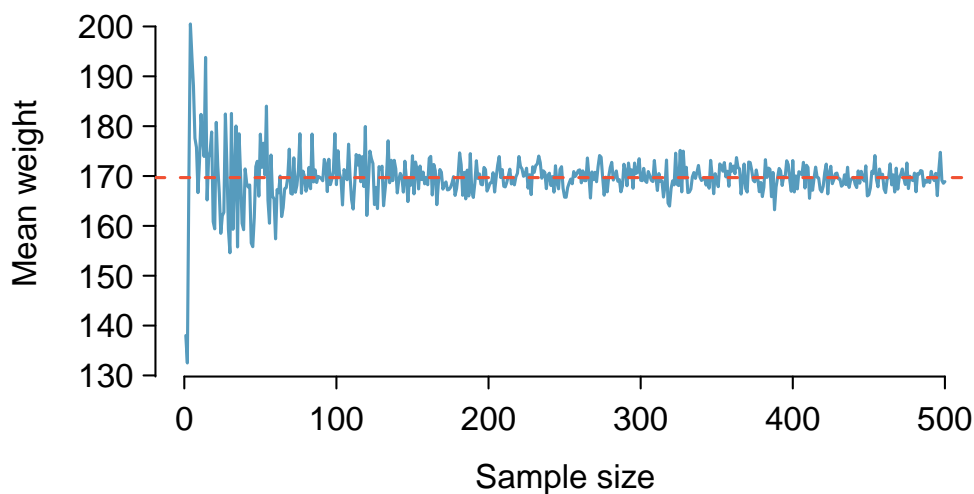


Figure 3.9: The mean weight computed for a random sample from `cdc`, increasing sample size one at a time until $n = 500$. The sample mean approaches the population mean (i.e., mean weight in `cdc`) as sample size increases.

Note how a sample size around 50 may produce a sample mean that is as much as 10 lbs higher or lower than the population mean. As sample size increases, the fluctuations around the population mean decrease; in other words, as sample size increases, the sample mean becomes less variable and provides a more reliable estimate of the population mean.

¹⁹Other population parameters, such as population median or population standard deviation, can also be estimated using sample versions.

²⁰Given that $\bar{x}_{\text{men}} = 185.1$ lbs and $\bar{x}_{\text{women}} = 162.3$ lbs, the difference of the two sample means, $185.1 - 162.3 = 22.8$ lbs, is a point estimate of the difference. The data in the random sample suggests that adult males are, on average, about 23 lbs heavier than adult females.

²¹It is not exactly the mean weight of all US adults, but will be very close since `cdc` is so large.

3.2.1 The sampling distribution for the mean

The sample mean weight calculated from `cdc.samp` is 173.3 lbs. Another random sample of 60 participants might produce a different value of \bar{x} , such as 169.5 lbs; repeated random sampling could result in additional different values, perhaps 172.1 lbs, 168.5 lbs, and so on. Each sample mean \bar{x} can be thought of as a single observation from a random variable \bar{X} . The distribution of \bar{X} is called the **sampling distribution of the sample mean**, and has its own mean and standard deviation like the random variables discussed in Chapter 3. The concept of a sampling distribution can be illustrated by taking repeated random samples from `cdc`. Figure 3.10 shows a histogram of sample means from 1,000 random samples of size 60 from `cdc`. The histogram provides an approximation of the theoretical sampling distribution of \bar{X} for samples of size 60.

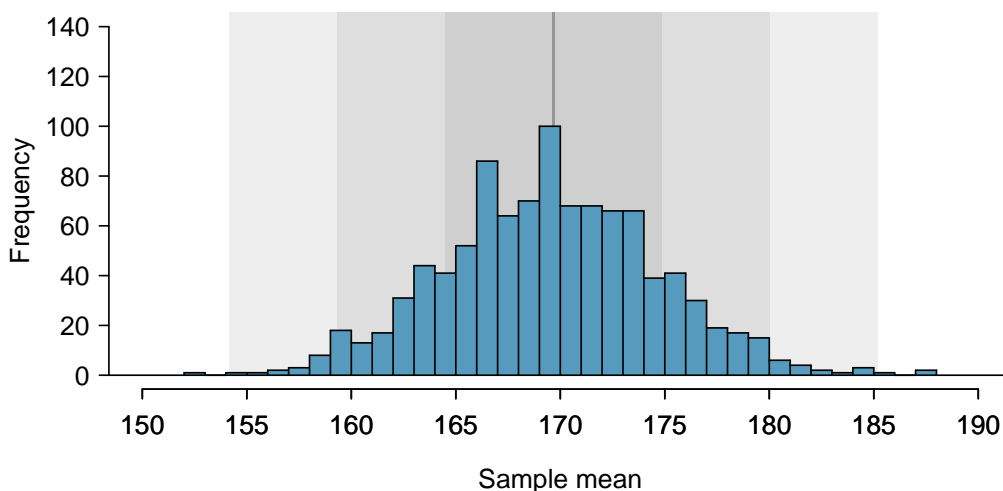


Figure 3.10: A histogram of 1000 sample means for weight among US adults, where the samples are of size $n = 60$.

SAMPLING DISTRIBUTION

The sampling distribution is the distribution of the point estimates based on samples of a fixed size from a certain population. It is useful to think of a particular point estimate as being drawn from a sampling distribution.

Since the complete sampling distribution consists of means for all possible samples of size 60, drawing a much larger number of samples provides a more accurate view of the distribution; the left panel of Figure 3.11 shows the distribution calculated from 100,000 sample means.

A normal probability plot of these sample means is shown in the right panel of Figure 3.11. All of the points closely fall around a straight line, implying that the distribution of sample means is nearly normal (see Section 2.3). This result follows from the Central Limit Theorem.

CENTRAL LIMIT THEOREM, INFORMAL DESCRIPTION

If a sample consists of at least 30 independent observations and the data are not strongly skewed, then the distribution of the sample mean is well approximated by a normal model.

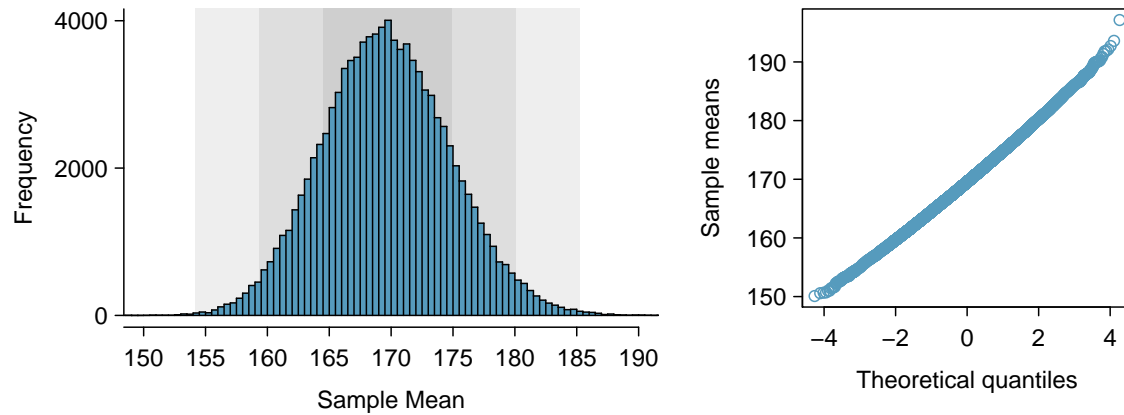


Figure 3.11: The left panel shows a histogram of the sample means for 100,000 random samples. The right panel shows a normal probability plot of those sample means.

The sampling distribution for the mean is unimodal and symmetric around the mean of the random variable \bar{X} . Statistical theory can be used to show that the mean of the sampling distribution for \bar{X} is exactly equal to the population mean μ .

However, in almost any study, conclusions about a population parameter must be drawn from the data collected from a single sample. The sampling distribution of \bar{X} is a theoretical concept, since obtaining repeated samples by conducting a study many times is not possible. In other words, it is not feasible to calculate the population mean μ by finding the mean of the sampling distribution for \bar{X} .

3.2.2 Standard error of the mean

The **standard error (SE)** of the sample mean measures the sample-to-sample variability of \bar{X} , the extent to which values of the repeated sample means oscillate around the population mean. The theoretical standard error of the sample mean is calculated by dividing the population standard deviation (σ_x) by the square root of the sample size n . Since the population standard deviation σ is typically unknown, the sample standard deviation s is often used in the definition of a standard error; s is a reasonably good estimate of σ . If \bar{X} represents the sample mean weight, its standard error (denoted by SE) is

$$SE_{\bar{X}} = \frac{s_x}{\sqrt{n}} = \frac{49.04}{\sqrt{60}} = 6.33.$$

This estimate tends to be sufficiently good when the sample size is at least 30 and the population distribution is not strongly skewed. In the case of skewed distributions, a larger sample size is necessary.

The probability tools of Section 2.2 can be used to derive the formula $\sigma_{\bar{X}} = \sigma_x/\sqrt{n}$, but the derivation is not shown here. Larger sample sizes produce sampling distributions that have lower variability. Increasing the sample size causes the distribution of \bar{X} to be clustered more tightly around the population mean μ , allowing for more accurate estimates of μ from a single sample, as shown in Figure 3.12. When sample size is large, it is more likely that any particular sample will have a mean close to the population mean.

SE
standard
error

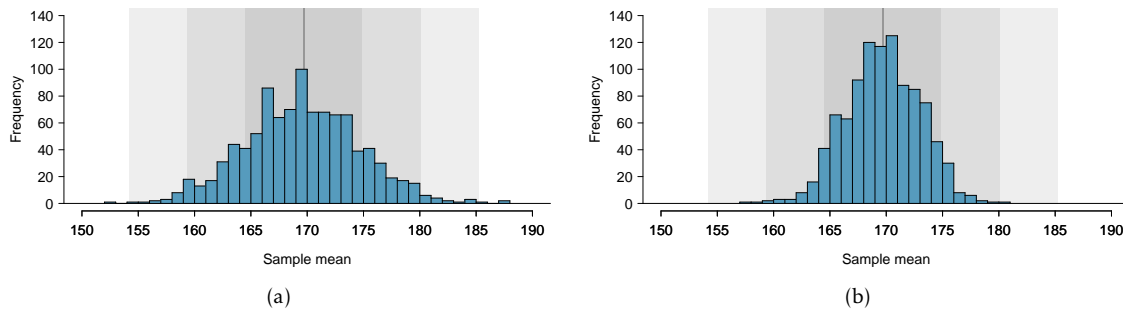


Figure 3.12: (a) Reproduced from Figure 3.10, an approximation of the sampling distribution of \bar{X} with $n = 60$. (b) An approximation of the sampling distribution of \bar{X} with $n = 200$.

THE STANDARD ERROR (SE) OF THE SAMPLE MEAN

Given n independent observations from a population with standard deviation σ , the standard error of the sample mean is equal to

$$SE_{\bar{X}} = \frac{s_x}{\sqrt{n}}. \quad (3.6)$$

This is an accurate estimate of the theoretical standard deviation of \bar{X} when the sample size is at least 30 and the population distribution is not strongly skewed.

SUMMARY: POINT ESTIMATE TERMINOLOGY

- The population mean and standard deviation are denoted by μ and σ .
- The sample mean and standard deviation are denoted by \bar{x} and s .
- The distribution of the random variable \bar{X} refers to the collection of sample means if multiple samples of the same size were repeatedly drawn from a population.
- The mean of the random variable \bar{X} equals the population mean μ . In the notation of Chapter 2, $\mu_{\bar{X}} = E(\bar{X}) = \mu$.
- The standard deviation of \bar{X} ($\sigma_{\bar{X}}$) is called the standard error (SE) of the sample mean.
- The theoretical standard error of the sample mean, as calculated from a single sample of size n , is equal to $\frac{\sigma}{\sqrt{n}}$. The standard error is abbreviated by SE and is usually estimated by using s , the sample standard deviation, such that $SE = \frac{s}{\sqrt{n}}$.

THE STANDARD ERROR (SE) OF THE PROPORTION

Given n independent observations from the population and a nominal variable with two groups (positive and negative), the standard error of the proportion of positives is equal to

$$SE_P = \sqrt{\frac{p \cdot (1 - p)}{n}} \quad (3.7)$$

where P is the proportion of positive cases in the population

If the sample size is large enough, we consider the point estimator \hat{p} , the standard error can be approximated using \hat{p} instead of the true proportion P .

3.2.3 Basic properties of point estimates

A point estimator of a parameter is said to be **unbiased** if the expected value of the point estimator is the value of the parameter. The sample mean and sample variance are unbiased estimators of the population mean and variance.

A point estimator is said to be **precise** if different samples of the population give similar values. In Figure 3.13, there are several graphical representation of unbiased and precise for a sampling distribution.

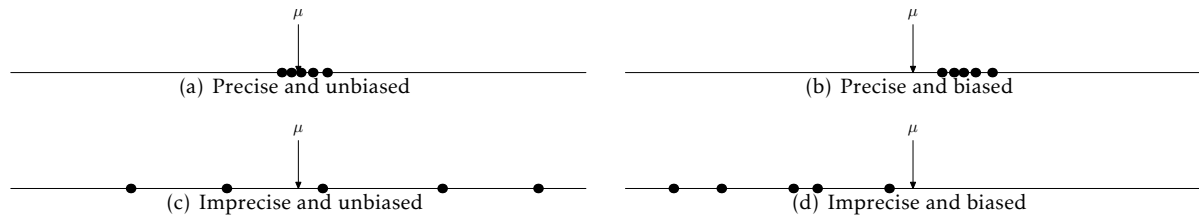
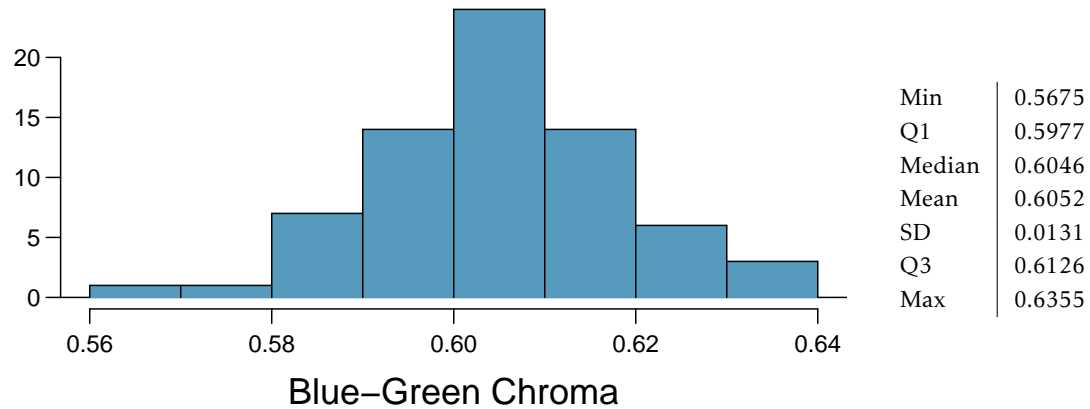


Figure 3.13: Point estimates of a parameter μ . Depending on the way the sampling distribution of the estimator approximates the parameter is said to be biased or precise

We achieved three goals in this unit. First, we determined that point estimates from a sample may be used to estimate population parameters. We also determined that these point estimates are not exact: they vary from one sample to another. Lastly, we quantified the uncertainty of the sample mean using what we call the standard error, mathematically represented in Equation (3.6).

3.3 Exercises

3.1 Egg coloration. The evolutionary role of variation in bird egg coloration remains mysterious to biologists. One hypothesis suggests that egg color may play a role in sexual selection. For example, perhaps healthier females are able to deposit more blue-green pigment into eggshells instead of using it themselves as an antioxidant. Researchers measured the blue-green chroma (BGC) of 70 different collared flycatcher nests in an area of the Czech Republic.



- What is the point estimate for the average BGC of nests?
- What is the point estimate for the standard deviation of the BGC of eggs across nests?
- Would a nest with average BGC of 0.63 be considered unusually high? Explain your reasoning.
- Compute the standard error of the sample mean using the summary statistics.

Chapter 4

Confidence Intervals

4.1 Confidence intervals

4.2 Single-sample inference with the t -distribution

4.3 Confidence interval for a single proportion

4.4 Exercises

A point estimate provides a single plausible value for a parameter. However, a point estimate is rarely perfect; usually there is some error in the estimate. Instead of supplying just a point estimate of a parameter, a next logical step would be to provide a plausible *range of values* for the parameter.

A plausible range of values for the population parameter is called a **confidence interval**.

Using only a point estimate is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net. We can throw a spear where we saw a fish, but we will probably miss. On the other hand, if we toss a net in that area, we have a good chance of catching the fish.

If we report a point estimate, we probably will not hit the exact population parameter. On the other hand, if we report a range of plausible values – a confidence interval – we have a good shot at capturing the parameter.



For labs, slides, and other resources, please visit
www.openintro.org/book/biostat

4.1 Confidence intervals

4.1.1 Interval estimates for a population parameter

While a point estimate consists of a single value, an interval estimate provides a plausible range of values for a parameter. When estimating a population mean μ , a **confidence interval** for μ has the general form

$$(\bar{x} - m, \bar{x} + m) = \bar{x} \pm m,$$

where m is the **margin of error**. Intervals that have this form are called **two-sided confidence intervals** because they provide both lower and upper bounds, $\bar{x} - m$ and $\bar{x} + m$, respectively.

The standard error of the sample mean is the standard deviation of its distribution; additionally, the distribution of sample means is nearly normal and centered at μ . Under the normal model, the sample mean \bar{x} will be within 1.96 standard errors (i.e., standard deviations) of the population mean μ approximately 95% of the time.¹ Thus, if an interval is constructed that spans 1.96 standard errors from the point estimate in either direction, a data analyst can be 95% **confident** that the interval

$$\bar{x} \pm 1.96 \times \text{SE} \quad (4.1)$$

contains the population mean. The value 95% is an approximation, accurate when the sampling distribution for the sample mean is close to a normal distribution. This assumption holds when the sample size is sufficiently large.

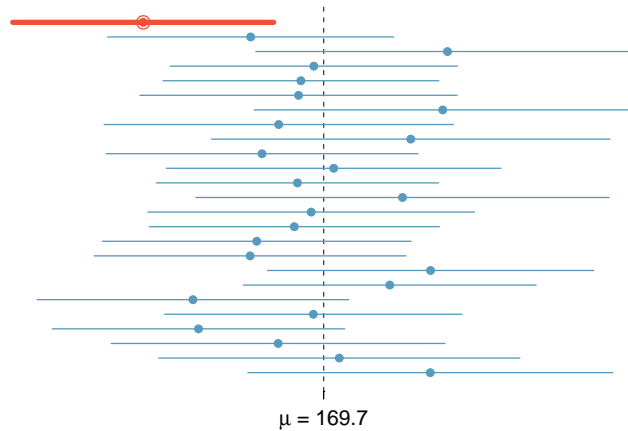


Figure 4.1: Twenty-five samples of size $n = 60$ were taken from cdc. For each sample, a 95% confidence interval was calculated for the population average adult weight. Only 1 of these 25 intervals did not contain the population mean, $\mu = 169.7$ lbs.

The phrase "95% confident" has a subtle interpretation: if many samples were drawn from a population, and a confidence interval is calculated from each one using Equation 4.1, about 95% of those intervals would contain the population mean μ . Figure 4.1 illustrates this process with 25 samples taken from cdc. Of the 25 samples, 24 contain the mean weight in cdc of 169.7 lbs, while one does not.

Just as with the sampling distribution of the sample mean, the interpretation of a confidence

¹In other words, the Z-score of 1.96 is associated with 2.5% area to the right (and $Z = -1.96$ has 2.5% area to the left); this can be found on normal probability tables or from using statistical software.

interval relies on the abstract construct of repeated sampling. A data analyst, who can only observe one sample, does not know whether the population mean lies within the single interval calculated. The uncertainty is due to random sampling—by chance, it is possible to select a sample from the population that has unusually high (or low) values, resulting in a sample mean \bar{x} that is relatively far from μ , and by extension, a confidence interval that does not contain μ .

EXAMPLE 4.2

The sample mean adult weight from the 60 observations in `cdc.samp` is $\bar{x}_{\text{weight}} = 173.3$ lbs, and the standard deviation is $s_{\text{weight}} = 49.04$ lbs. Use Equation 4.1 to calculate an approximate 95% confidence interval for the average adult weight in the US population.

The standard error for the sample mean is $SE_{\bar{x}} = \frac{49.04}{\sqrt{60}} = 6.33$ lbs. The 95% confidence interval is

$$\bar{x}_{\text{weight}} \pm 1.96SE_{\bar{x}} = 173.3 \pm (1.96)(6.33) = (160.89, 185.71) \text{ lbs.}$$

The data support the conclusion that, with 95% confidence, the average weight of US adults is between approximately 161 and 186 lbs.

Figure 3.11 visually shows that the sampling distribution is nearly normal. To assess normality of the sampling distribution without repeated sampling, it is necessary to check whether the data are skewed. Although Figure 4.2 shows some skewing, the sample size is large enough that the confidence interval should be reasonably accurate.

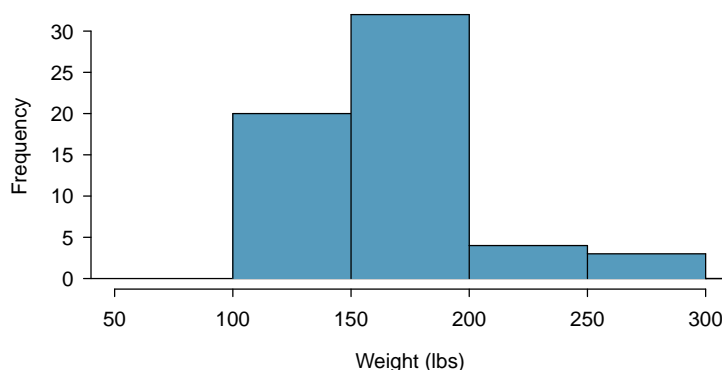


Figure 4.2: Histogram of weight in `cdc.samp`

GUIDED PRACTICE 4.3

There are 31 females in the sample of 60 US adults, and the average and standard deviation of weight for these individuals are 162.3 lbs and 57.74 lbs, respectively. A histogram of weight for the 31 females is shown in Figure 4.3. Calculate an approximate 95% confidence interval for the average weight of US females. Is the interval likely to be accurate?²

²Applying Equation 4.1: $162.3 \pm (1.96)(57.74/\sqrt{31}) \rightarrow (149.85, 174.67)$. The usual interpretation would be that a data analyst can be about 95% confident the average weight of US females is between approximately 150 and 175 lbs. However, the histogram of female weights shows substantial right skewing, and several females with recorded weights larger than 200 lbs. The confidence interval is probably not accurate; a larger sample should be collected in order for the sampling distribution of the mean to be approximately normal. Next section will introduce the *t*-distribution, which is more reliable with small sample sizes than the *z*-distribution.

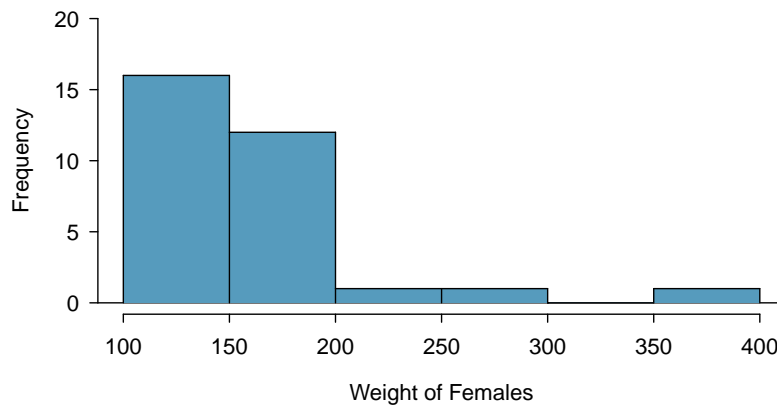


Figure 4.3: Histogram of weight for the 31 females in cdc.samp.

4.1.2 Changing the confidence level

Ninety-five percent confidence intervals are the most commonly used interval estimates, but intervals with confidence levels other than 95% can also be constructed. The general formula for a confidence interval (for the population mean μ) is given by

$$\bar{x} \pm z^* \times SE, \quad (4.4)$$

where z^* is chosen according to the confidence level. When calculating a 95% confidence level, z^* is 1.96, since the area within 1.96 standard deviations of the mean captures 95% of the distribution.

To construct a 99% confidence interval, z^* must be chosen such that 99% of the normal curve is captured between $-z^*$ and z^* .

EXAMPLE 4.5

Let Y be a normally distributed random variable. Ninety-nine percent of the time, Y will be within how many standard deviations of the mean?

(E)

This is equivalent to the z -score with 0.005 area to the right of z and 0.005 to the left of $-z$. In the normal probability table, this is the z -value that with 0.005 area to its right and 0.995 area to its left. The closest two values are 2.57 and 2.58; for convenience, round up to 2.58. The unobserved random variable Y will be within 2.58 standard deviations of μ 99% of the time, as shown in Figure 4.4.

A 99% confidence interval will have the form

$$\bar{x} \pm 2.58 \times SE, \quad (4.6)$$

and will consequently be wider than a 95% interval for μ calculated from the same data, since the margin of error m is larger.

EXAMPLE 4.7

Create a 99% confidence interval for the average adult weight in the US population using the data in cdc.samp. The point estimate is $\bar{x}_{weight} = 173.3$ and the standard error is $SE_{\bar{x}} = 6.33$.

(E)

Apply the 99% confidence interval formula: $\bar{x}_{weight} \pm 2.58 \times SE_{\bar{x}} \rightarrow (156.97, 189.63)$. A data analyst can be 99% confident that the average adult weight is between 156.97 and 189.63 lbs.

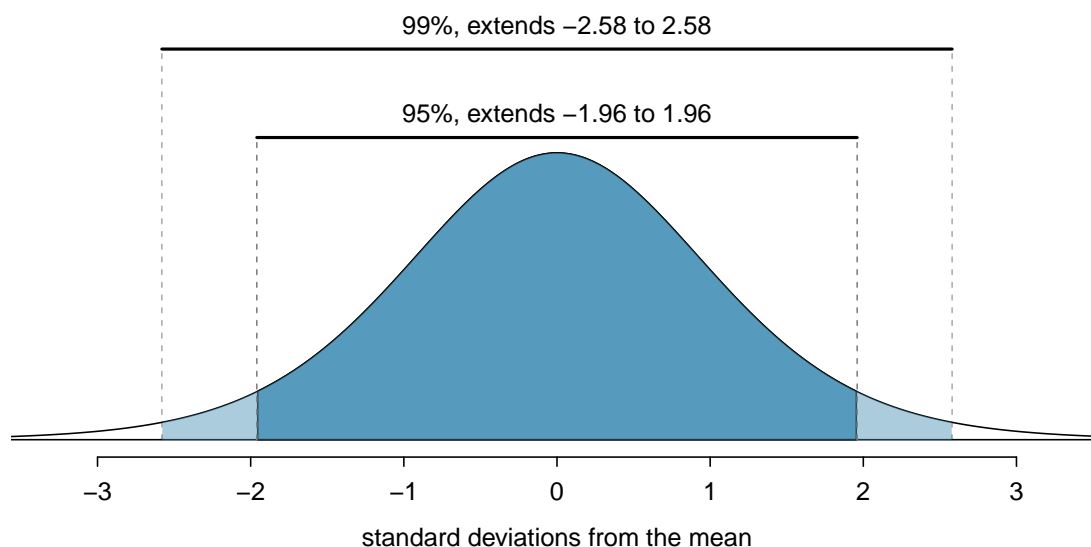


Figure 4.4: The area between $-z^*$ and z^* increases as $|z^*|$ becomes larger. If the confidence level is 99%, z^* is chosen such that 99% of the normal curve is between $-z^*$ and z^* , which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: $z^* = 2.58$.

The 95% confidence interval for the average adult weight is (160.89, 185.71) lbs. Increasing the confidence level to 99% results in the interval (156.97, 189.63) lbs; this wider interval is more likely to contain the population mean μ . However, increasing the confidence level comes at a cost: a wider interval is less informative in providing a precise estimate of the population mean. Consider the extreme: to be "100% confident" that an interval contains μ , the interval must span all possible values of μ . For example, with 100% confidence the average weight is between 0 and 1000 lbs; while this interval necessarily contains μ , it has no interpretive value and is completely uninformative.³

Decreasing the confidence level produces a narrower interval; the estimate is more precise, but also more prone to inaccuracy. For example, consider a 50% confidence interval for average adult weight using `cdc.samp`: the z^* value is 0.67, and the confidence interval is (169.06, 177.54) lbs. This interval provides a more precise estimate of the population average weight μ than the 99% or 95% confidence intervals, but the increased precision comes with less confidence about whether the interval contains μ . In a theoretical setting of repeated sampling, if 100 50% confidence intervals were computed, only half could be expected to contain μ .

The choice of confidence level is a trade-off between obtaining a precise estimate and calculating an interval that can be reasonably expected to contain the population parameter. In published literature, the most used confidence intervals are the 90%, 95%, and 99%.

4.1.3 Interpreting confidence intervals

The correct interpretation of an XX% confidence interval is, "We are XX% confident that the population parameter is between ..." While it may be tempting to say that a confidence interval captures the population parameter with a certain probability, this is a common error. The confidence level only quantifies how plausible it is that the parameter is within the interval; there is no probability associated with whether a parameter is contained in a specific confidence interval. The

³Strictly speaking, to be 100% confident requires an interval spanning all positive numbers; 1000 lbs has been arbitrarily chosen as an upper limit for human weight.

confidence coefficient reflects the nature of a procedure that is correct XX% of the time, given that the assumptions behind the calculations are true.

The conditions regarding the validity of the normal approximation can be checked using the numerical and graphical summaries discussed in Chapter 1. However, the condition that data should be from a random sample is sometimes overlooked. If the data are not from a random sample, then the confidence interval no longer has interpretive value, since there is no population mean to which the confidence interval applies. For example, while only simple arithmetic is needed to calculate a confidence interval for BMI from the `famuss` dataset in Chapter 1, the participants in the study are almost certainly not a random sample from some population; thus, a confidence interval should not be calculated in this setting.

EXAMPLE 4.8

Body mass index (BMI) is one measure of body weight that adjusts for height. The National Health and Nutrition Examination Survey (NHANES) consists of a set of surveys and measurements conducted by the US CDC to assess the health and nutritional status of adults and children in the United States. The dataset `nhanes.samp` contains 76 variables and is a random sample of 200 individuals from the measurements collected in the years 2009-2010 and 2012-2013.⁴ Use `nhanes.samp` to calculate a 95% confidence interval for adult BMI in the US population, and assess whether the data suggest Americans tend to be overweight.

In the random sample of 200 participants, BMI is available for all 135 of the participants that are 21 years of age or older. As shown in the histogram (Figure 4.5), the data are right-skewed, with one large outlier. The outlier corresponds to an implausibly extreme BMI value of 69.0; since it seems likely that the value represents an error from when the data was recorded, this data point is excluded from the following analysis.

The mean and standard deviation in this sample of 134 are 28.8 and 6.7 kg/meter², respectively. The sample size is large enough to justify using the normal approximation when computing the confidence interval. The standard error of the mean is $SE = 6.7/\sqrt{134} = 0.58$, so the 95% confidence interval is given by

$$\begin{aligned}\bar{x}_{\text{BMI}} \pm (1.96)(SE) &= 28.8 \pm (1.96)(0.58) \\ &= (27.7, 29.9).\end{aligned}$$

Based on this sample, a data analyst can be 95% confident that the average BMI of US adults is between 27.7 and 29.9 kg/m².

The World Health Organization (WHO) and other agencies use BMI to set normative guidelines for body weight. The current guidelines are shown in Figure 4.6.

The confidence interval (27.7, 29.9) kg/m² certainly suggests that the average BMI in the US population is higher than 21.7, the middle of the range for normal BMIs, and even higher than 24.99, the upper limit of the normal weight category. These data indicate that Americans tend to be overweight.

⁴The sample was drawn from a larger sample of 20,293 participants in the **NHANES** package, available from The Comprehensive R Archive Network (CRAN). The CDC uses a complex sampling design that samples some demographic subgroups with larger probabilities, but `nhanes.samp` has been adjusted so that it can be viewed as a random sample of the US population.

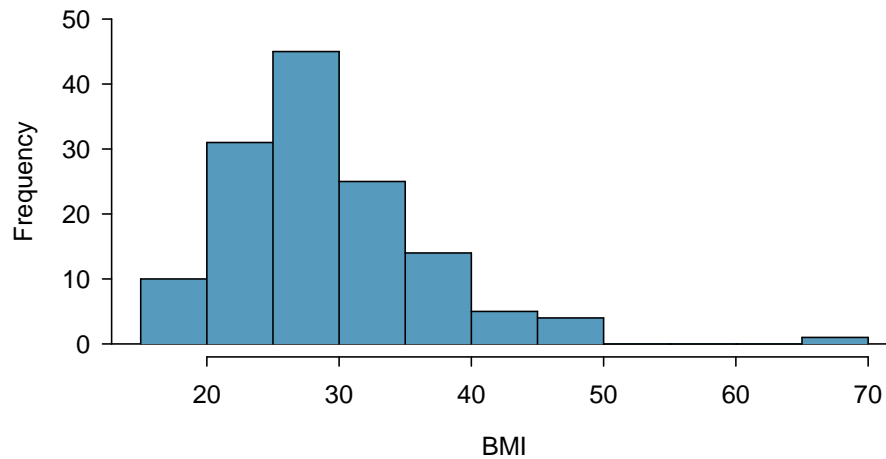


Figure 4.5: The distribution of BMI for the 135 adults in `nhanes.samp`.

Category	BMI range
Underweight	< 18.50
Normal (healthy weight)	18.5-24.99
Overweight	≥ 25
Obese	≥ 30

Figure 4.6: WHO body weight categories based on BMI.

4.1.4 Sample size calculation to estimate population mean

Before starting to sample the population to estimate a parameter such as the mean, it is convenient to know how many individuals you need to achieve an accurate estimation with a confidence that the error is lower than a given threshold. Following the previous estimations given in this chapter, the error is the difference between the sample mean (point estimator) and the population mean (true parameter). Confidence interval whose mid point is the sample mean is likely to include the population mean with the given probability $1 - \alpha$. Consequently, the error is going to be less than $z_{1-\alpha/2}\sigma/\sqrt{n}$ with that probability.

If the error of your estimation should be lower than a given threshold or maximum error, then sample size n must be taken large enough such that $z_{1-\alpha/2}\sigma/\sqrt{n}$ is lower than max. error. Using elementary algebraic reasoning, n should be larger than $z_{1-\alpha/2}^2\sigma^2/\text{max. error}^2$.

Sample size calculation of μ

The sample size needed to estimate the population mean, μ with a maximum error max.error at a confidence $100(1 - \alpha)$ and a population variance σ^2 is $n > z_{1-\alpha/2}^2 \frac{\sigma^2}{\text{max.error}^2}$.

● Example 4.9

Calculate the sample size you need to estimate the height of male senior in high schools with a maximum error of 0.1 inches at a confidence of 95%. We assume that the population standard deviation is $\sigma = 3$ inches

$$n > z_{1-\alpha/2}^2 \frac{\sigma^2}{\text{max. error}^2} = 1.96^2 \frac{3^2}{0.1^2} = 3457.44$$

Therefore, you should take at least 3458 individuals for your sample.

4.2 Confidence interval with the t -distribution

The tools studied in the previous section all made use of the t -statistic from a sample mean,

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}},$$

where the parameter μ is a population mean, \bar{x} and s are the sample mean and standard deviation, and n is the sample size. Tests and confidence intervals were restricted to samples of at least 30 independent observations from a population where there was no evidence of strong skewness. This allowed for the Central Limit Theorem to be applied, justifying use of the normal distribution to calculate probabilities associated with the t -statistic.

In sample sizes smaller than 30, if the data are approximately symmetric and there are no large outliers, the t -statistic has what is called a t -distribution. When the normal distribution is used as the sampling distribution of the t -statistic, s is essentially being treated as a good replacement for the unknown population standard deviation σ . However, the sample standard deviation s , as an estimate of σ , has its own inherent variability like \bar{x} . The t density function adjusts for the variability in s by having more probability in the left and right tails than the normal distribution.

4.2.1 The t -distribution

Figure 4.7 shows a t -distribution and normal distribution. Like the standard normal distribution, the t -distribution is unimodal and symmetric about zero. However, the tails of a t -distribution are thicker than for the normal, so observations are more likely to fall beyond two standard deviations from the mean than under the normal distribution.⁵ While the estimate of the standard error will be less accurate with smaller sample sizes, the thick tails of the t -distribution correct for the variability in s .

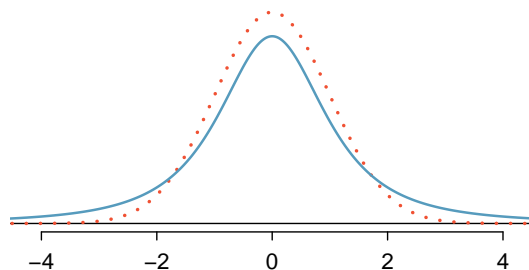


Figure 4.7: Comparison of a t -distribution (solid line) and a normal distribution (dotted line).

The t -distribution can be described as a family of symmetric distributions with a single parameter: degrees of freedom, which equals $n - 1$. Several t -distributions are shown in Figure 4.8. When there are more degrees of freedom, the t -distribution looks very much like the standard normal distribution. With degrees of freedom of 30 or more, the t -distribution is nearly indistinguishable from the normal distribution. Since the t -statistics in Chapter 3 were associated with sample sizes of at least 30, the degrees of freedom for the corresponding t -distributions were large enough to justify use of the normal distribution to calculate probabilities.

⁵The standard deviation of the t -distribution is actually a little more than 1. However, it is useful to think of the t -distribution as having a standard deviation of 1 in the context of using it to conduct inference.

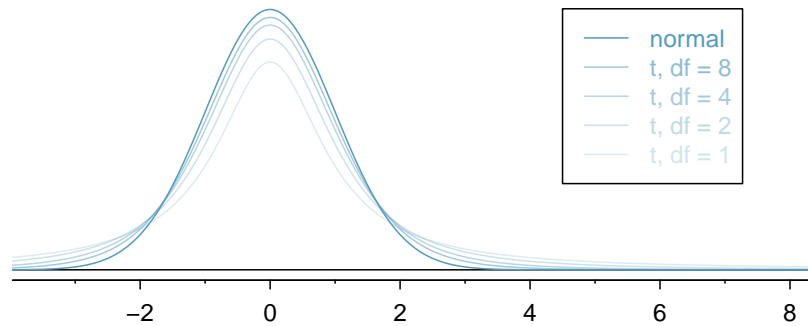


Figure 4.8: The larger the degrees of freedom, the more closely the t -distribution resembles the standard normal model.

DEGREES OF FREEDOM (DF)

The degrees of freedom characterize the shape of the t -distribution. The larger the degrees of freedom, the more closely the distribution approximates the normal model.

Probabilities for the t -distribution can be calculated either by using distribution tables or using statistical software. The use of software has become the preferred method because it is more accurate, allows for complete flexibility in the choice of t -values on the horizontal axis, and is not limited to a small range of degrees of freedom. The remainder of this section illustrates the use of a **t -table**, partially shown in Figure 4.9, in place of the normal probability table. A larger t -table is in Appendix B.2 on page 221. The R labs illustrate the use of software to calculate probabilities for the t -distribution. Readers intending to use software can skip to the next section.

df/p	0.9	0.95	0.975	0.99	0.995
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
⋮	⋮	⋮	⋮	⋮	⋮
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
⋮	⋮	⋮	⋮	⋮	⋮
400	1.28	1.65	1.97	2.34	2.59
500	1.28	1.65	1.96	2.33	2.59
∞	1.28	1.64	1.96	2.33	2.58

Figure 4.9: An abbreviated look at the t -table. Each row represents a different t -distribution. The columns describe the cutoffs for specific tail areas. The row with $df = 18$ has been highlighted.

Each row in the t -table represents a t -distribution with different degrees of freedom. The columns correspond to percentiles. For instance, for a t -distribution with $df = 18$, row 18 is used (highlighted in Figure 4.9). The value in this row that identifies the cutoff for an upper tail of 5% is found in the column where *one tail* is 0.950. This cutoff is 1.73. The cutoff for the lower 5% is -1.73; just like the normal distribution, all t -distributions are symmetric.

EXAMPLE 4.10

What proportion of the t -distribution with 18 degrees of freedom falls below -2.10?

E

Just like a normal probability problem, we first draw the picture in Figure 4.10 and shade the area below -2.10. The area on the left of -2.10 is the same area that on the right of 2.10. In the t distribution are the percentiles corresponding to positive values. To find this area, we identify the appropriate row: $df = 18$. Then we identify the column containing the absolute value of -2.10; a very similar value is at the last but two column and the corresponding quantile is 0.975. Hence, on the left of 2.10 the area is 0.975 and, consequently, the area on the right is 0.025. About 2.5% of the distribution falls below -2.10. In the next example we encounter a case where the exact t value is not listed in the table.

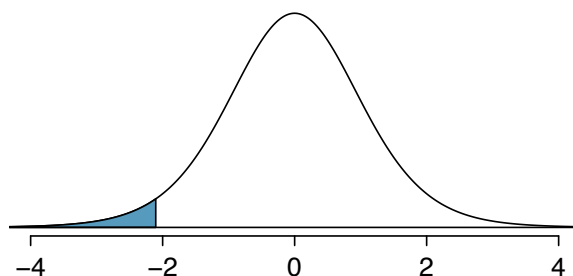


Figure 4.10: The t -distribution with 18 degrees of freedom. The area below -2.10 has been shaded.

EXAMPLE 4.11

A t -distribution with 20 degrees of freedom is shown in the left panel of Figure 4.11. Estimate the proportion of the distribution falling above 1.65 and below -1.65.

E

We identify the row in the t table using the degrees of freedom: $df = 20$. Then we look for 1.65; it is not listed. It falls between the columns corresponding to 0.900 and 0.9500. Since these values bound 1.65, their tail areas will bound the tail area corresponding to 1.65. These are the values of area on the left but the question is about values on the right. So, the area on the right is between 0.050 and 0.10, and we conclude that between 5% and 10% of the distribution is more than 1.65 standard deviations above the mean. If we like, we can identify the precise area using statistical software: 0.0573.

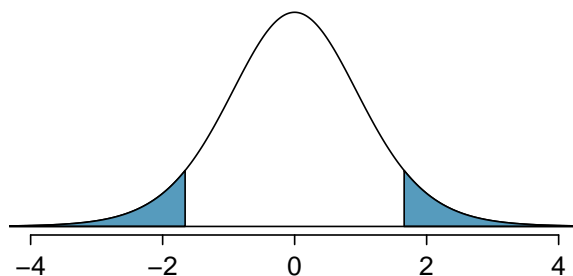


Figure 4.11: The t -distribution with 20 degrees of freedom, with the area further than 1.65 away from 0 shaded.

4.2.2 Using the t -distribution for tests and confidence intervals for a population mean

Section 4.1 provided formulas for confidence intervals for population means in random samples large enough for the t -statistic to have a nearly normal distribution. In samples smaller than 30 from approximately symmetric distributions without large outliers, the t -statistic has a t -distribution with degrees of freedom equal to $n - 1$. Just like inference in larger samples, inference using the t -distribution also requires that the observations in the sample be independent. Random samples from very large populations always produce independent observations; in smaller populations, observations will be approximately independent as long as the size of the sample is no larger than 10% of the population.

In summary, to proceed with the t distribution for inference about a single mean, we must check two conditions.

Independence of observations. We verify this condition just as we did before. We collect a simple random sample from less than 10% of the population, or if it was an experiment or random process, we carefully check to the best of our abilities that the observations were independent.

Observations come from a nearly normal distribution. This second condition is difficult to verify with small data sets. We often (i) take a look at a plot of the data for obvious departures from the normal model, and (ii) consider whether any previous experiences alert us that the data may not be nearly normal.

Formulas for tests and intervals using the t -distribution are very similar to those using the normal distribution. For a sample of size n with sample mean \bar{x} and standard deviation s , two-sided confidence intervals with confidence coefficient $100(1 - \alpha)\%$ have the form

$$\bar{x} \pm t_{df}^* \times SE,$$

where SE is the standard error of the sample mean (s/\sqrt{n}) and t_{df}^* is the point on a t -distribution with $n - 1$ degrees of freedom and area $(1 - \alpha/2)$ to its left.

A one-sided interval with the same confidence coefficient will have the form

$$\begin{aligned} &\bar{x} + t_{df}^* \times SE \text{ (one-sided upper confidence interval), or} \\ &\bar{x} - t_{df}^* \times SE \text{ (one-sided lower confidence interval),} \end{aligned}$$

except that in this case t_{df}^* is the point on a t -distribution with $n - 1$ degrees of freedom and area $(1 - \alpha)$ to its left.

With the ability to conveniently calculate t^* for any sample size or associated α via computing software, the t -distribution can be used by default over the normal distribution. The rule of thumb that $n > 30$ qualifies as a large enough sample size to use the normal distribution dates back to when it was necessary to rely on distribution tables.

EXAMPLE 4.12

Dolphins are at the top of the oceanic food chain; as a consequence, dangerous substances such as mercury tend to be present in their organs and muscles at high concentrations. In areas where dolphins are regularly consumed, it is important to monitor dolphin mercury levels. This example uses data from a random sample of 19 Risso's dolphins from the Taiji area in Japan.⁶ Calculate the 95% confidence interval for average mercury content in Risso's dolphins from the Taiji area using the data in Figure 4.12.

The observations are a simple random sample consisting of less than 10% of the population, so independence of the observations is reasonable. The summary statistics in Figure 4.12 do not suggest any skew or outliers; all observations are within 2.5 standard deviations of the mean. Based on this evidence, the approximate normality assumption seems reasonable.

E

Use the t -distribution to calculate the confidence interval:

$$\begin{aligned}\bar{x} \pm t_{df}^* \times SE &= \bar{x} \pm t_{18}^* \times s/\sqrt{n} \\ &= 4.4 \pm 2.10 \times 2.3/\sqrt{19} \\ &= (3.29, 5.51) \text{ } \mu\text{g/wet g.}\end{aligned}$$

The t^* point can be read from the t -table on page 93, in the column with area totaling 0.05 in the two tails (third column) and the row with 18 degrees of freedom. Based on these data, one can be 95% confident the average mercury content of muscles in Risso's dolphins is between 3.29 and 5.51 $\mu\text{g/wet gram}$.

Alternatively, the t^* point can be calculated in R with the function `qt`, which returns a value of 2.1009.

n	\bar{x}	s	minimum	maximum
19	4.4	2.3	1.7	9.2

Figure 4.12: Summary of mercury content in the muscle of 19 Risso's dolphins from the Taiji area. Measurements are in $\mu\text{g/wet g}$ (micrograms of mercury per wet gram of muscle).

GUIDED PRACTICE 4.13**G**

The FDA's webpage provides some data on mercury content of various fish species.⁷ From a sample of 15 white croaker (Pacific), a sample mean and standard deviation were computed as 0.287 and 0.069 ppm (parts per million), respectively. The 15 observations ranged from 0.18 to 0.41 ppm. Assume that these observations are independent. Based on summary statistics, does the normality assumption seem reasonable? If so, calculate a 90% confidence interval for the average mercury content of white croaker (Pacific).⁸

⁶Taiji is a significant source of dolphin and whale meat in Japan. Thousands of dolphins pass through the Taiji area annually; assume that these 19 dolphins represent a simple random sample. Data reference: Endo T and Haraguchi K. 2009. High mercury levels in hair samples from residents of Taiji, a Japanese whaling town. *Marine Pollution Bulletin* 60(5):743-747.

⁷www.fda.gov/food/foodborneillnesscontaminants/metals/ucm115644.htm

⁸There are no obvious outliers; all observations are within 2 standard deviations of the mean. If there is skew, it is not evident. There are no red flags for the normal model based on this (limited) information. $\bar{x} \pm t_{14}^* \times SE \rightarrow 0.287 \pm 1.76 \times 0.0178 \rightarrow (0.256, 0.318)$. We are 90% confident that the average mercury content of croaker white fish (Pacific) is between 0.256 and 0.318 ppm.

4.3 Confidence interval for a single proportion

Advanced melanoma is an aggressive form of skin cancer that until recently was almost uniformly fatal. In rare instances, a patient's melanoma stopped progressing or disappeared altogether when the patient's immune system successfully mounted a response to the cancer. Those observations led to research into therapies that might trigger an immune response in cancer. Some of the most notable successes have been in melanoma, particularly with two new therapies, nivolumab and ipilimumab.⁹

A 2013 report in the New England Journal of Medicine by Wolchok et al. reported the results of a study in which patients were treated with both nivolumab and ipilimumab.¹⁰ Fifty-three patients were given the new regimens concurrently, and the response to therapy could be evaluated in 52 of the 53. Of the 52 evaluable patients, 21 (40%) experienced a response according to commonly accepted criteria. In previous studies, the proportion of patients responding to one of these agents was 30% or less. How might one compare the new data to past results?

The data from this study are binomial data, with success defined as a response to therapy. Suppose the number of patients who respond in a study like this is represented by the random variable X , where X is binomial with parameters n (the number of trials, where each trial is represented by a patient) and p (the unknown population proportion of response). From formulas discussed in Chapter 2, the mean of X is np and the standard deviation of X is $\sqrt{np(1-p)}$.

Inference about p is based on the sample proportion \hat{p} , where $\hat{p} = X/n$. In this case, $\hat{p} = 21/52 = 0.404$. If the sample proportion is nearly normally distributed, the normal approximation to the binomial distribution can be used to conduct inference; this method is commonly used.

4.3.1 Inference using the normal approximation

A **sample proportion** can be described as a sample mean. If each success in the melanoma data is represented as a 1 and each failure as a 0, then the sample proportion is the mean of the 52 numerical outcomes:

$$\hat{p} = \frac{0 + 1 + 1 + \cdots + 0}{52} = 0.404.$$

The distribution of \hat{p} is nearly normal when the distribution of successes and failures is not too strongly skewed.

⁹The -mab suffix in these therapies stands for monoclonal antibody, a therapeutic agent made by identical immune cells that are all clones of a unique parent cell from a patient.

¹⁰N Engl J Med 2013;369:122-33. DOI: 10.1056/NEJMoa1302369

CONDITIONS FOR THE SAMPLING DISTRIBUTION OF \hat{p} BEING NEARLY NORMAL

The sampling distribution for \hat{p} , calculated from a sample of size n from a population with a success proportion p , is nearly normal when

1. the sample observations are independent and
2. at least 10 successes and 10 failures are expected in the sample, i.e. $np \geq 10$ and $n(1-p) \geq 10$. This is called the **success-failure condition**.

If these conditions are met, then the sampling distribution of \hat{p} is approximately normal with mean p and standard error

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}. \quad (4.14)$$

When conducting inference, the population proportion p is unknown. Thus, to construct a confidence interval, the sample proportion \hat{p} can be substituted for p to check the success-failure condition and compute the standard error.

Confidence intervals for a proportion

When using the normal approximation to the sampling distribution of \hat{p} , a confidence interval for a proportion has the same structure as a confidence interval for a mean; it is centered at the point estimate, with a margin of error calculated from the standard error and appropriate z^* value. The formula for a 95% confidence interval is

$$\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

EXAMPLE 4.15

Using the normal approximation, construct an approximate 95% confidence interval for the response probability for patients with advanced melanoma who were administered the combination of nivolumab and ipilimumab.

The independence and success-failure assumptions should be checked first. Since the outcome of one patient is unlikely to influence that of other patients, the observations are independent. The success-failure condition is satisfied since $n\hat{p} = (52)(.404) = 21 > 10$ and $n\hat{p}(1-\hat{p}) = (52)(.596) = 31 > 10$.

The point estimate for the response probability, based on a sample of size $n = 52$, is $\hat{p} = 0.404$. For a 95% confidence interval, $z^* = 1.96$. The standard error is estimated as: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.404)(1-0.404)}{52}} = 0.068$. The confidence interval is

$$0.404 \pm 1.96(0.068) \rightarrow (0.27, 0.54)$$

The approximate 95% confidence interval for p , the population response probability of melanoma patients to the combination of these new drugs, is $(0.27, 0.54)$ or $(27\%, 54\%)$.

\hat{p}
sample
proportion

p
population
proportion

E

GUIDED PRACTICE 4.16**G**

In New York City on October 23rd, 2014, a doctor who had recently been treating Ebola patients in Guinea went to the hospital with a slight fever and was subsequently diagnosed with Ebola. Soon after, a survey conducted by the Marist Poll, an organization with a carefully designed methodology for drawing random samples from identified populations, found that 82% of New Yorkers favored a "mandatory 21-day quarantine for anyone who has come in contact with an Ebola patient."¹¹ a) Verify that the sampling distribution of \hat{p} is nearly normal. b) Construct a 95% confidence interval for p , the proportion of New York adults who supported a quarantine for anyone who has come into contact with an Ebola patient.¹²

Did the participants in the melanoma trial constitute a random sample? Patients who participate in clinical trials are unlikely to be a random sample of patients with the disease under study since the patients or their physicians must be aware of the trial, and patients must be well enough to travel to a major medical center and be willing to receive an experimental therapy that may have serious side effects.

Investigators in the melanoma trial were aware that the observed proportion of patients responding in a clinical trial may be different than the hypothetical response probability in the population of patients with advanced melanoma. Study teams try to minimize these systematic differences by following strict specifications for deciding whether patients are eligible for a study. However, there is no guarantee that the results observed in a sample will be replicated in the general population.

Small, initial studies in which there is no control group, like the one described here, are early steps in exploring the value of a new therapy and are used to justify further study of a treatment when the results are substantially different than expected. The largest observed response rate in previous trials of 30% was close to the lower bound of the confidence interval from the study (27%, 54%), so the results were considered adequate justification for continued research on this treatment.

4.3.2 Sample size calculation of population proportion

Before sampling your population to estimate a proportion is convenient to know the sample size you will need. The theoretical reasoning to get the formula is similar to the seen in the section 4.1.4. However, for the standard error of the proportion is depending on the population proportion, which is the unknown value to estimate. To prevent this problem, a pre-estimated value p_0 of the population proportion is used.

Formula to estimate the sample size The sample size to estimate the population proportion with a maximum error, max.error, at a confidence $100(1 - \alpha)\%$ with a previously estimated value p_0 .

$$n > \frac{z_{1-\alpha/2}^2 p_0(1-p_0)}{\text{max.error}^2}$$

● Example 4.17

¹¹ Poll ID NY141026 on maristpoll.marist.edu.

¹²a) The poll is based on a simple random sample and consists of fewer than 10% of the adult population of New York, which makes independence a reasonable assumption. The success-failure condition is satisfied since, $1042(0.82) > 5$ and $1042(1 - 0.82) > 5$. b) $0.82 \pm 1.96 \sqrt{\frac{0.82(1-0.82)}{1042}} \rightarrow (0.796, 0.844)$.

What is the number of dogs we need to estimate the proportion of stray dogs vaccinated of leptospirosis with a confidence of 95% of the (absolute) error being less than 5%? From the data in a vet clinic, we have a rough idea that they are going to be around 65% with a confidence of 95%

Our confidence is 95%, so $\alpha = 0.05$ and $1 - \alpha/2 = 0.975$. The z-score to use is $z_{0.975} = 1.96$ and applying the formula with $p_0 = 0.65$

$$n > \frac{1.96^2 \cdot 0.65 \cdot 0.35}{0.05^2} = 349.5856$$

So, we need at least 350 dogs.

In some settings a preliminary estimate for p can be used to calculate n . When no estimate is available, calculus can be used to show that $p(1 - p)$ has its largest value when $p = 0.50$, and that conservative value for p is often used to ensure that n is sufficiently large regardless of the value of the unknown population proportion p . In that case, n satisfies

$$n \geq \frac{(z_{1-\alpha/2})^2(0.50)(1 - 0.50)}{m^2} = \frac{(z_{1-\alpha/2})^2}{4m^2}.$$

EXAMPLE 4.18

Donor organs for organ transplant are scarce. Studies are conducted to explore whether the population of eligible organs can be expanded. Suppose a research team is studying the possibility of transplanting lungs from hepatitis C positive individuals; recipients can be treated with one of the new drugs that cures hepatitis C. Preliminary studies in organ transplant are often designed to estimate the probability of a successful organ graft 6 months after the transplant. How large should a study be so that the 95% confidence interval for the probability of a successful graft at 6 months is no wider than 20%?

E

A confidence interval no wider than 20% has a margin of error of 10%, or 0.10. Using the conservative value $p = 0.50$,

$$n = \frac{(1.96)^2}{(4)(0.10^2)} = 96.04.$$

Sample sizes are always rounded up, so the study should have 97 patients.

Since the study will likely yield a value \hat{p} different from 0.50, the final margin of error will be smaller than ± 0.10 .

When the confidence coefficient is 95%, 1.96 can be replaced by 2 and the sample size formula reduces to

$$n = 1/m^2.$$

This remarkably simple formula is often used by practitioners for a quick estimate of sample size.

GUIDED PRACTICE 4.19

G

A 2015 estimate of Congress' approval rating was 19%.¹³ Using this estimate, how large should an additional survey be to produce a margin of error of 0.04 with 95% confidence?¹⁴

Sample size calculation is an essential step for an experimental design not only for estimation, but also for hypothesis testings, which are going to be studied during next units. However, these

¹³www.gallup.com/poll/183128/five-months-gop-congress-approval-remains-low.aspx

¹⁴Apply the formula

sample calculate are out of the scope of this introductory courses and we are not going to provide more details in units to come.

$$1.96 \times \sqrt{\frac{p(1-p)}{n}} \approx 1.96 \times \sqrt{\frac{0.19(1-0.19)}{n}} \leq 0.04 \quad \rightarrow \quad n \geq 369.5.$$

A sample size of 370 or more would be reasonable.

4.4 Exercises

Exercise. 4.1

Suppose X is a random variable representing the amount of organic sulfur contained in 500mg methyl sulfonyl methane (MSM) tablets. Organic sulfur plays an important role in the maintenance of joint cartilage. The formula for MSM requires that each tablet should contain 85mg of organic sulfur. To assess how well the requirement is being followed, 24 MSM tablets were randomly selected. Analytic results show that on the average the tablets contain about 80.5mg of organic sulfur. If, based on experience, the amount of organic sulfur found in MSM tablets is normally distributed with $\sigma = 14.2$ mg, construct a 95%–confidence interval of the population mean μ .¹⁵

Exercise. 4.2

It is reasonable to assume that the level of triglycerides (milligrams per decaliter) is normally distributed. A previous study indicates that the variance σ^2 among male subjects is $86.4\text{mg}^2/\text{dL}^2$. We want to use the sample mean of a random sample to estimate the population mean of the triglyceride level and want to be able to assert with probability 0.85 that our error will be no more than (at most) 2.75mg/dL. How large a sample will we need?¹⁶

Exercise. 4.3

An experiment has been designed in order to estimate the number of heartbeat per minute in 5-year-old dogs. The experiment was done with 49 dogs randomly selected from a public register and their pulses were measured. It was found that the averaged number of heartbeat per minute was 90. If the variance of number of beat per minute is $\sigma^2 = 100$, what is the 95% confidence interval of the mean of the number of beats per minute.

Exercise. 4.4

Gas analysis of arterial blood in a sample of 15 lab rats is measured using oximetry. The sample provides the following values of the arterial oxygen tension PaO_2 for standing gas

75, 80, 84, 74, 84, 78, 89, 72, 83, 76, 75, 87, 78, 79, 88

Calculate a 95% confidence interval of the mean of the population if we can assume that values of the oxygen tension is normally distributed.

Exercise. 4.5

As a part of a study of the queue time in a vet hospital, it was reported that a sample of 100 patients were at the waiting room an averaged time of 23 minutes. The sample standard deviation was 10 minutes. Calculate a 90% confidence interval of the mean waiting time of a patient at the vet clinic and interpret the result.

Exercise. 4.6

A sample of 25 ten year-old male gorillas provided a mean weight and a standard deviation of 36.5 kg and 5 kg, respectively. Find a 90% confidence interval of the mean weight of male gorillas of the population and give an interpretation of that interval.

¹⁵J.S. Kim and R.J. Dailey. *Biostatistics for Oral Healthcare*. Wiley, 2008. ISBN: 9780470388273. URL: <https://books.google.es/books?id=n1fR0LF1jhsC>.

¹⁶Ibidem

Exercise. 4.7

The following values are bilirubin concentrations of a sample of 20 dogs which suffer from hepatitis:

20, 5, 14, 8, 21, 3, 12, 7, 15, 2,
26, 6, 23, 4, 22, 9, 15, 7, 19, 2

Assuming that values of bilirubin concentration is normally distributed. Construct a 95% confidence interval of the mean bilirubin concentration for dogs suffering from hepatitis. According to all the studies bilirubin concentration is normally distributed.

Exercise. 4.8

A study is conducted to test the hypothesis that people with glaucoma have higher-than-average blood pressure. The study includes 200 people with glaucoma whose mean SBP is 140mmHg with a standard deviation of 25mmHg.

1. Construct a 95%-confidence interval for the true mean SBP among people with glaucoma.
2. If the average SBP for people of comparable age is 130mmHg, is there an association between glaucoma and blood pressure?¹⁷

Exercise. 4.9

We assume that the length of the Petal for the iris-setosa is normally distributed. If the length of 20 flower is the following:

1.4, 1.4, 1.3, 1.5, 1.4, 1.7, 1.4, 1.5, 1.4, 1.5,
1.5, 1.6, 1.4, 1.1, 1.2, 1.5, 1.3, 1.4, 1.7, 1.5

Find a 95%-confidence interval for the mean μ of the petal length.

Exercise. 4.10

Cardiologists want to estimate the average duration of chest pain experienced by AMI (acute myocardial infarction) patients. Suppose that the duration is normally distributed with variance $\sigma^2 = 424$. Chest pain data were collected from 24 AMI patients who have been admitted to a coronary care unit. With what probability can you assert that the maximum error of estimate will be no greater than 10 min?

Exercise. 4.11

Suppose a clinical trial is conducted to test the efficacy of a new drug, spectinomycin, for treating gonorrhea in females. Forty-six patients are given a 4-g daily dose of the drug and are seen a week later, at which time 6 of the patients still have gonorrhea. What is a 95%-confidence interval for p the probability of a failure with the drug?¹⁸

Exercise. 4.12

A university clinic offers a flu vaccine each year. The flu vaccine does not protect individuals from being infected by the flu virus. It only reduces the probability of infection. Of 375 individuals who were given a flu vaccine, 64 had the flu during the last flu season. Construct a 90% confidence interval for the proportion of the people who will not get the flu after receiving a flu shot.¹⁹

¹⁷B. Rosner. *Fundamentals of Biostatistics*. Cengage Learning, 2015. ISBN: 9781305465510. URL: <https://books.google.es/books?id=yn4yBgAAQBAJ>.

¹⁸B. Rosner. *Fundamentals of Biostatistics*. Cengage Learning, 2015. ISBN: 9781305465510. URL: <https://books.google.es/books?id=yn4yBgAAQBAJ>.

¹⁹J.S. Kim and R.J. Dailey. *Biostatistics for Oral Healthcare*. Wiley, 2008. ISBN: 9780470388273. URL: <https://books.google.es/books?id=n1fR0LF1jhsC>.

Chapter 5

One-sample hypothesis testing

5.1 Hypothesis testing

5.2 Hypothesis testing of the mean

5.3 Hypothesis testing of proportions

5.4 Understanding hypothesis testing

5.5 Notes

5.6 Exercises

This chapter introduces a method for testing scientific hypotheses about μ . The concepts used in this chapter will appear throughout the rest of the book, for other settings of hypotheses.

In this chapter hypothesis testing will be applied to three cases related to veterinary :

1. Analyzing the effect of vegetarian diet in pigs and compare the weight of a vegetarian pigs and the weight of a pigs which eat typical animal feed.
 2. Comparing heart rate of dogs living close to a heavy traffic road with the usual heart rate.
 3. Checking whether there exists any difference between the prevalence of a parasitic disease in a country with another country whose data are corroborated by several studies.
-



For labs, slides, and other resources, please visit
www.openintro.org/book/biostat

5.1 Hypothesis testing

Important decisions in science, such as whether a new treatment for a disease should be approved for the market, are primarily data-driven. For example, does a clinical study of a new cholesterol-lowering drug provide robust evidence of a beneficial effect in patients at risk for heart disease? A confidence interval can be calculated from the study data to provide a plausible range of values for a population parameter, such as the population average decrease in cholesterol levels. A drug is considered to have a beneficial effect on a population of patients if the population average effect is large enough to be clinically important. It is also necessary to evaluate the strength of the evidence that a drug is effective; in other words, is the observed effect larger than would be expected from chance variation alone?

Hypothesis testing is a method for calculating the probability of making a specific observation under a working hypothesis, called the null hypothesis. By assuming that the data come from a distribution specified by the null hypothesis, it is possible to calculate the likelihood of observing a value as extreme as the one represented by the sample. If the chances of such an extreme observation are small, there is enough evidence to reject the null hypothesis in favor of an alternative hypothesis.

NULL AND ALTERNATIVE HYPOTHESES

The **null hypothesis** (H_0) often represents either a skeptical perspective or a claim to be tested. The **alternative hypothesis** (H_A) is an alternative claim and is often represented by a range of possible parameter values.

Generally, an investigator suspects that the null hypothesis is not true and performs a hypothesis test in order to evaluate the strength of the evidence against the null hypothesis. The logic behind rejecting or failing to reject the null hypothesis is similar to the principle of presumption of innocence in many legal systems. In the United States, a defendant is assumed innocent until proven guilty; a verdict of guilty is only returned if it has been established beyond a reasonable doubt that the defendant is not innocent. In the formal approach to hypothesis testing, the null hypothesis (H_0) is not rejected unless the evidence contradicting it is so strong that the only reasonable conclusion is to reject H_0 in favor of H_A .

We could use a reasoning based on confidence intervals as it is shown in the next example.

EXAMPLE 5.1

The heart rate is known to have a mean value of 90 for Great Dane dogs in normal conditions.^a We try to verify if there is any difference between the heart rate for dogs living in a noisy neighbourhood and another dog living in a quieter neighbourhood. With this purpose, we have randomly selected 10 Great Danes living close to heavy traffic road the number of beats per minutes are the following:

113, 95, 97, 107, 85, 100, 100, 115, 98, 112

We need to assume that the heart rate is normally distributed in order to be able to calculate a confidence interval. We set up the hypothesis where μ is the average heart rate of dogs living close to heavy traffic area:

H_0 : The heart rate of Great Danes living close to heavy traffic area is the same that the average heart rate of the general population, $\mu = 90$

H_A : The heart rate of Great Danes living close to heavy traffic area is different from the average heart rate of the general population, $\mu \neq 90$

The sample size is 10, the sample mean is 102.2 and the sample variance is 82.0667. Hence, the estimation of the standard error is $SE = \sqrt{82.0667/10} = 2.9884$. We can construct a 95% confidence interval by means of the t distribution since population variance is unknown.

$$\bar{x} \pm t_{9, .975} SE = 102.2 \pm 2.262 \cdot 2.9884 \rightarrow (95.4402, 108.9598)$$

The value 90 does not belong to the interval and if the null hypothesis were true and the average heart rate were 90, this event would happen less than 5% of the times. Therefore, there exists evidence that the average heart rate is different from 90.

^aDisclaim: These data are not real. They are generated in order to provide an example, but has nothing to do with the true heart rate of a dog.

In health and biological science is preferred to follow the steps in formal hypothesis testing presented in the next section, which is applied when data are analyzed to support a decision or make a scientific claim.

5.1.1 The Formal Approach to Hypothesis Testing

In this section, hypothesis testing will be used to address the question of whether Americans generally wish to be heavier or lighter than their current weight. In the cdc data, the two variables weight and wt desire are, respectively, the recorded actual and desired weights for each respondent, measured in pounds.

Suppose that μ is the population average of the difference weight – wt desire. Using the observations from cdc.samp, assess the strength of the claim that, on average, there is no systematic preference to be heavier or lighter.

There are two techniques to apply hypothesis testing, one of them is based on calculating a p-value and the other one an acceptance interval. The technique of the p-value is difficult to apply when using tables except for a big sample, but statistical software, such as R-Commander, display this value and the conclusion must be drawn from it.

Step 1: Formulating null and alternative hypotheses

The claim to be tested is that the population average of the difference between actual and desired weight for US adults is equal to 0.

$$H_0 : \mu = 0.$$

In the absence of prior evidence that people typically wish to be lighter (or heavier), it is reasonable to begin with an alternative hypothesis that allows for differences in either direction.

$$H_A : \mu \neq 0.$$

The alternative hypothesis $H_A : \mu \neq 0$ is called a **two-sided alternative**. A one-sided alternative could be used if, for example, an investigator felt there was prior evidence that people typically wish to weigh less than they currently do: $H_A : \mu > 0$.

More generally, when testing a hypothesis about a population mean μ , the null and alternative hypotheses are written as follows

- For a two-sided alternative:

$$H_0 : \mu = \mu_0, H_A : \mu \neq \mu_0.$$

- For a one-sided alternative:

$$H_0 : \mu = \mu_0, H_A : \mu < \mu_0 \quad \text{or} \quad H_0 : \mu = \mu_0, H_A : \mu > \mu_0.$$

The symbol μ denotes a population mean, while μ_0 refers to the numeric value specified by the null hypothesis; in this example, $\mu_0 = 0$ (no difference between desired and real weight). Note that null and alternative hypotheses are statements about the underlying population, not the observed values from a sample. The **response variable** is the used to calculate the mean or any other parameter, such as proportion.

Step 2: Specifying a significance level, α

It is important to specify how rare or unlikely an event must be in order to represent sufficient evidence against the null hypothesis. This should be done during the design phase of a study, to prevent any bias that could result from defining 'rare' only after analyzing the results.

When testing a statistical hypothesis, an investigator specifies a **significance level**, α , that defines a 'rare' event. Typically, α is chosen to be 0.05, though it may be larger or smaller, depending on context; this is discussed in more detail in Section 5.4.1. An α level of 0.05 implies that an event occurring with probability lower than 5% will be considered sufficient evidence against H_0 .

Step 3: Calculating the test statistic

A *test statistic* is a special summary statistic that is particularly useful for evaluating a hypothesis test or identifying the p-value. In general, it has the form

$$\frac{\text{point estimate} - \text{null value}}{SE_{\text{point estimate}}}$$

Therefore, calculating the test statistic t is analogous to standardizing observations with Z-scores. For a hypothesis testing of the mean, the mean test statistic quantifies the number of standard deviations between the sample mean \bar{x} and the population mean μ :

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

where s is the sample standard deviation and n is the number of observations in the sample. If $x = \text{weight} - \text{wt desire}$, then for the 60 recorded differences in `cdc.samp`, $\bar{x} = 18.2$ and $s = 33.46$. In this sample, respondents weigh on average about 18 lbs more than they wish. The test statistic is

$$t = \frac{18.2 - 0}{33.46/\sqrt{60}} = 4.22.$$

The observed sample mean is 4.22 standard deviations to the right of $\mu_0 = 0$.

Step 4 (for large sample): Calculating the p -value

The **p -value** is the probability of observing a sample mean as or more extreme than the observed value, under the assumption that the null hypothesis is true. In samples of size 40 or more, the t -statistic will have a standard normal distribution unless the data are strongly skewed or extreme outliers are present. Recall that a standard normal distribution has mean 0 and standard deviation 1.

For two-sided tests, with $H_A : \mu \neq \mu_0$, the p -value is the sum of the area of the two tails defined by the t -statistic: $2P(Z \geq |t|) = P(Z \leq -|t|) + P(Z \geq |t|)$ (Figure 5.1).

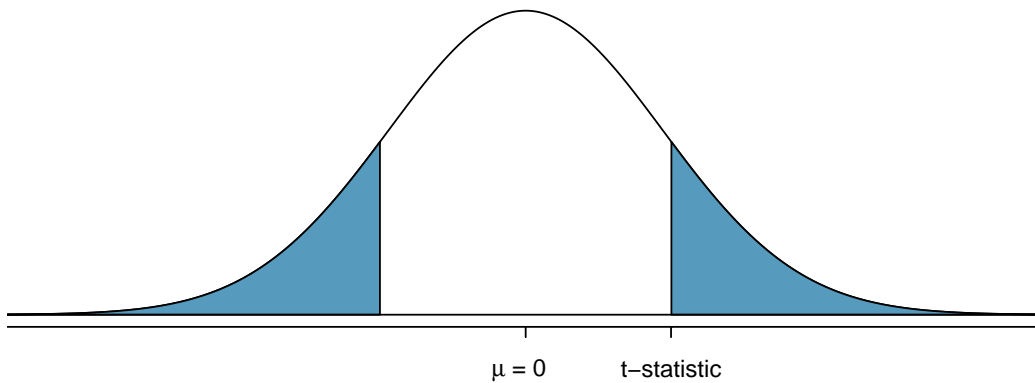


Figure 5.1: A two-sided p -value for $H_A : \mu \neq \mu_0$ on a standard normal distribution. The shaded regions represent observations as or more extreme than \bar{x} in either direction.

For one-sided tests with $H_A : \mu > \mu_0$, the p -value is given by $P(Z \geq t)$, as shown in Figure 5.2. If $H_A : \mu < \mu_0$, the p -value is the area to the left of the t -statistic, $P(Z \leq t)$.

The p -value can either be calculated from software or from the normal probability tables. For the weight-difference example, the p -value is vanishingly small: $p = P(Z \leq -4.22) + P(Z > 4.22) < 0.001$.

Step 4 (for any sample): Calculate acceptance interval

Instead of calculating p -value, which is always possible to calculate with an adequate software, set an interval containing the values for which the null hypothesis is failed to reject. The decision is made depending on the value being inside or outside the interval. This interval is calculated with percentiles of the distribution of the test statistic, pay attention that the distribution of the test statistic is not the same as the response variable.

The critical value(s) define the *rejection area* is the range the values of the test statistic for which is said the test result is statistically significant and the test hypothesis rejected.

The *acceptance region* is the range of values of a test statistics for which there is not enough statistical evidence to reject the null hypotheses. So, the null hypothesis is failed to reject For the two-sided tests, if the sample is small, the acceptance interval is given by percentiles of the

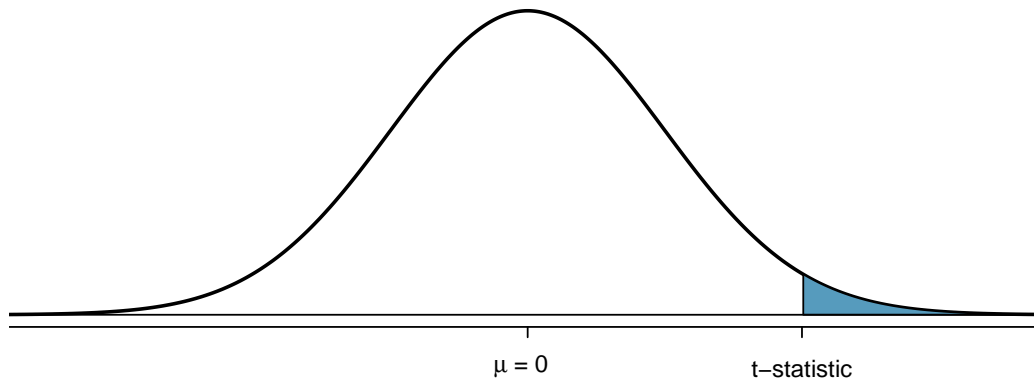


Figure 5.2: A one-sided p -value for $H_A : \mu > \mu_0$ on a standard normal distribution is represented by the shaded area to the right of the t -statistic. This area equals the probability of making an observation as or more extreme than \bar{x} , if the null hypothesis is true.

t distribution $(-t_{n-1, 1-\alpha/2}, t_{n-1, 1-\alpha/2})$ and if the sample is large enough, the acceptance interval is given by percentile of the normal distribution, $(-z_{1-\alpha/2}, z_{1-\alpha/2})$.

Step 5: Drawing a conclusion

To reach a conclusion about the null hypothesis, directly compare p and α . Note that for a conclusion to be informative, it must be presented in the context of the original question; it is not useful to only state whether or not H_0 is rejected.

If $p > \alpha$, the observed sample mean is not extreme enough to warrant rejecting H_0 ; more formally stated, there is insufficient evidence to reject H_0 . A high p -value suggests that the difference between the observed sample mean and μ_0 can reasonably be attributed to random chance.

If $p \leq \alpha$, there is sufficient evidence to reject H_0 and accept H_A . In the `cdc.samp weight-difference` data, the p -value is very small, with the t -statistic lying to the right of the population mean. The chance of drawing a sample with mean as large or larger than 18.2 if the distribution were centered at 0 is less than 0.001. Thus, the data support the conclusion that on average, the difference between actual and desired weight is not 0 and is positive; people generally seem to feel they are overweight.

If the acceptance interval is calculated in step 4, null hypothesis is rejected if test statistic is outside the interval. Otherwise, the null hypothesis is failed to reject.

GUIDED PRACTICE 5.2



Suppose that the mean weight difference in the sampled group of 60 adults had been 7 pounds instead of 18.2 pounds, but with the same standard deviation of 33.46 pounds. Would there still be enough evidence at the $\alpha = 0.05$ level to reject $H_0 : \mu = 0$ in favor of $H_A : \mu \neq 0$?¹

¹Re-calculate the t -statistic: $(7 - 0)/(33.46/\sqrt{60}) = 1.62$. The p -value $P(Z \leq -1.62) + P(Z \geq 1.62) = 0.105$. Since $p > \alpha$, there is insufficient evidence to reject H_0 . In this case, a sample average difference of 7 is not large enough to discount the possibility that the observed difference is due to sampling variation, and that the observations are from a distribution centered at 0.

5.2 Hypothesis testing of the mean

If the null hypothesis refers to the mean of a response variable to a **reference value** μ_0 , the test statistic to use is $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$, called t -test statistic. We have to pay attention to the response variable and the sample size to determine its distribution. If the sample size is large enough, namely $n > 30$ and not very skewed, the t -test statistic is very close to follow a standard normal distribution. If the response variable is normally distributed, then the t -test statistic follows a Student's t distribution with $n - 1$ degrees of freedom. Consequently, if the sample is small and the response variable is clearly not normally distributed, we cannot apply neither normal distribution nor t -distribution. However, if it is a large sample and the response variable is normally distributed, using normal distribution or t -distribution are almost equivalent.

5.2.1 Two examples with large samples

EXAMPLE 5.3

In 2015, the National Sleep Foundation published new guidelines for the amount of sleep recommended for adults: 7-9 hours of sleep per night.² The NHANES survey includes a question asking respondents about how many hours per night they sleep; the responses are available in `nhanes.samp`. In the sample of 134 adults used in the BMI example, the average reported hours of sleep is 6.90, with standard deviation 1.39. Is there evidence that American adults sleep less than 7 hours per night?

Let μ be the population average of hours of sleep per night for US adults. Conduct a one-sided test, since the question asks whether the average amount of sleep per night might be less than 7 hours.

Formulate the null and alternative hypotheses. $H_0 : \mu = 7$ hours vs. $H_A : \mu < 7$ hours.

Specify the significance level, α . Let $\alpha = 0.05$, since the question does not reference a different value.

Calculate the test statistic. The t -statistic has value

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{6.90 - 7.00}{1.33/\sqrt{134}} = -0.864.$$

Calculate the p -value.

For this one-sided alternative $H_A : \mu < 7$, the p -value is

$$P(Z \leq t) = P(Z < -0.864) = 0.19.$$

Since the alternative states that μ_0 is less than 7, the p -value is represented by the area to the left of $t = -0.864$, as shown in Figure 5.3.

Draw a conclusion. The p -value is larger than the specified significance level α . The null hypothesis is not rejected since the data do not represent sufficient evidence to support the claim that American adults sleep less than 7 hours per night.

²Sleep Health: Journal of the National Sleep Foundation, Vol. 1, Issue 1, pp. 40 - 43

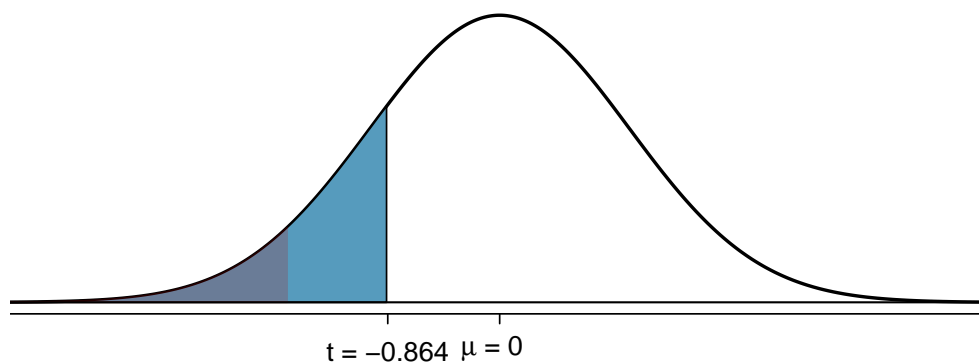


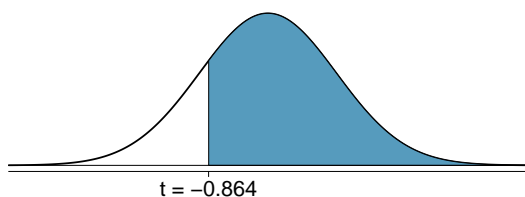
Figure 5.3: The large blue shaded region represents the p -value, the area to the left of $t = -0.864$. The smaller grey shaded region represents the rejection region of area 0.05 in the left tail.

GUIDED PRACTICE 5.4

G

From these data, is there sufficient evidence at the $\alpha = 0.10$ significance level to support the claim that American adults sleep more than 7 hours per night?³

³The t -statistic does not change from 1.65. Re-calculate the p -value since the alternative hypothesis is now $H_A : \mu > 7$: $P(Z \geq -0.864) = 0.81$. Since $p > \alpha$, there is insufficient evidence to reject H_0 at $\alpha = 0.10$. A common error when conducting one-sided tests is to assume that the p -value will always be the area in the smaller of the two tails to the right or left of the observed value. It is important to remember that the area corresponding to the p -value is in the direction specified by the alternative hypothesis.



EXAMPLE 5.5

It is known that the general population of pigs bred on farms have a average weight of 57 kg after 6 month. What happen with pigs which are feed with a vegetarian diet. He have collected 100 pigs 6 months-old fed with a vegetarian diet and the average weight was 55.9 kg and the standard deviation was 5 Kg. Is there any significant difference?

We can apply the normal model since we have the conditions of independent^a and the sample size being larger than 30. We assume that the distribution is unskewed, in case of being provided the data set a histogram should be generated to check that they are not strongly skewed.

The null hypothesis would be the skeptical point of view that the type of diet has nothing to do with the weight and the alternative would be that pigs fed with vegetarian diet have a lower weight.

H_0 The average weight of pigs fed with vegetarian diet is the same that pigs fed with the usual animal feed, $\mu = 57$.

H_1 The average weight of pigs fed with vegetarian diet is smaller than average weight for pigs fed with the usual animal feed, $\mu < 57$.

The t -statistics test will be $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{55.9 - 57}{5/\sqrt{100}} = -2.2$. The p-value is the probability of getting a value lower than this value if the null hypothesis is true. Hence, the p-value is the probability of a standard normal variable will be lower than -2.2 , so is the lower tail area of the normal distribution.

$$P(Z < -2.2) = 1 - P(Z < 2.2) = 1 - 0.9861 = 0.0139$$

This p-value is lower that $\alpha = 0.05$. We can consider that there exists a significant difference and that there is evidence that the average weight of pigs fed with vegetarian diet is smaller than the averaged weight of pigs eating the usual animal feed.

^aA remark of caution: Independence could be compromised if our sample were extracted from just a few farms instead of all the farms which are providing vegetarian diet. Since they are sharing conditions like temperatute, room, etc, two pigs from the same farm would not be completely independent.

E

5.2.2 Examples with small samples

EXAMPLE 5.6

While fish and other types of seafood are important for a healthy diet, nearly all fish and shellfish contain traces of mercury. Dietary exposure to mercury can be particularly dangerous for young children and unborn babies. Regulatory organizations such as the US Food and Drug Administration (FDA) provide guidelines as to which types of fish have particularly high levels of mercury and should be completely avoided by pregnant women and young children; additionally, certain species known to have low mercury levels are recommended for consumption. While there is no international standard that defines excessive mercury levels in saltwater fish species, general consensus is that fish with levels above 0.50 parts per million (ppm) should not be consumed. A study conducted to assess mercury levels for saltwater fish caught off the coast of New Jersey found that a sample of 23 bluefin tuna had mean mercury level of 0.52 ppm, with standard deviation 0.16 ppm.^a Based on these data, should the FDA add bluefin tuna from New Jersey to the list of species recommended for consumption, or should a warning be issued about their mercury levels?

Let μ be the population average mercury content for bluefin tuna caught off the coast of New Jersey. Conduct a two-sided test of the hypothesis $\mu = 0.50$ ppm in order to assess the evidence for either definitive safety or potential danger.

Formulate the null and alternative hypotheses. $H_0 : \mu = 0.50$ ppm vs. $H_A : \mu \neq 0.50$ ppm

Specify the significance level, α . A significance level of $\alpha = 0.05$ seems reasonable.

Calculate the test statistic. The t -statistic has value

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{0.52 - 0.50}{0.16/\sqrt{23}} = 0.599.$$

Calculate the acceptance interval.^b For a small sample, a response variable normally distributed and a two-sided alternative $H_A : \mu \neq 0.50$, the response variable is following the Student's t -distribution with $df = n - 1 = 22$ degrees of freedom, on the table $t_{df, 1-\alpha/2} = t_{22, 0.975} = 2.074$. So, the acceptance interval is $(-2.074, 2.074)$.

Draw a conclusion.^c The t -test statistic is not included in the acceptance interval, therefore there is not enough evidence to reject that the mean mercury level for the New Jersey coastal population of bluefin tuna is 0.50 ppm.

Note that "failure to reject" is not equivalent to "accepting" the null hypothesis. Recall the earlier analogy related to the principle of "innocent until proven guilty". If there is not enough evidence to prove that the defendant is guilty, the official decision must be "not guilty", since the defendant may not necessarily be innocent. Similarly, while there is not enough evidence to suggest that μ is not equal to 0.5 ppm, it would be incorrect to claim that the evidence states that μ is 0.5 ppm.

From these data, there is not statistically significant evidence to either recommend these fish as clearly safe for consumption or to warn consumers against eating them. Based on these data, the Food and Drug Administration might decide to monitor this species more closely and conduct further studies.

^aJ. Burger, M. Gochfeld, Science of the Total Environment 409 (2011) 1418–1429

^bThe p -value could be calculated using statistical software. The value is $2 \times P(t_{22} > 0.599) = 2 \times 0.2776 = 0.5552$

^cThe p -value is larger than the specified significance level α , as shown in Figure 5.4. The data do not show that the mercury content of bluefin tuna caught off the coast of New Jersey differs significantly from 0.50 ppm. Since $p > \alpha$, there is insufficient evidence to reject the null hypothesis

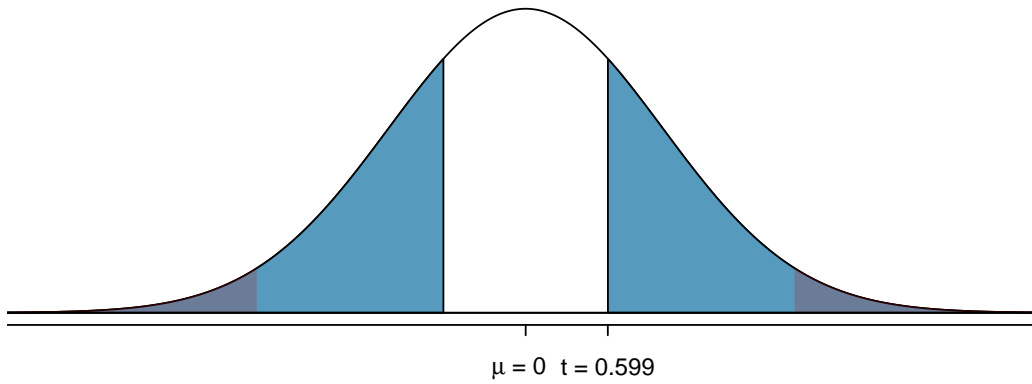


Figure 5.4: The large blue shaded regions represent the p -value, the area to the right of $t = 0.599$ and to the left of $-t = -0.599$. The smaller grey shaded regions represents the **rejection region** as defined by α ; in this case, an area of 0.025 in each tail. The t -statistic calculated from \bar{x} would have to lie within either of the extreme tail areas to constitute sufficient evidence against the null hypothesis. The grey shaded regions are bounded by -2.074 and 2.074, capturing 95% of the distribution of the test statistic.

EXAMPLE 5.7

Use the data in exercise 5.1 to check whether the heart rate of Great Danes living close to a heavy traffic road is different from the average heart rate. With a significance level of $\alpha = 0.05$

We assume that the distribution of the heart rate is normally distributed^a. We standardize with the same way than z , but we use the t statistic in order to indicate that this test statistic is using Student's t distribution.

E

$$t = \frac{\bar{x} - \text{null value}}{SE} = \frac{102.2 - 90}{9.43/\sqrt{10}} = \frac{102.2 - 90}{2.98} = 4.09$$

The acceptance interval is $(-t_{9, .975}, t_{9, .975}) = (-2.262, 2.262)$. The value of the test statistic does not belong to the interval. Therefore, null hypothesis is rejected and we have enough statistical evidence that there exists a difference between the average value of the heart rate of Great Danes living near heavy traffic roads and the general average heart rate for these dogs.

^aIn general, this procedure for hypothesis testing works if the distribution of the variable is symmetric.

5.3 Hypothesis testing of proportions

Just as with inference for population means, confidence intervals for population proportions can be used when deciding whether to reject a null hypothesis. It is useful in most settings, however, to calculate the p -value for a test as a measure of the strength of the evidence contradicting the null hypothesis.

When using the normal approximation for the distribution of \hat{p} to conduct a hypothesis test, one should always verify that \hat{p} is nearly normal under H_0 by checking the independence and success-failure conditions. Since a hypothesis test is based on the distribution of the test statistic under the null hypothesis, the success-failure condition is checked using the null proportion p_0 , not the estimate \hat{p} .

According to the normal approximation to the binomial distribution, the number of successes in n trials is normally distributed with mean np_0 and standard deviation $\sqrt{np_0(1-p_0)}$. This approximation is valid when np_0 and $n(1-p_0)$ are both at least 10.

Under the null hypothesis, the sample proportion $\hat{p} = X/n$ is approximately distributed as

$$N\left(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}\right).$$

The test statistic z for the null hypothesis $H_0 : p = p_0$ based on a sample of size n is

$$z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{\hat{p} - p_0}{\sqrt{\frac{(p_0)(1-p_0)}{n}}}.$$

EXAMPLE 5.8

Suppose that out of a cohort of 120 patients with stage 1 lung cancer at the Dana-Farber Cancer Institute (DFCI) treated with a new surgical approach, 80 of the patients survive at least 5 years, and suppose that National Cancer Institute statistics indicate that the 5-year survival probability for stage 1 lung cancer patients nationally is 0.60. Do the data collected from 120 patients support the claim that the DFCI population treated with this new form of surgery has a different 5-year survival probability than the national population? Let $\alpha = 0.10$, since this is an early study of the new surgery.

Test the hypothesis $H_0 : p = 0.60$ versus the alternative, $H_A : p \neq 0.60$, using $\alpha = 0.10$. If we assume that the outcome of one patient at DFCI does not influence the outcome of other patients, the independence condition is met, and the success-failure condition is satisfied since $(120)(0.60) = 80 > 5$ and $(120)(1 - 0.60) = 40 > 5$. The test statistic is the z -score of the point estimate:

$$z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.67 - 0.60}{\sqrt{\frac{(0.60)(1-0.60)}{120}}} = 1.57.$$

The p -value^a is the probability that a standard normal variable is larger than 1.57 or smaller than -1.57, $P(|Z| > 1.57) = 0.12$; since the p -value is greater than 0.10, there is insufficient evidence to reject H_0 in favor of H_A . There is not convincing evidence that the survival probability at DFCI differs from the national survival probability. Had a more traditional 0.05 significance level been used, the data would be even less convincing.

^aThe technique of the acceptance interval could be used by means of percentiles of the normal distribution $(-z_{1-\alpha/2}, z_{1-\alpha/2})$. Since $z_{0.95} = 1.645$, the acceptance interval is $(-1.645, 1.645)$.

EXAMPLE 5.9

Using the data from the study in advanced melanoma, use the normal approximation to the sampling distribution of \hat{p} to test the null hypothesis that the response probability to the novel combined therapy is 30% against a one-sided alternative that the response proportion is greater than 30%. Let $\alpha = 0.10$.

E

The test statistic has value

$$z = (0.404 - 0.30) / \sqrt{(0.30)(0.70)/52} = 1.64.$$

The one-sided p -value is $P(Z \geq 1.64) = 0.05$; there is sufficient evidence to reject the null hypothesis at $\alpha = 0.10$. This is an example of where a two-sided test and a one-sided test yield different conclusions.

GUIDED PRACTICE 5.10**G**

One of the questions on the National Health and Nutrition Examination Survey (introduced in Chapter 3) asked participants whether they participated in moderate or vigorous intensity sports, fitness, or recreational activities. In a random sample of 135 adults, 76 answered "Yes" to the question. Based on this evidence, are a majority of American adults physically active?⁴

⁴The observations are independent. Check success-failure: $np_0 = n(1 - p_0) = 135(0.5) > 10$. $H_0 : p = 0.5$; $H_A : p > 0.5$. Calculate the z -score: $z = \frac{0.56 - 0.50}{\sqrt{\frac{0.5(1-0.5)}{135}}} = 1.39$. The p -value is 0.08. Since the p -value is larger than 0.05, there is insufficient evidence to reject H_0 ; there is not convincing evidence that a majority of Americans are physically active, although the data suggest that may be the case.

EXAMPLE 5.11

Chagas disease or American Trypanosomiasis is a tropical parasitic disease. The infectious agent is the protozoan *Trypanosoma cruzi* and it is spread by insects, parasitizing dogs, cats and other pets. In Costa Rica has been performed a study in order to know if the prevalence of infected dogs by *Trypanosoma* is different from the proportion in USA. A large study in USA detected that in 8% of dogs were found antibody against this infectious agent. In Costa Rica a random sample of 176 dogs was selected and 11 in the sample were found antibody against *Trypanosoma cruzi*.

Is there any difference between USA and Costa Rica about the prevalence of infection by *Trypanosoma cruzi*?

The null and alternative hypothesis are

$$H_0 : p = 0.08 \text{ (8\%)}$$

$$H_A : p \neq 0.08$$

E

where p is the proportion of infected dogs in Costa Rica and $p_0 = 0.08$ is the proportion of infected dogs in USA.

The sample proportion is $\hat{p} = 11/176 = 0.0625$ and the standard error of this point estimation is $SE_{\hat{p}} = \sqrt{0.08 \cdot (1 - 0.08)/176} = 0.02$

The test statistics is

$$z = \frac{p - p_0}{SE_{\hat{p}}} = \frac{\hat{p} - p_0}{\sqrt{p_0 \cdot (1 - p_0)/n}} = \frac{0.0625 - 0.08}{0.02} = -0.8557$$

For a significance of 5% ($\alpha = 0.05$), the acceptance region is $(-z_{1-\alpha/2}, z_{1-\alpha/2}) = (-z_{.975}, z_{.975}) = (-1.96, 1.96)$. The test statistics belongs to the acceptance region. We are not able to reject the null hypothesis. In conclusion, there is not a statistical significant difference between the proportions of infected dogs in Costa Rica and USA. We have not evidence with this sample that there exists any difference between the proportion of infected dogs in Costa Rica and USA.

5.4 Understanding hypothesis testing

5.4.1 Decision errors

Hypothesis tests can potentially result in incorrect decisions, such as rejecting the null hypothesis when the null is actually true. Figure 5.5 shows the four possible ways that the conclusion of a test can be right or wrong.

		Test conclusion	
		Fail to reject H_0	Reject H_0 in favor of H_A
Reality	H_0 True	Correct Decision	Type 1 Error
	H_A True	Type 2 Error	Correct Decision

Figure 5.5: Four different scenarios for hypothesis tests.

Rejecting the null hypothesis when the null is true represents a **Type I error**, while a **Type II error** refers to failing to reject the null hypothesis when the alternative is true.

EXAMPLE 5.12

In a trial, the defendant is either innocent (H_0) or guilty (H_A). After hearing evidence from both the prosecution and the defense, the court must reach a verdict. What does a Type I Error represent in this context? What does a Type II Error represent?

E

If the court makes a Type I error, this means the defendant is innocent, but wrongly convicted (rejecting H_0 when H_0 is true). A Type II error means the court failed to convict a defendant that was guilty (failing to reject H_0 when H_0 is false).

The probability of making a Type I error is the same as the significance level α , since α determines the cutoff point for rejecting the null hypothesis. For example, if α is chosen to be 0.05, then there is a 5% chance of incorrectly rejecting H_0 .

The rate of Type I error can be reduced by lowering α (e.g., to 0.01 instead of 0.05); doing so requires an observation to be more extreme to qualify as sufficient evidence against the null hypothesis. However, this inevitably raises the rate of Type II errors, since the test will now have a higher chance of failing to reject the null hypothesis when the alternative is true.

EXAMPLE 5.13

In a courtroom setting, how might the rate of Type I errors be reduced? What effect would this have on the rate of Type II errors?

E

Lowering the rate of Type I error is equivalent to raising the standards for conviction such that fewer people are wrongly convicted. This increases Type II error, since higher standards for conviction leads to fewer convictions for people who are actually guilty.

GUIDED PRACTICE 5.14

G

In a courtroom setting, how might the rate of Type II errors be reduced? What effect would this have on the rate of Type I errors?⁵

⁵To lower the rate of Type II error, the court could lower the standards for conviction, or in other words, lower the bar

Choosing a significance level

Reducing the error probability of one type of error increases the chance of making the other type. As a result, the significance level is often adjusted based on the consequences of any decisions that might follow from the result of a significance test.

By convention, most scientific studies use a significance level of $\alpha = 0.05$; small enough such that the chance of a Type I error is relatively rare (occurring on average 5 out of 100 times), but also large enough to prevent the null hypothesis from almost never being rejected. If a Type I error is especially dangerous or costly, a smaller value of α is chosen (e.g., 0.01). Under this scenario, it is better to be cautious about rejecting the null hypothesis, so very strong evidence against H_0 is required in order to reject the null and accept the alternative. Conversely, if a Type II error is relatively dangerous, then a larger value of α is chosen (e.g., 0.10). Hypothesis tests with larger values of α will reject H_0 more often.

For example, in the early stages of assessing a drug therapy, it may be important to continue further testing even if there is not very strong initial evidence for a beneficial effect. If the scientists conducting the research know that any initial positive results will eventually be more rigorously tested in a larger study, they might choose to use $\alpha = 0.10$ to reduce the chances of making a Type II error: prematurely ending research on what might turn out to be a promising drug.

A government agency responsible for approving drugs to be marketed to the general population, however, would likely be biased towards minimizing the chances of making a Type I error—approving a drug that turns out to be unsafe or ineffective. As a result, they might conduct tests at significance level 0.01 in order to reduce the chances of concluding that a drug works when it is in fact ineffective. The US FDA and the European Medical Agency (EMA) customarily require that two independent studies show the efficacy of a new drug or regimen using $\alpha = 0.05$, though other values are sometimes used.

5.4.2 Choosing between one-sided and two-sided tests

In some cases, the choice of a one-sided or two-sided test can influence whether the null hypothesis is rejected. For example, consider a sample for which the t -statistic is 1.80. If a two-sided test is conducted at $\alpha = 0.05$, the p -value is

$$P(Z \leq -|t|) + P(Z \geq |t|) = 2P(Z \geq 1.80) = 0.072.$$

There is insufficient evidence to reject H_0 , since $p > \alpha$. However, what if a one-sided test is conducted at $\alpha = 0.05$, with $H_A : \mu > \mu_0$? In this case, the p -value is

$$P(Z \geq t) = P(Z \geq 1.80) = 0.036.$$

The conclusion of the test is different: since $p < \alpha$, there is sufficient evidence to reject H_0 in favor of the alternative hypothesis. Figure 5.6 illustrates the different outcomes from the tests.

Two-sided tests are more "conservative" than one-sided tests; it is more difficult to reject the null hypothesis with a two-sided test. The p -value for a one-sided test is exactly half the p -value for a two-sided test conducted at the same significance level; as a result, it is easier for the p -value from a one-sided test to be smaller than α . Additionally, since the rejection region for a two-sided test is divided between two tails, a test statistic needs to be more extreme in order to fall within a rejection region. While the t -statistic of 1.80 is not within the two-sided rejection region, it is within the one-sided rejection region.⁶

for what constitutes sufficient evidence of guilt (increase α , e.g. to 0.10 instead of 0.05). This will result in more guilty people being convicted, but also increase the rate of wrongful convictions, increasing the Type I error.

⁶The two-sided rejection regions are bounded by -1.96 and 1.96, while the one-sided rejection region begins at 1.65.

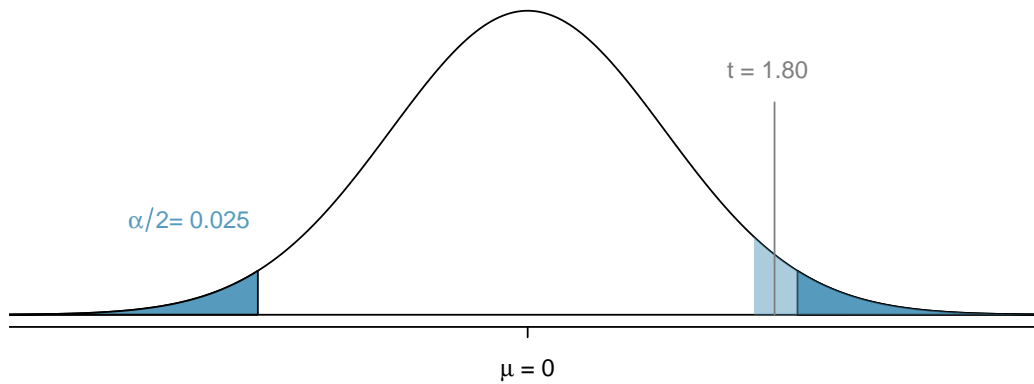


Figure 5.6: Under a one-sided test at significance level $\alpha = 0.05$, a t -statistic of 1.80 is within the rejection region (shaded light blue). However, it would not be within the rejection region under a two-sided test with $\alpha = 0.05$ (darker blue).

For a fixed sample size, a one-tailed test will have a smaller probability of Type II error in comparison to a two-tailed test conducted at the same α level. In other words, with a one-sided test, it is easier to reject the null hypothesis if the alternative is actually true.

The choice of test should be driven by context, although it is not always clear which test is appropriate. Since it is easier to reject H_0 with the one-tailed test, it might be tempting to always use a one-tailed test when a significant result in a particular direction would be interesting or desirable.

However, it is important to consider the potential consequences of missing a significant difference in the untested direction. Generally, a two-sided test is the safest option, since it does not incorporate any existing biases about the direction of the results and can detect a difference at either the upper or lower tail. In the 1980s, researchers were interested in assessing a new set of drugs expected to be more effective at reducing heart arrhythmias than previously available therapies. They designed a one-sided clinical trial, convinced that the newer therapy would reduce mortality. The trial was quickly terminated due to an unanticipated effect of the drug; an independent review board found that the newer therapy was almost 4 times as likely to kill patients as a placebo! In a clinical research setting, it can be dangerous and even unethical to conduct a one-sided test under the belief that there is no possibility of patient harm from the drug intervention being tested.

One-sided tests are appropriate if the consequences of missing an effect in the untested direction are negligible, or if a large observed difference in the untested direction and a conclusion of "no difference" lead to the same decision. For example, suppose that a company has developed a drug to reduce blood pressure that is cheaper to produce than current options available on the market. If the drug is shown to be equally effective or more effective than an existing drug, the company will continue investing in it. Thus, they are only interested in testing the alternative hypothesis that the new drug is less effective than the existing drug, in which case, they will stop the project. It is acceptable to conduct a one-sided test in this situation since missing an effect in the other direction causes no harm.

The decision as to whether to use a one-sided or two-sided test must be made before data analysis begins, in order to avoid biasing conclusions based on the results of a hypothesis test. In particular, changing to a one-sided test after discovering that the results are "almost" significant for the two-sided test is unacceptable. Manipulating analyses in order to achieve low p -values leads to invalid results that are often not replicable. Unfortunately, this kind of "significance-chasing" has become widespread in published science, leading to concern that most current published research findings are false.

5.4.3 The informal use of p -values

Formal hypothesis tests are designed for settings where a decision or a claim about a hypothesis follows a test, such as in scientific publications where an investigator wishes to claim that an intervention changes an outcome. However, progress in science is usually based on a collection of studies or experiments, and it is often the case that the results of one study are used as a guide for the next study or experiment.

Sir Ronald Fisher was the first to propose using p -values as one of the statistical tools for evaluating an experiment. In his view, an outcome from an experiment that would only happen 1 in 20 times ($p = 0.05$) was worth investigating further. The use of p -values for formal decision making came later. While valuable, formal hypothesis testing can often be overused; not all significant results should lead to a definitive claim, but instead prompt further analysis.

The formal use of p -values is emphasized here because of its prominence in the scientific literature, and because the steps outlined are fundamental to the scientific method for empirical research: specify hypotheses, state in advance how strong the evidence should be to constitute sufficient evidence against the null, specify the method of analysis and compute the test statistic, draw a conclusion. These steps are designed to avoid the pitfall of choosing a hypothesis or method of analysis that is biased by the data and hence reaches a conclusion that may not be reproducible.

5.5 Notes

Confidence intervals and hypothesis testing are two of the central concepts in inference for a population based on a sample. The confidence interval shows a range of population parameter values consistent with the observed sample, and is often used to design additional studies. Hypothesis testing is a useful tool for evaluating the strength of the evidence against a working hypothesis according to a pre-specified standard for accepting or rejecting hypotheses.

The calculation of p -values and confidence intervals is relatively straightforward; given the necessary summary statistics, α , and confidence coefficients, finding any p -value or confidence interval simply involves a set of formulaic steps. However, the more difficult parts of any inference problem are the steps that do not involve any calculations. Specifying appropriate null and alternative hypotheses for a test relies on an understanding of the problem context and the scientific setting of the investigation. Similarly, a choice about a confidence coefficient for an interval relies on judgment as to balancing precision against the chance of possible error. It is also not necessarily obvious when a significance level other than $\alpha = 0.05$ should be applied. These choices represent the largest distinction between a true statistics problem as compared to a purely mathematical exercise.

Furthermore, in order to rely on the conclusions drawn from making inferences, it is necessary to consider factors such as study design, measurement quality, and the validity of any assumptions made. For example, is it valid to use the normal approximation to calculate p -values? In small to moderate sample sizes ($30 \leq n \leq 50$), it may not be clear that the normal model is accurate. It is even necessary to be cautious about the use and interpretation of the p -value. For example, an article published in *Nature* about the mis-use of p -values references a published study that showed people who meet their spouses online are more likely to have marital satisfaction, with p -value less than 0.001. However, statistical significance does not measure the importance or practical relevance of a result; in this case, the change in happiness moved from 5.48 to 5.64 on a 7-point scale. A p -value reported without context or other evidence is uninformative and potentially deceptive.

These nuanced issues cannot be adequately covered in any introduction to statistics. It is unrealistic to encourage students to use their own judgment with aspects of inference that even experienced investigators find challenging. At the same time, it would also be misleading to suggest that the choices are always clear-cut in practice. It seems best to offer some practical guidance for getting started:

- The default choice of α is 0.05; similarly, the default confidence coefficient for a confidence interval is 95%.
- Unless it is clear from the context of a problem that change in only one direction from the null hypothesis is of interest, the alternative hypothesis should be two-sided.
- The use of a standard normal distribution to calculate p -values is reasonable for sample sizes of 30 or more if the distribution of data are not strongly skewed and there are no large outliers. If there is skew or a few large outliers, sample sizes of 50 or more are usually sufficient.
- Pay attention to the context of a problem, particularly when formulating hypotheses and drawing conclusions.

The next chapters will discuss methods of inference in specific settings, such as comparing two groups. These settings expand on the concepts discussed in this chapter and offer additional opportunities to practice calculating tests and intervals, reading problems for context, and checking underlying assumptions behind methods of inference.

5.6 Exercises

Exercise. 5.1

Although obstructive sleep apnea (OSA) patients tend to be obese, a significant number of them are not. A recent study suggested that fat deposition around the neck may be a factor associated with OSA in non-obese patients. Suppose that apnea hypopnea index (AHI), an index used to quantify the degree of obstructive sleep apnea, is normally distributed and that the average AHI for non-obese patients, whose body mass index is less than 25, is 7.65. It has been Dr. Johnston's experience that AHI for obese patients (say, BMI > 35) is not much different from that of the non-obese. To validate his assertion Dr. Johnston took a sample of 15 obese patients and measured their AHI. The sample mean (\bar{x}) and variance (s^2) of his measurements are 9.77 and 11.14, respectively. State the hypotheses and perform a test at the significance level $\alpha = 0.05$.⁷

Exercise. 5.2

Coronary heart disease is the leading cause of morbidity and mortality throughout the world. A research group investigated the possible association between periodontal health and coronary heart disease in patients with acute myocardial infarction (AMI) and chronic coronary heart disease (CCHD). Based on the physical examination done prior to admission into a clinical trial, it was determined that the average level of triglycerides for the patients with AMI is 196.49 mg/dl. Suppose the investigators claimed that the triglyceride level for the patients with the CCHD should be lower than that value for the patients in the AMI group. To confirm their claim, the investigators randomly selected 27 patients with CCHD and observed their triglyceride levels. The data yielded $\bar{x} = 155.23$ mg/dl and $s = 88.04$ mg/dl. Assume that the triglyceride levels are normally distributed. How would you state the hypotheses for the investigators, and what could you conclude? ⁸

Exercise. 5.3

Body mass index is calculated by dividing a person's weight by the square of his or her height; it is a measure of the extent to which the individual is overweight. For the population of middle-aged men who later develop diabetes mellitus, the distribution of baseline body mass indices is approximately normal with an unknown mean μ and standard deviation σ . A sample of 58 men selected from this group has mean $\bar{x} = 25.0$ kg/m² and standard deviation $s = 2.7$ kg/m².

1. Construct a 95% confidence interval for the population mean μ .
2. At the 0.05 level of significance, test whether the mean baseline body mass index for the population of middle-aged men who do develop diabetes is equal to 24.0 kg/m², the mean for the population of men who do not. What is the p -value of the test?
3. What do you conclude?
4. Based on the 95% confidence interval, would you have expected to reject or not to reject the null hypothesis? Why?

Exercise. 5.4

The population of male industrial workers in London who have never experienced a major coronary event has mean systolic blood pressure 136 mmHg and mean diastolic blood pressure 84 mmHg. You might be interested in determining whether these values are the same as those for the population of industrial workers who have suffered a coronary event.

⁷ibidem

⁸ibidem

1. A sample of 86 workers who have experienced a major coronary event has mean systolic blood pressure $\bar{x}=143$ mmHg and standard deviation $s=24.4$ mmHg. Test the null hypothesis that the mean systolic blood pressure for the population of industrial workers who have experienced such an event is identical to the mean for the workers who have not, using a two-sided test at the $\alpha=0.10$ level.
2. The same sample of men has mean diastolic blood pressure $\bar{x}=87$ mmHg and standard deviation $s=16.0$ mmHg. Test the null hypothesis that the mean systolic blood pressure for the population of workers who have experienced a major coronary event is identical to the mean for the workers who have not.
3. How do the two groups of workers compare?

Exercise. 5.5

Suppose a survey with practicing Medicinae Doctors (MDs) showed about 43% have a solo practice and the remainder 57% have a partnership, association, are employed with HMO or government, etc. Upon reading this survey report, a researcher was curious about the proportion of veterinary physician who have a solo practice. She has randomly selected 500 veterinary physician listed in a directory and mailed a questionnaire to them. By the end of the month she had received 223 responses. The sample proportion of a solo practice based on these responses was $\hat{p} = 0.32$. Is the proportion of a solo practice among veterinarian physicians different from that among MDs?

1. State the hypotheses.
2. Test the hypothesis at the significance level $\alpha = 0.05$.

Exercise. 5.6

In a random sample of 11 race horses in a Physical Condition Test the strength of the lateral digital extensor muscle was measured. The following values in Kg were got from the sample

58, 62, 64, 67, 69, 70, 72, 73, 73, 75, 80

Assuming that the muscle strength is normally distributed, test whether the population mean is different from 65 with a significance $\alpha = 0.05$. Calculate the value of the test statistics, the acceptance interval and draw your conclusions.

Exercise. 5.7

A study is being developed about the percentage of wild cats infected by toxoplasma (toxoplasmosis). It is known that the percentage of infected cats in Europe is 25%. A random sample of 125 wild cats has been selected to perform the study. In this sample, a blood analysis has concluded that 35 cats have suffered from this disease. What conclusion can be drawn from this study with respect to the percentage of infected cats in Spain and in Europe? Before answering set the adequate hypothesis testing with a significance of $\alpha = 0.05$ and calculate the p-value.

Exercise. 5.8

We are interested in the height of Arabian horses fed with an experimental diet. This horse breed has an average height of 160 cm. With the objective of assessing whether the horses get the same average height, we have selected 15 horses fed with this new type of diet and the mean and standard deviation of the sample were 162.5 cm and 5 cm respectively. Assume that the horse height is normally distributed. Can you conclude that the mean height of the horses with this diet is higher than 160 cm? Before answering, do the hypothesis testing with a significance of $\alpha = 0.05$, calculate the acceptance interval and draw your conclusions.

Exercise. 5.9

In a sample of 1500 dogs from a neighbour of a city, 125 blood test were positive about macrocytic anemia Do these data provide enough evidence that the proportion of dogs with this disease is larger than

6%? Before answering this question, explain the null and alternative hypothesis and perform the test with a significance level $\alpha = 0.05$ and calculate the p-value of this test.

Exercise. 5.10

A group of veterinarians is trying to determine whether the chloride level in the water in a agricultural area with a large number of pork farms is adequate. The ideal level is 325 units. A sample of water of 150 taps coming from random selected farms was collected by the team. The mean and the standard deviation is 332 and 52, respectively. Do the veterinarian group state whether the chloride level is adequate or not? Explain your reasons by means of a hypothesis testing.

Chapter 6

Two-sample Hypothesis testing

6.1 Inference for the difference of two proportions

6.2 Two-sample test for paired data

6.3 Testing the equality of two variances

6.4 Two-sample test for independent data with identical variances

6.5 Two-sample test for independent data with non-identical variances

6.6 Notes

Chapter 4.1 and 5 introduced a framework for statistical inference based on confidence intervals and hypotheses. In this chapter, we encounter several new point estimates and scenarios. In contrast with the previous situation, there is not a reference value, but a comparison between two groups. In each case, the inference ideas remain the same:

1. Determine which point estimate or test statistic is useful.
 2. Identify an appropriate distribution for the point estimate or test statistic.
-



For labs, slides, and other resources, please visit
www.openintro.org/book/biostat

6.1 Inference for the difference of two proportions

Just as inference can be done for the difference of two population means, conclusions can also be drawn about the difference of two population proportions: $p_1 - p_2$.

6.1.1 Sampling distribution of the difference of two proportions

The normal model can be applied to $\hat{p}_1 - \hat{p}_2$ if the sampling distribution for each sample proportion is nearly normal and if the samples are independent random samples from the relevant populations.

CONDITIONS FOR THE SAMPLING DISTRIBUTION OF $\hat{p}_1 - \hat{p}_2$ TO BE APPROXIMATELY NORMAL

The difference $\hat{p}_1 - \hat{p}_2$ tends to follow a normal model when

- each of the two samples are random samples from a population,
- the two samples are independent of each other, and
- each sample proportion follows (approximately) a normal model. This condition is satisfied when $n_1 p_1, n_1(1 - p_1), n_2 p_2$ and $n_2(1 - p_2)$ are all ≥ 10 .

The standard error of the difference in sample proportions is

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2} = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}, \quad (6.1)$$

where p_1 and p_2 are the population proportions, and n_1 and n_2 are the two sample sizes.

6.1.2 Confidence intervals for $p_1 - p_2$

When calculating confidence intervals for a difference of two proportions using the normal approximation to the binomial, the two sample proportions are used to verify the success-failure condition and to compute the standard error.

EXAMPLE 6.2

The way a question is phrased can influence a person's response. For example, Pew Research Center conducted a survey with the following question:¹

As you may know, by 2014 nearly all Americans will be required to have health insurance. [People who do not buy insurance will pay a penalty] while [People who cannot afford it will receive financial help from the government]. Do you approve or disapprove of this policy?

For each randomly sampled respondent, the statements in brackets were randomized: either they were kept in the order given above, or the order of the two statements was reversed. Figure 6.1 shows the results of this experiment. Calculate and interpret a 90% confidence interval of the difference in the probability of approval of the policy.

First the conditions for the use of a normal model must be verified. The Pew Research Center uses sampling methods that produce random samples of the US population (at least approximately) and because each group was a simple random sample from less than 10% of the population, the observations are independent, both within the samples and between the samples. The success-failure condition also holds for each sample, so the normal model can be used for confidence intervals for the difference in approval proportions. The point estimate of the difference in support, where \hat{p}_1 corresponds to the original ordering and \hat{p}_2 to the reversed ordering:

$$\hat{p}_1 - \hat{p}_2 = 0.47 - 0.34 = 0.13.$$

The standard error can be computed from Equation (6.1) using the sample proportions:

$$SE \approx \sqrt{\frac{0.47(1-0.47)}{771} + \frac{0.34(1-0.34)}{732}} = 0.025.$$

For a 90% confidence interval, $z^* = 1.65$:

$$\text{point estimate} \pm z^* \times SE \rightarrow 0.13 \pm 1.65 \times 0.025 \rightarrow (0.09, 0.17).$$

With 90% confidence, the proportion approving the 2010 health care law ranged between 9% and 17% depending on the phrasing of the question. The Pew Research Center interpreted this modestly large difference as an indication that for most of the public, opinions were still fluid on the health insurance mandate. The law eventually passed as the Affordable Health Care Act (ACA).

	Sample size (n_i)	Approve (%)	Disapprove (%)	Other
Original ordering	771	47	49	3
Reversed ordering	732	34	63	3

Figure 6.1: Results for a Pew Research Center poll where the ordering of two statements in a question regarding healthcare were randomized.

6.1.3 Hypothesis testing for $p_1 - p_2$

Hypothesis tests for $p_1 - p_2$ are usually testing the null hypothesis of no difference between p_1 and p_2 ; i.e. $H_0 : p_1 - p_2 = 0$. Under the null hypothesis, $\hat{p}_1 - \hat{p}_2$ is normally distributed with mean 0

¹www.people-press.org/2012/03/26/public-remains-split-on-health-care-bill-opposed-to-mandate. Sample sizes for each polling group are approximate.

and standard deviation $\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}$, where under the null hypothesis $p = p_1 = p_2$.

Since p is unknown, an estimate is used to compute the standard error of $\hat{p}_1 - \hat{p}_2$; p can be estimated by \hat{p} , the weighted average of the sample proportions \hat{p}_1 and \hat{p}_2 :

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2},$$

where x_1 is the number of observed events in the first sample and x_2 is the number of observed events in the second sample. This **pooled proportion** \hat{p} is also used to check the success-failure condition.

The test statistic z for testing $H_0 : p_1 = p_2$ versus $H_A : p_1 \neq p_2$ equals:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}.$$

EXAMPLE 6.3

We investigate whether there is an increased risk of cancer in dogs that are exposed to the herbicide 2,4-dichlorophenoxyacetic acid (2,4-D). A study in 1994 examined 491 dogs that had developed cancer and 945 dogs as a control group.^a Of these two groups, researchers identified which dogs had been exposed to 2,4-D in their owner's yard. The results are shown in Table 6.2.

Is this study an experiment or an observational study? The owners were not instructed to apply or not apply the herbicide, so this is an observational study. This question was especially tricky because one group was called the *control group*, which is a term usually seen in experiments.

Set up hypotheses to test whether 2,4-D and the occurrence of cancer in dogs are related. Use a one-sided test and compare across the cancer and no cancer groups. Using the proportions within the cancer and no cancer groups may seem odd. We intuitively may desire to compare the fraction of dogs with cancer in the 2,4-D and no 2,4-D groups, since the herbicide is an explanatory variable. However, the cancer rates in each group do not necessarily reflect the cancer rates in reality due to the way the data were collected. For this reason, computing cancer rates may greatly alarm dog owners.

H_0 : the proportion of dogs with exposure to 2,4-D is the same in "cancer" and "no cancer" dogs, $p_c - p_n = 0$.

H_A : dogs with cancer are more likely to have been exposed to 2,4-D than dogs without cancer, $p_c - p_n > 0$.

Under the assumption of independence, we can use the normal model and make statements regarding the canine population based on the data.

The point estimate of the difference in sample proportions is $\hat{p}_c - \hat{p}_n = 0.067$. First, we compute the pooled proportion:

$$\hat{p} = \frac{\# \text{ of "successes" }}{\# \text{ of cases }} = \frac{191 + 304}{191 + 300 + 304 + 641} = 0.345$$

The estimate for the standard error is $SE = 0.026$. Compute the test statistic:

$$z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.067 - 0}{0.026} = 2.58$$

We leave the picture to you. Looking up $z = 2.58$ in the normal probability table: 0.9951. However this is the lower tail, and the upper tail represents the p-value: $1 - 0.9951 = 0.0049$. We reject the null hypothesis and conclude that dogs getting cancer and owners using 2,4-D are associated.

^aHayes HM, Tarone RE, Cantor KP, Jessen CR, McCurnin DM, and Richardson RC. 1991. Case-Control Study of Canine Malignant Lymphoma: Positive Association With Dog Owner's Use of 2, 4-Dichlorophenoxyacetic Acid Herbicides. Journal of the National Cancer Institute 83(17):1226-1231.

	cancer	no cancer
2,4-D	191	304
no 2,4-D	300	641

Table 6.2: Summary results for cancer in dogs and the use of 2,4-D by the dog's owner.

EXAMPLE 6.4

The use of screening mammograms for breast cancer has been controversial for decades because the overall benefit on breast cancer mortality is uncertain. Several large randomized studies have been conducted in an attempt to estimate the effect of mammogram screening. A 30-year study to investigate the effectiveness of mammograms versus a standard non-mammogram breast cancer exam was conducted in Canada with 89,835 female participants.² During a 5-year screening period, each woman was randomized to either receive annual mammograms or standard physical exams for breast cancer. During the 25 years following the screening period, each woman was screened for breast cancer according to the standard of care at her health care center.

At the end of the 25 year follow-up period, 1,005 women died from breast cancer. The results by intervention are summarized in Figure 6.3.

Assess whether the normal model can be used to analyze the study results.

E

Since the participants were randomly assigned to each group, the groups can be treated as independent, and it is reasonable to assume independence of patients within each group. Participants in randomized studies are rarely random samples from a population, but the investigators in the Canadian trial recruited participants using a general publicity campaign, by sending personal invitation letters to women identified from general population lists, and through contacting family doctors. In this study, the participants can reasonably be thought of as a random sample.

The pooled proportion \hat{p} is

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{500 + 505}{500 + 44,425 + 505 + 44,405} = 0.0112.$$

Checking the success-failure condition for each group:

$$\begin{aligned} \hat{p} \times n_{mgm} &= 0.0112 \times 44,925 = 503 & (1 - \hat{p}) \times n_{mgm} &= 0.9888 \times 44,925 = 44,422 \\ \hat{p} \times n_{ctrl} &= 0.0112 \times 44,910 = 503 & (1 - \hat{p}) \times n_{ctrl} &= 0.9888 \times 44,910 = 44,407 \end{aligned}$$

All values are at least 10.

The normal model can be used to analyze the study results.

	Death from breast cancer?	
	Yes	No
Mammogram	500	44,425
Control	505	44,405

Figure 6.3: Summary results for the mammogram study.

²Miller AB. 2014. *Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial*. BMJ 2014;348:g366 doi: 10.1136/bmj.g366

EXAMPLE 6.5

Do the results from the study provide convincing evidence of a difference in the proportion of breast cancer deaths between women who had annual mammograms during the screening period versus women who received annual screening with physical exams?

The null hypothesis is that the probability of a breast cancer death is the same for the women in the two groups. If group 1 represents the mammogram group and group 2 the control group, $H_0 : p_1 = p_2$ and $H_A : p_1 \neq p_2$. Let $\alpha = 0.05$.

Calculate the test statistic z :

$$z = \frac{0.01113 - 0.01125}{\sqrt{(0.0112)(1 - 0.0112)\left(\frac{1}{44,925} + \frac{1}{44,910}\right)}} = -0.17.$$

E

The two-sided p -value is $P|Z| \geq 0.17 = 0.8650$, which is greater than 0.05. There is insufficient evidence to reject the null hypothesis; the observed difference in breast cancer death rates is reasonably explained by chance.

Evaluating medical treatments typically requires accounting for additional evidence that cannot be evaluated from a statistical test. For example, if mammograms are much more expensive than a standard screening and do not offer clear benefits, there is reason to recommend standard screenings over mammograms. This study also found that a higher proportion of diagnosed breast cancer cases in the mammogram screening arm (3250 in the mammogram group vs 3133 in the physical exam group), despite the nearly equal number of breast cancer deaths. The investigators inferred that mammograms may cause over-diagnosis of breast cancer, a phenomenon in which a breast cancer diagnosed with mammogram and subsequent biopsy may never become symptomatic. The possibility of over-diagnosis is one of the reasons mammogram screening remains controversial.

EXAMPLE 6.6

Calculate a 95% confidence interval for the difference in proportions of deaths from breast cancer from the Canadian study.

The independence and random sampling conditions have already been discussed. The success failure condition should be checked for each sample, since this is not a hypothesis testing context (i.e., there is no null hypothesis). For the mammogram group, $\hat{p}_1 = 0.01113$; $n_1\hat{p}_1 = (0.01113)(44,925) = 500$ and $n_1(1 - \hat{p}_1) = 39,925$. It is easy to show that the success failure condition is holds for the control group as well.

The point estimate for the difference in the probability of death is

$$\hat{p}_1 - \hat{p}_2 = 0.01113 - 0.01125 = -0.00012,$$

or 0.012%.

The standard error for the estimated difference uses the individual estimates of the probability of a death:

$$SE \approx \sqrt{\frac{0.01113(1 - 0.01113)}{44,925} + \frac{0.01125(1 - 0.01125)}{44,910}} = 0.0007.$$

The 95% confidence interval is given by

$$-0.00012 \pm (1.96)(0.0007) = (-0.0015, 0.0013).$$

With 95% confidence, the difference in the probability of death is between -0.15% and 0.13%. As expected from the large p -value, the confidence interval contains the null value 0.

E

6.2 Two-sample test for paired data

In the 2000 Olympics, was the use of a new wetsuit design responsible for an observed increase in swim velocities? In a study designed to investigate this question, twelve competitive swimmers swam 1500 meters at maximal speed, once wearing a wetsuit and once wearing a regular swimsuit.³ The order of wetsuit versus swimsuit was randomized for each of the 12 swimmers. Figure 6.4 shows the average velocity recorded for each swimmer, measured in meters per second (m/s).⁴

	swimmer.number	wet.suit.velocity	swim.suit.velocity	velocity.diff
1	1	1.57	1.49	0.08
2	2	1.47	1.37	0.10
3	3	1.42	1.35	0.07
4	4	1.35	1.27	0.08
5	5	1.22	1.12	0.10
6	6	1.75	1.64	0.11
7	7	1.64	1.59	0.05
8	8	1.57	1.52	0.05
9	9	1.56	1.50	0.06
10	10	1.53	1.45	0.08
11	11	1.49	1.44	0.05
12	12	1.51	1.41	0.10

Figure 6.4: Paired Swim Suit Data

The swimsuit velocity data are an example of **paired data**, in which two sets of observations are uniquely paired so that an observation in one set matches an observation in the other; in this case, each swimmer has two measured velocities, one with a wetsuit and one with a swimsuit. A natural measure of the effect of the wetsuit on swim velocity is the difference between the measured maximum velocities ($\text{velocity.diff} = \text{wet.suit.velocity} - \text{swim.suit.velocity}$). Even though there are two measurements per swimmer, using the difference in velocities as the variable of interest allows for the problem to be approached like those in Section 5.1. Although it was not explicitly noted, the data used in Section 5.1.1 were paired; each respondent had both an actual and desired weight.

Suppose the parameter δ is the population average of the difference in maximum velocities during a 1500m swim if all competitive swimmers recorded swim velocities with each suit type. A hypothesis test can then be conducted with the null hypothesis that the mean population difference in swim velocities between suit types equals 0 (i.e., there is no difference in population average swim velocities), $H_0 : \delta = 0$, against the alternative that the difference is non-zero, $H_A : \delta \neq 0$.

STATING HYPOTHESES FOR PAIRED DATA

When testing a hypothesis about paired data, compare the groups by testing whether the population mean of the differences between the groups equals 0.

- For a two-sided test, $H_0 : \delta = 0$; $H_A : \delta \neq 0$.
- For a one-sided test, either $H_0 : \delta = 0$; $H_A : \delta > 0$ or $H_0 : \delta = 0$; $H_A : \delta < 0$.

³De Lucas et. al, The effects of wetsuits on physiological and biomechanical indices during swimming. *Journal of Science and Medicine in Sport*, 2000; 3(1): 1-8

⁴The data are available as swim in the oibioestat R package. The data are also used in Lock et. al *Statistics, Unlocking the Power of Data*, Wiley, 2013.

Some important assumptions are being made. First, it is assumed that the data are a random sample from the population. While the observations are likely independent, it is more difficult to justify that this sample of 12 swimmers is randomly drawn from the entire population of competitive swimmers. Nevertheless, it is often assumed in problems such as these that the participants are reasonably representative of competitive swimmers. Second, it is assumed that the population of differences is normally distributed. This is a small sample, one in which normality would be difficult to confirm. The dot plot for the difference in velocities in Figure 6.5 shows approximate symmetry.

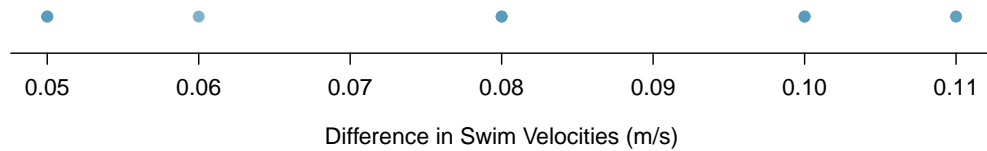


Figure 6.5: A dot plot of differences in swim velocities.

Let \bar{x}_{diff} denote the sample average of the differences in maximum velocity, s_{diff} the sample standard deviation of the differences, and n the number of pairs in the dataset. The t -statistic used to test H_0 vs. H_A is:

$$\frac{\bar{x}_{\text{diff}} - \delta_0}{s_{\text{diff}}/\sqrt{n}},$$

where in this case $\delta_0 = 0$.⁵

EXAMPLE 6.7

Using the data in Figure 6.4, conduct a two-sided hypothesis test at $\alpha = 0.05$ to assess whether there is evidence to suggest that wetsuits have an effect on swim velocities during a 1500m swim.

The hypotheses are $H_0 : \delta = 0$ and $H_A : \delta \neq 0$. Let $\alpha = 0.05$.

Calculate the t -statistic:

$$t = \frac{\bar{x}_{\text{diff}} - \delta_0}{s_{\text{diff}}/\sqrt{n}} = \frac{0.078 - 0}{0.022/\sqrt{12}} = 12.32$$

The two-sided p -value is

$$p = P(T < -12.32) + P(T > 12.32),$$

where t has a t -distribution with $n - 1 = 11$ degrees of freedom. The t -table shows that $p < 0.01$. Software can be used to show that $p = 2.3 \times 10^{-7}$, a very small value indeed.

The data support the claim that the wetsuits changed swim velocity in a 1500m swim. The observed average increase of 0.078 m/s is significantly different than the null hypothesis of no change, and suggests that swim velocities are higher when swimmers wear wetsuits as opposed to swimsuits.

Calculating confidence intervals for paired data is also based on the differences between the values in each pair; the same approach as for single-sample data can be applied on the differences. For example, a two-sided 95% confidence interval for paired data has the form:

$$\left(\bar{x}_{\text{diff}} - t_{df}^* \times \frac{s_{\text{diff}}}{\sqrt{n}}, \bar{x}_{\text{diff}} + t_{df}^* \times \frac{s_{\text{diff}}}{\sqrt{n}} \right),$$

where t^* is the point on a t -distribution with $df = n - 1$ for n pairs, with area 0.025 to its right.

⁵This value is specified by the null hypothesis of no difference.

GUIDED PRACTICE 6.8**G**

Using the data in Figure 6.4, calculate a 95% confidence interval for the average difference in swim velocities during a 1500m swim. Is the interval consistent with the results of the hypothesis test?⁶

The general approach when analyzing paired data is to first calculate the differences between the values in each pair, then use those differences in methods for confidence intervals and tests for a single sample. Any conclusion from an analysis should be stated in terms of the original paired measurements.

⁶Use the values of \bar{x}_{diff} and s_{diff} as calculated previously: 0.078 and 0.022. The t^* value of 2.20 has $df = 11$ and 0.025 area to the right. The confidence interval is $(0.078 \pm \frac{0.022}{\sqrt{12}}) \rightarrow (0.064, 0.091)$ m/s. With 95% confidence, δ lies between 0.064 m/s and 0.09 m/s. The interval does not include 0 (no change), which is consistent with the result of the hypothesis test.

6.3 Testing the equality of two variances

The difference of the variances between two groups is frequently used to study the quality of two different procedures or treatment. Less variability is associated to greater quality. For this course, another use is to confirm that the variances of a continuous variable for two groups are identical. This is going to be a previous step before studying hypothesis testing of the mean of independent sample. Later, we are going to distinguish between two-sample test for independent data with non-identical variances (6.4) and with identical variances (6.5).

Given two normally distributed random variables, two sample are extracted. Let s_1^2 be the variance of a sample with n_1 following a normal distribution mean μ_1 and standard deviation σ_1 and let s_2^2 be the variance of a sample with n_2 following a normal distribution mean μ_2 and standard deviation σ_2 . The quotient

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

follows a distribution called Snedecor's F distribution with a numerator with degree of freedom $n_1 - 1$ and a denominator with degree of freedom $n_2 - 1$. In the Figure 6.6 several probability density function are displayed.

A table with 95 and 97.5 percentiles of the Snedecor's F distribution can be found on the Appendix B. The 5 and 2.5 percentiles can be found by means of the following formula, based on interchanging the roles of the numerator and denominator:

$$F_{n_1, n_2, \alpha} = \frac{1}{F_{n_2, n_1, 1-\alpha}}$$

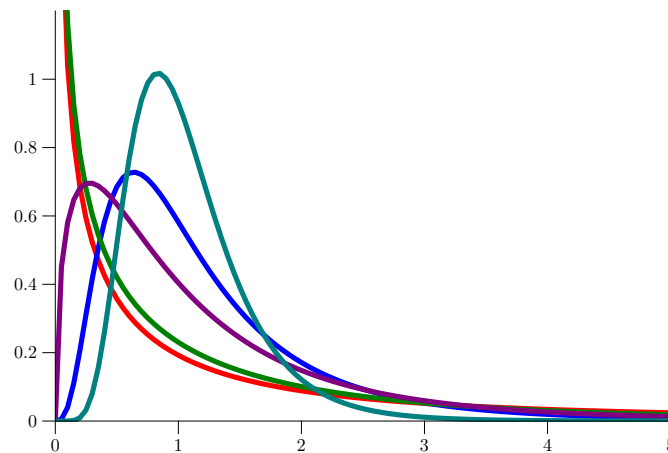


Figure 6.6: Snedecor's F distributions. Several probability density functions are displayed. The red curve corresponds to degree of freedom 1 in the numerator and 2 in the denominator. The blue curve corresponds to degree of freedom 9 in the numerator and 9 in the denominator. The red curve corresponds to degree of freedom 1 in the numerator and 2 in the denominator. The green curve corresponds to degree of freedom 1 in the numerator and 10 in the denominator. The dark green curve corresponds to degree of freedom 20 in the numerator and 25 in the denominator.

EXAMPLE 6.9

The time of a surgical intervention in horses using two different techniques is being estimated to improve the quality of the treatment. It's assume that the time of surgical intervention is a normal distribution for both of the techniques. The sample variance is $s_1 = 50$ for 31 horses which technique was 1 and $s_2 = 24$ for 25 horses operated with the technique 2. Test whether the variance of these variable is significantly different.

The null-hypothesis is

$$H_0 : \sigma_A = \sigma_B$$

$$H_A : \sigma_A \neq \sigma_B$$

and the sample size are $n_A = 31$ and $n_B = 25$, and the variances $s_A^2 = 50$ and $s_B^2 = 24$.

The test statistics is $F = s_A^2/s_B^2 = 50/24 = 2.08$

The acceptance region is $(F_{30,24,.025}, F_{30,24,.975}) = (\frac{1}{F_{24,30,.975}}, F_{30,24,.975}) = (\frac{1}{2.1359}, 2.2090) = (0.4681, 2.2090)$.

The value of the test statistic is inside a acceptance region and the null hypothesis is failed to reject. The study does not provide enough evidence to say that the two groups have different variances.

6.4 Two-sample test for independent data (identical variances)

In this section we consider a difference in two population means, $\mu_1 - \mu_2$, under the condition that the data are not paired. If the data are not paired and there is not a natural link between the individuals in the two groups, it is said to be **independent samples**. In this section, the variances are assumed to be identical. This assumption can be considered reasonable if we fail to reject the null hypothesis of the test of two variances, which we have studied on the previous section. When variances cannot be assumed identical, the test statistic is different and it will be studied on the next section. The paired and independent samples methods are similar in theory but different in the details. Just as with a single sample, we identify conditions to ensure a point estimate of the difference $\bar{x}_1 - \bar{x}_2$ is nearly normal. Next we introduce a formula for the standard error.

6.4.1 Pooled standard deviation estimate (if variances are identical for both groups)

Occasionally, two populations will have standard deviations that are so similar that they can be treated as identical. For example, historical data or a well-understood biological mechanism may justify this strong assumption. In such cases, we can make our t distribution approach slightly more precise by using a pooled standard deviation.

The **pooled standard deviation** of two groups is a way to use data from both samples to better estimate the standard deviation and standard error. If s_1 and s_2 are the standard deviations of groups 1 and 2 and there are good reasons to believe that the population standard deviations are equal, then we can obtain an improved estimate of the group variances by pooling their data:

$$s_{pooled}^2 = \frac{s_1^2 \times (n_1 - 1) + s_2^2 \times (n_2 - 1)}{n_1 + n_2 - 2}$$

where n_1 and n_2 are the sample sizes, as before. To use this new statistic, we substitute s_{pooled}^2 in place of s_1^2 and s_2^2 in the standard error formula, and we use an updated formula for the degrees of freedom:

$$df = n_1 + n_2 - 2$$

The benefits of pooling the standard deviation are realized through obtaining a better estimate of the standard deviation for each group and using a larger degrees of freedom parameter for the t distribution. Both of these changes may permit a more accurate model of the sampling distribution of $\bar{x}_1 - \bar{x}_2$.

Pooling standard deviations should be done only after careful research. A pooled standard deviation is only appropriate when background research indicates the population standard deviations are nearly equal. When the sample size is large and the condition may be adequately checked with data, the benefits of pooling the standard deviations greatly diminishes.

6.4.2 Hypothesis testing with equal variance

Assume that the variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$ are equal but unknown in both groups. The sample pool variance is considered the point estimate of the variance for both groups. Therefore, the standard error is the pool standard error multiplied by the square root of $1/n_1$ plus $1/n_2$.

In order to test whether the mean population are identical in both groups, the following t statistic is defined

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

If the mean population are equal in both group, this test statistic follows a Student's t distribution with degree of freedom $n_1 + n_2 - 2$.

EXAMPLE 6.10

Two simple random samples are selected from the adult male population of Beijing and Shanghai, respectively. The variable height (cm) of the individuals in each sample has the following data:

Beijing	$n_B = 10$	$\bar{x}_B = 169$	$s_B^2 = 1800$
Shanghai	$n_S = 21$	$\bar{x}_S = 171$	$s_S^2 = 1400$

Is there a significant difference between the height of male population?

First, test whether the groups have identical variance $\sigma_B = \sigma_S$:

The statistic $F = \frac{1800}{1400} = 1.2857$ and the acceptance region with numeratos 10 and 21 is $(F_{9,20,.025}, F_{9,20,.975}) = (0.27, 2.84)$ Therefore, it is failed to reject that the variances are different. Let us assume that the variances are identical.

Second, test if the mean are identical in both groups.

(E)

The pooled variance (s_{pooled}^2) in this case is

$$S_{\text{pooled}}^2 = \frac{(n_B - 1)S_B^2 + (n_S - 1)S_S^2}{n_B + n_S - 2} = \frac{9 \cdot 1800 + 20 \cdot 1400}{29} = 1524.138$$

The test statistic is

$$t = \frac{(\bar{x}_B - \bar{x}_S)}{S_p \sqrt{\frac{1}{n_B} + \frac{1}{n_S}}} = \frac{169 - 171}{\sqrt{1524.138} \sqrt{\frac{1}{10} + \frac{1}{21}}} = -0.1333358$$

The acceptance region is $(-t_{29,.975}, t_{29,.975}) = (-2.04523, 2.04523)$.

We fail to reject the null hypothesis. The difference between the adult male people in both city is not statistically significant.

6.5 Two-sample test for independent data (non-identical variances)

In this section, we study the situation when identical variances cannot be assumed.

Does treatment using embryonic stem cells (ESCs) help improve heart function following a heart attack? New and potentially risky treatments are sometimes tested in animals before studies in humans are conducted. In a 2005 paper in *Lancet*, Menard, et al. describe an experiment in which 18 sheep with induced heart attacks were randomly assigned to receive cell transplants containing either ESCs or inert material.⁷ Various measures of cardiac function were measured 1 month after the transplant.

This design is typical of an intervention study. The analysis of such an experiment is an example of drawing inference about the difference in two population means, $\mu_1 - \mu_2$, when the data are independent, i.e., not paired. The point estimate of the difference, $\bar{x}_1 - \bar{x}_2$, is used to calculate a t -statistic that is the basis of confidence intervals and tests.

6.5.1 Confidence interval for a difference of means

Figure 6.7 contains summary statistics for the 18 sheep.⁸ Percent change in heart pumping capacity was measured for each sheep. A positive value corresponds to increased pumping capacity, which generally suggests a stronger recovery from the heart attack. Is there evidence for a potential treatment effect of administering stem cells?

	n	\bar{x}	s
ESCs	9	3.50	5.17
control	9	-4.33	2.76

Figure 6.7: Summary statistics of the embryonic stem cell study.

Figure 6.8 shows that the distributions of percent change do not have any prominent outliers, which would indicate a deviation from normality; this suggests that each sample mean can be modeled using a t -distribution. Additionally, the sheep in the study are independent of each other, and the sheep between groups are also independent. Thus, the t -distribution can be used to model the difference of the two sample means.

USING THE t -DISTRIBUTION FOR A DIFFERENCE IN MEANS

The t -distribution can be used for inference when working with the standardized difference of two means if (1) each sample meets the conditions for using the t -distribution and (2) the samples are independent.

A confidence interval for a difference of two means has the same basic structure as previously discussed confidence intervals:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{df}^* \times SE.$$

The following formula is used to calculate the standard error of $\bar{x}_1 - \bar{x}_2$. Since σ is typically unknown, the standard error is estimated by using s in place of σ .

⁷Menard C, et al., Transplantation of cardiac-committed mouse embryonic stem cells to infarcted sheep myocardium: a preclinical 2005; 366:1005-12, doi [https://doi.org/10.1016/S0140-6736\(05\)67380-1](https://doi.org/10.1016/S0140-6736(05)67380-1)

⁸The data are accessible as the dataset `stem.cells` in the `openintro` R package.

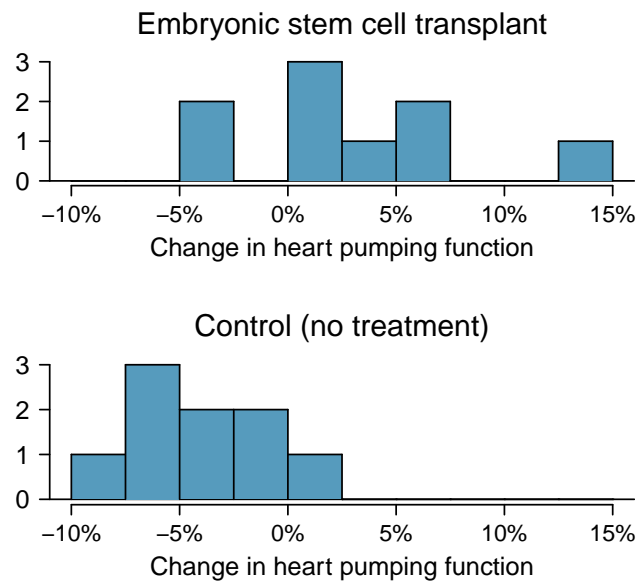


Figure 6.8: Histograms for both the embryonic stem cell group and the control group. Higher values are associated with greater improvement.

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

In this setting, the t -distribution has a somewhat complicated formula for the degrees of freedom that is usually calculated with software.⁹ An alternative approach uses the smaller of $n_1 - 1$ and $n_2 - 1$ as the degrees of freedom.¹⁰

DISTRIBUTION OF A DIFFERENCE OF SAMPLE MEANS

The sample difference of two means, $\bar{x}_1 - \bar{x}_2$, can be modeled using the t -distribution and the standard error

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (6.11)$$

when each sample mean can itself be modeled using a t -distribution and the samples are independent. To calculate the degrees of freedom without using software, use the smaller of $n_1 - 1$ and $n_2 - 1$.

⁹See Section 6.6 for the formula.

¹⁰This technique for degrees of freedom is conservative with respect to a Type 1 Error; it is more difficult to reject the null hypothesis using this approach for degrees of freedom.

EXAMPLE 6.12

Calculate and interpret a 95% confidence interval for the effect of ESCs on the change in heart pumping capacity of sheep following a heart attack.

The point estimate for the difference is $\bar{x}_1 - \bar{x}_2 = \bar{x}_{\text{esc}} - \bar{x}_{\text{control}} = 7.83$.

The standard error is:

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{5.17^2}{9} + \frac{2.76^2}{9}} = 1.95.$$

Since $n_1 = n_2 = 9$, use $df = 8$; $t_8^* = 2.31$ for a 95% confidence interval. Alternatively, computer software can provide more accurate values: $df = 12.225$, $t^* = 2.174$.

The confidence interval is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{df}^* \times SE \rightarrow 7.83 \pm 2.31 \times 1.95 \rightarrow (3.38, 12.38).$$

With 95% confidence, the average amount that ESCs improve heart pumping capacity lies between 3.38% to 12.38%.¹¹ The data provide evidence for a treatment effect of administering stem cells.

6.5.2 Hypothesis tests for a difference in means

Is there evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who do not smoke? The dataset `births` contains data from a random sample of 150 cases of mothers and their newborns in North Carolina over a year; there are 50 cases in the smoking group and 100 cases in the nonsmoking group.¹²

	fAge	mAge	weeks	weight	sexBaby	smoke
1	NA	13	37	5.00	female	nonsmoker
2	NA	14	36	5.88	female	nonsmoker
3	19	15	41	8.13	male	smoker
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	
150	45	50	36	9.25	female	nonsmoker

Figure 6.9: Four cases from the `births` dataset.

EXAMPLE 6.13

Evaluate whether it is appropriate to apply the t -distribution to the difference in sample means between the two groups.

Since the data come from a simple random sample and consist of less than 10% of all such cases, the observations are independent. While each distribution is strongly skewed, the large sample sizes of 50 and 100 allow for the use of the t -distribution to model each mean separately. Thus, the difference in sample means may be modeled using a t -distribution.

A hypothesis test can be conducted to evaluate whether there is a relationship between mother's smoking status and average newborn birth weight. The null hypothesis represents the case of no

¹¹From software, the confidence interval is (3.58, 12.08).

¹²This dataset is available in the `openintro` R package.

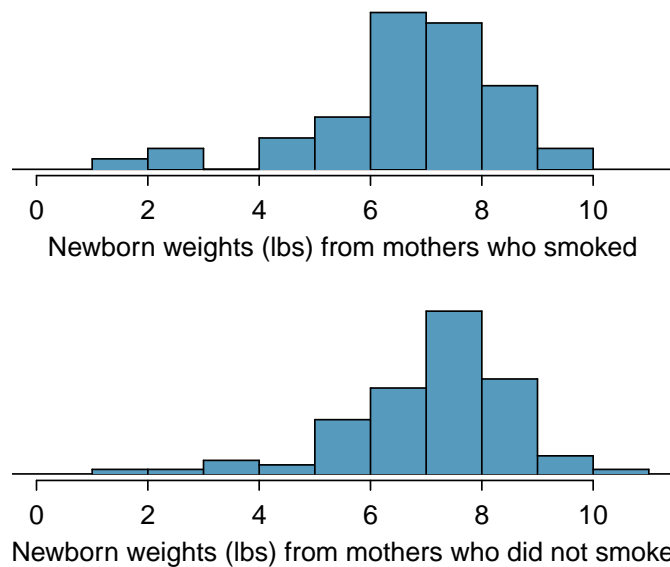


Figure 6.10: The top panel represents birth weights for infants whose mothers smoked. The bottom panel represents the birth weights for infants whose mothers who did not smoke. The distributions exhibit moderate-to-strong and strong skew, respectively.

difference between the groups, $H_0 : \mu_{ns} - \mu_s = 0$, where μ_{ns} represents the population mean of newborn birthweight for infants with mothers who did not smoke, and μ_s represents mean newborn birthweight for infants with mothers who smoked. Under the alternative hypothesis, there is some difference in average newborn birth weight between the groups, $H_A : \mu_{ns} - \mu_s \neq 0$. The hypotheses can also be written as $H_0 : \mu_{ns} = \mu_s$ and $H_A : \mu_{ns} \neq \mu_s$.

STATING HYPOTHESES FOR TWO-GROUP DATA

When testing a hypothesis about two independent groups, directly compare the two population means and state hypotheses in terms of μ_1 and μ_2 .

- For a two-sided test, $H_0 : \mu_1 = \mu_2$; $H_A : \mu_1 \neq \mu_2$.
- For a one-sided test, either $H_0 : \mu_1 = \mu_2$; $H_A : \mu_1 > \mu_2$ or $H_0 : \mu_1 = \mu_2$; $H_A : \mu_1 < \mu_2$.

In this setting, the formula for a t -statistic is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

Under the null hypothesis of no difference between the groups, $H_0 : \mu_1 - \mu_2 = 0$, the formula simplifies to

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

EXAMPLE 6.14

Using Figure 6.11, conduct a hypothesis test to evaluate whether there is evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who do not smoke.

The hypotheses are $H_0 : \mu_1 = \mu_2$ and $H_A : \mu_1 \neq \mu_2$, where μ_1 represents the average newborn birth weight for nonsmoking mothers and μ_2 represents average newborn birth weight for mothers who smoke. Let $\alpha = 0.05$.

Calculate the t -statistic:

E

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{7.18 - 6.78}{\sqrt{\frac{1.60^2}{100} + \frac{1.43^2}{50}}} = 1.54.$$

Approximate the degrees of freedom as $50 - 1 = 49$. The t -score of 1.49 falls between the first and second columns in the $df = 49$ row of the t -table, so the two-sided p -value is between 0.10 and 0.20.¹³

This p -value is larger than the significance value, 0.05, so the null hypothesis is not rejected. There is insufficient evidence to state there is a difference in average birth weight of newborns from North Carolina mothers who did smoke during pregnancy and newborns from North Carolina mothers who did not smoke during pregnancy.

	smoker	nonsmoker
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

Figure 6.11: Summary statistics for the births dataset.

6.5.3 The paired test vs. independent group test

In the two-sample setting, students often find it difficult to determine whether a paired test or an independent group test should be used. The paired test applies only in situations where there is a natural pairing of observations between groups, such as in the swim data. Pairing can be obvious, such as the two measurements for each swimmer, or more subtle, such as measurements of respiratory function in twins, where one member of the twin pair is treated with an experimental treatment and the other with a control. In the case of two independent groups, there is no natural way to pair observations.

A common error is to overlook pairing in data and assume that two groups are independent. The swimsuit data can be used to illustrate the possible harm in conducting an independent group test rather than a paired test. In Section 6.2, the paired t -test showed a significant difference in the swim velocities between swimmers wearing wetsuits versus regular swimsuits. Suppose the analysis had been conducted without accounting for the fact that the measurements were paired.

The mean and standard deviation for the 12 wet suit velocities are 1.51 and 0.14 (m/sec),

¹³From R, $df = 89.277$ and $p = 0.138$.

respectively, and 1.43 and 0.14 (m/sec) for the 12 swim suit velocities. A two-group test statistic is:

$$t = \frac{1.52 - 1.43}{\sqrt{0.14^2/12 + 0.14^2/12}} = 1.37.$$

If the degrees of freedom are approximated as $11 = 12 - 1$, the two-sided p -value as calculated from software is 0.20. According to this method, the null hypothesis of equal mean velocities for the two suit types would not be rejected.

It is not difficult to show that the numerator of the paired test (the average of the within swimmer differences) and the numerator of the two-group test (the difference of the average times for the two groups) are identical. The values of the test statistics differ because the denominators are different—specifically, the standard errors associated with each statistic are different. For the paired test statistic, the standard error uses the standard deviation of the within pair differences (0.22) and has value $0.022/\sqrt{12} = 0.006$. The two-group test statistic combines the standard deviations for the original measurements and has value $\sqrt{0.14^2/12 + 0.14^2/12} = 0.06$. The standard error for the two-group test is 10-fold larger than for the paired test.

This striking difference in the standard errors is caused by the much lower variability of the individual velocity differences compared to the variability of the original measurements. Due to the correlation between swim velocities for a single swimmer, the differences in the two velocity measurements for each swimmer are consistently small, resulting in low variability. Pairing has allowed for increased precision in estimating the difference between groups.

The swim suit data illustrates the importance of context, which distinguishes a statistical problem from a purely mathematical one. While both the paired and two-group tests are numerically feasible to calculate, without an apparent error, the context of the problem dictates that the correct approach is to use a paired test.

GUIDED PRACTICE 6.15



Propose an experimental design for the embryonic stem cell study in sheep that would have required analysis with a paired t -test.¹⁴

¹⁴The experiment could have been done on pairs of siblings, with one assigned to the treatment group and one assigned to the control group. Alternatively, sheep could be matched up based on particular characteristics relevant to the experiment; for example, sheep could be paired based on similar weight or age. Note that in this study, a design involving two measurements taken on each sheep would be impractical.

6.6 Notes

The material in this chapter is particularly important. For many applications, t -tests and Analysis of Variance (ANOVA), which will study on the next chapter, are an essential part of the core of statistics in medicine and the life sciences. The comparison of two or more groups is often the primary aim of experiments both in the laboratory and in studies with human subjects. More generally, the approaches to interpreting and drawing conclusions from testing demonstrated in this chapter are used throughout the rest of the text and, indeed, in much of statistics.

While it is important to master the details of the techniques of testing for differences in two or more groups, it is even more critical to not lose sight of the fundamental principles behind the tests. A statistically significant difference in group means does not necessarily imply that group membership is the reason for the observed association. A significant association does not necessarily imply causation, even if it is highly significant; confounding variables may be involved. In most cases, causation can only be inferred in controlled experiments when interventions have been assigned randomly. It is also essential to carefully consider the context of a problem. For instance, students often find the distinction between paired and independent group comparisons confusing; understanding the problem context is the only reliable way to choose the correct approach.

It is generally prudent to use the form of the t -test that does not assume equal standard deviations, but you have less power. The formulas are simpler when standard deviations are equal, and software is more widely available for that case. The differences in sample sizes are usually minor and less important than assumptions about target differences or the values of the standard deviations. If the standard deviations are expected to be very different, then more specialized software for computing sample size and power should be used. The analysis done after the study has been completed should then use the t -test for unequal standard deviations.

Tests for significant differences are sometimes overused in science, with not enough attention paid to estimates and confidence intervals. Confidence intervals for the difference of two population means show a range of underlying differences in means that are consistent with the data, and often lead to insights not possible from only the test statistic and p -value. Wide confidence intervals may show that a non-significant test is the result of high variability in the test statistic, perhaps caused by a sample size that was too small. Conversely, a highly significant p -value may be the result of such a large sample size that the observed differences are not scientifically meaningful; that may be evident from confidence intervals with very narrow width.

Finally, the formula used to approximate degrees of freedom ν for the independent two-group t -test that does not assume equal variance is

$$\nu = \frac{\left[(s_1^2/n_1) + (s_2^2/n_2) \right]^2}{\left[(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1) \right]},$$

where n_1, s_1 are the sample size and standard deviation for the first sample, and n_2, s_2 are the corresponding values for the second sample. Since ν is routinely provided in the output from statistical software, there is rarely any need to calculate it by hand. The approximate formula $\text{df} = \min(n_1 - 1, n_2 - 1)$ always produces a smaller value for degrees of freedom and hence a larger p -value.

6.7 Exercises

Exercise. 6.1

A study is performed to compare the prevalence of the Iron Deficiency Anemia in two breeds of swine: Landrace and Spotted Poland. A random sample of 450 Landrace swines were selected and 105 showed positive indicators of IDA. Meanwhile, 120 out of 375 spotted Poland pigs have positive indicators of IDA. Has this study got enough evidence to point out that there exists a difference of the IDA prevalence in both groups? (Use $\alpha = 0.05$)

Exercise. 6.2

Two physiotherapeutic treatments are being tested to relieve some rheumatologic symptoms in horses. The first treatment (T_1) has been tested in 150 horses and 87 of them improved their conditions after one month of treatment. In the group receiving the second treatment T_2 90 out of 170 horses have improved their conditions. Is there any difference in the efficacy of the treatments? Calculate the p-value.

Exercise. 6.3

A group of researchers suppose that a relevant factor in hay fever in horses is the climate. They design an experiment with two type of climate zones (one region with temperate climate and a cold region) in the same country. In a sample of 1350 horses in the warm zone, 95 of them were positive in the hay fever test. In the other region, a sample of 2010 horses was selected and 113 in the sample were positive in the allergy test. Does this experiment provide enough evidence that the prevalence of hay fever is not the same in both region? Calculate the p-value.

Exercise. 6.4

A cardiovascular disease study for adult dogs was performed to confirm whether there exists any relation between pollution and this type of diseases. A sample of 1350 dogs was selected from several industrial areas and 95 of them have some cardiovascular disorder. Meanwhile, 113 dogs in a sample of 2010 dogs living in unpolluted area shown cardiovascular diseases. Can this study get the conclusion that the polluted areas have a different proportion of cardiovascular diseases than the unpolluted areas? ($\alpha = 0.05$) Calculate the p-value.

Exercise. 6.5

The height of the pelvis is measured for gorillas. In a sample of 12 male gorillas the mean was 13.21 cm and the standard deviation was 1.05 cm. In a sample of 9 female gorillas the mean was 11.00 cm and the sample standard deviation was 1.01 cm. Assuming that the height of the pelvis is normally distributed, answer the following questions:

1. Is there any difference in the variance of the pelvis height between female and male gorillas?
2. Is there any difference between the means of the pelvis height between female and male gorillas?

Exercise. 6.6

A 4-week weight control program was developed by a team of nutrition scientists. To evaluate the efficacy of the program the investigators selected 8 subjects who have body mass index higher than 30. Subjects were given specific instructions to comply in order to control the variables that may affect their weight, such as physical exercise and snacks. The weight of each subject at baseline and at the end of the 4-week clinical trial was measured. Suppose the weight distribution is known to be normal. State and

test the hypothesis to determine if the program is effective.¹⁵

Subject	Baseline	Program
1	82.3	76.5
2	76.5	73.8
3	103.7	98.2
4	96.8	95.8
5	108.5	112.6
6	94.3	89.9
7	115.7	111.4
8	125.1	117.4

Exercise. 6.7

Social phobia has in recent years been recognized as a considerable public health concern. It was speculated by psychiatrists that socially phobic patients who are diagnosed as having chronic depression may have greater fear of social interaction than those who do not have chronic depression. Liebowitz social anxiety test was given to 16 socially phobic subjects who suffer from chronic depression (group I), and 21 socially phobic subjects who are not chronically depressed (group II). The investigators computed descriptive statistics from the social anxiety test scores: $\bar{X}_1 = 22.6$, $S_1^2 = 14.0$, $\bar{X}_2 = 20.1$, and $S_2^2 = 12.2$. If the distribution of Liebowitz social anxiety test scores are normally distributed, what can you conclude from the data?

1. Check whether it can be assumed that the variances in the two groups are identical.
2. State the appropriate hypothesis according to the objectives of this study.
3. Perform the test by using the p value.¹⁶

Exercise. 6.8

It has been suggested that smoking does not affect the risk of cardiovascular diseases in populations with low serum cholesterol levels. To determine whether cigarette smoking is an independent risk factor among men with low levels of serum cholesterol, a nationwide, multicentered study was conducted. At one of the study sites, Orange Crest Community Hospital, 25 smokers and 47 non-smokers signed the consent form to participate in the study. Their serum cholesterol measurements are summarized below. Suppose the distribution of serum cholesterol levels is approximately normal. What can you conclude from the data collected by the researchers?¹⁷

Serum Cholesterol	Non-Smokers	Current Smokers
Sample mean	$\bar{X}_1 = 209.1$	$\bar{X}_2 = 213.3$
Sample SD	$S_1 = 35.5$	$S_2 = 37.6$

¹⁵J.S. Kim and R.J. Dailey. *Biostatistics for Oral Healthcare*. Wiley, 2008. ISBN: 9780470388273. URL: <https://books.google.es/books?id=n1fR0LF1jhsC>.

¹⁶Ibidem

¹⁷Ibidem

Chapter 7

Analysis of Variance

7.1 Comparing means with ANOVA

7.2 Multiple comparisons and controlling Type I Error rate

7.3 Exercises

In some settings, it is useful to compare means across several groups. It might be tempting to do pairwise comparisons between groups; for example, if there are three groups (A, B, C), why not conduct three separate t -tests (A vs. B , A vs. C , B vs. C)?

The primary issue here is that we are inspecting the data before picking the groups that will be compared. It is inappropriate to examine all data by eye (informal testing) and only afterwards decide which parts to formally test. This is called **data snooping** or **data fishing**. Naturally we would pick the groups with the large differences for the formal test, leading to an inflation in the Type 1 Error rate. To understand this better, let's consider the following problem:

Suppose we are to measure the aptitude for students in 20 classes in a large elementary school at the beginning of the year. In this school, all students are randomly assigned to classrooms, so any differences we observe between the classes at the start of the year are completely due to chance. However, with so many groups, we will probably observe a few groups that look rather different from each other. If we select only these classes that look so different, we will probably make the wrong conclusion that the assignment wasn't random. While we might only formally test differences for a few pairs of classes, we informally evaluated the other classes by eye before choosing the most extreme cases for a comparison. For additional information on the ideas expressed in this example, we recommend reading about the **prosecutor's fallacy**.¹

In the next section we will learn how to use the F statistic and ANOVA to test whether observed differences in means could have happened just by chance even if there was no difference in the respective population means. Conducting multiple tests on the same data increases the rate of Type I error, making it more likely that a difference will be found by chance, even if there is no difference among the population means. Multiple testing is discussed further in Section 7.2.



For labs, slides, and other resources, please visit
www.openintro.org/book/biostat

¹See, for example, www.stat.columbia.edu/~cook/movabletype/archives/2007/05/the_prosecutors.html.

7.1 Comparing means with ANOVA

Instead, the methodology behind a t -test can be generalized to a procedure called **analysis of variance (ANOVA)**, which uses a single hypothesis test to assess whether the means across several groups are equal. Strong evidence favoring the alternative hypothesis in ANOVA is described by unusually large differences among the group means.

H_0 : The mean outcome is the same across all k groups. In statistical notation, $\mu_1 = \mu_2 = \dots = \mu_k$ where μ_i represents the mean of the outcome for observations in category i .

H_A : At least one mean is different.

There are three conditions on the data that must be checked before performing ANOVA: 1) observations are independent within and across groups, 2) the data within each group are nearly normal, and 3) the variability across the groups is about equal.

EXAMPLE 7.1

Examine Figure 7.1. Compare groups I, II, and III. Is it possible to visually determine if the differences in the group centers is due to chance or not? Now compare groups IV, V, and VI. Do the differences in these group centers appear to be due to chance?

E

It is difficult to discern a difference in the centers of groups I, II, and III, because the data within each group are quite variable relative to any differences in the average outcome. However, there appear to be differences in the centers of groups IV, V, and VI. For instance, group V appears to have a higher mean than that of the other two groups. The differences in centers for groups IV, V, and VI are noticeable because those differences are large relative to the variability in the individual observations within each group.

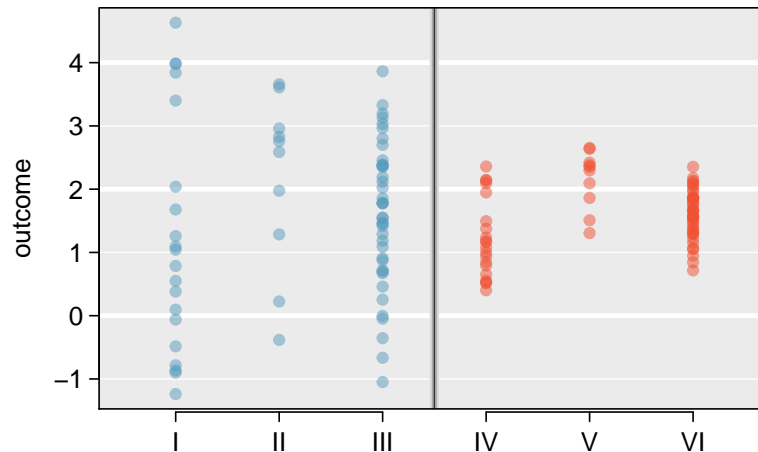


Figure 7.1: Side-by-side dot plot for the outcomes for six groups.

7.1.1 Diagnostics for ANOVA

As we have said, there are three conditions we must check for an ANOVA analysis: all observations must be independent, the data in each group must be nearly normal, and the variance

within each group must be approximately equal.

Independence. If the data are a simple random sample from less than 10% of the population, this condition is satisfied. If the group is assigned, of instance, in a clinical trial, then the group should be randomly assigned.

Approximately normal. As with one- and two-sample testing for means, the normality assumption is especially important when the sample size is quite small. Techniques to check that a distribution is normal are **Shapiro–Wilk** test, **Kolmogorov–Smirnov** test or Q-Q plot. Sometimes in ANOVA there are so many groups or so few observations per group that checking normality for each group isn't reasonable. For groups with a sample size large enough, namely greater than 30, the normal condition is not needed to be tested by these techniques.

Constant variance. The last assumption is that the variance in the groups is about equal from one group to the next. A method to check the constant variance is applying the **Levene's test**.

7.1.2 Analysis of variance (ANOVA) and the *F*-test

The famuss dataset was introduced in Chapter 1, Section 1.2.2. In the FAMuSS study, researchers examined the relationship between muscle strength and genotype at a location on the ACTN3 gene. The measure for muscle strength is percent change in strength in the non-dominant arm (ndrm.ch). Is there a difference in muscle strength across the three genotype categories (CC, CT, TT)?

GUIDED PRACTICE 7.2



The null hypothesis under consideration is the following: $\mu_{CC} = \mu_{CT} = \mu_{TT}$. Write the null and corresponding alternative hypotheses in plain language.²

Figure 7.2 provides summary statistics for each group. A side-by-side boxplot for the change in non-dominant arm strength is shown in Figure 7.3; Figure 7.4 shows the Q-Q plots by each genotype. Notice that the variability appears to be approximately constant across groups; nearly constant variance across groups is an important assumption that must be satisfied for using ANOVA. Based on the Q-Q plots, there is evidence of moderate right skew; the data do not follow a normal distribution very closely, but could be considered to 'loosely' follow a normal distribution.³ It is reasonable to assume that the observations are independent within and across groups; it is unlikely that participants in the study were related, or that data collection was carried out in a way that one participant's change in arm strength could influence another's.

	CC	CT	TT
Sample size (n_i)	173	261	161
Sample mean (\bar{x}_i)	48.89	53.25	58.08
Sample SD (s_i)	29.96	33.23	35.69

Figure 7.2: Summary statistics of change in non-dominant arm strength, split by genotype.

² H_0 : The average percent change in non-dominant arm strength is equal across the three genotypes. H_A : The average percent change in non-dominant arm strength varies across some (or all) groups.

³In a more advanced course, it can be shown that the ANOVA procedure still holds with deviations from normality when sample sizes are moderately large. Additionally, a more advanced course would discuss appropriate transformations to induce normality.

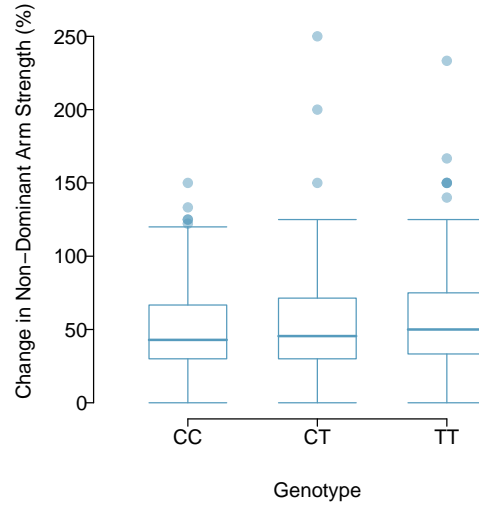


Figure 7.3: Side-by-side box plot of the change in non-dominant arm strength for 595 participants across three groups.

EXAMPLE 7.3

The largest difference between the sample means is between the CC and TT groups. Consider again the original hypotheses:

$$H_0: \mu_{CC} = \mu_{CT} = \mu_{TT}$$

H_A : The average percent change in non-dominant arm strength (μ_i) varies across some (or all) groups.

E

Why might it be inappropriate to run the test by simply estimating whether the difference of μ_{CC} and μ_{TT} is statistically significant at a 0.05 significance level?

It is inappropriate to informally examine the data and decide which groups to formally test. This is a form of **data fishing**; choosing the groups with the largest differences for the formal test will lead to an increased chance of incorrectly rejecting the null hypothesis (i.e., an inflation in the Type I error rate). Instead, all the groups should be tested using a single hypothesis test.

Analysis of variance focuses on answering one question: is the variability in the sample means large enough that it seems unlikely to be from chance alone? The variation between groups is referred to as the **mean square between groups** (MSG); the MSG is a measure of how much each group mean varies from the overall mean. Let \bar{x} represent the mean of outcomes across all groups, where \bar{x}_i is the mean of outcomes in a particular group i and n_i is the sample size of group i . The mean square between groups is:

$$MSG = \frac{1}{k-1} \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2 = \frac{1}{df_G} SSG,$$

where SSG is the **sum of squares between groups**, $\sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2$, and $df_G = k - 1$ is the degrees of freedom associated with the MSG when there are k groups.

Under the null hypothesis, any observed variation in group means is due to chance and there is no real difference between the groups. In other words, the null hypothesis assumes that the groupings are non-informative, such that all observations can be thought of as belonging to a single group. If this scenario is true, then it is reasonable to expect that the variability between the group

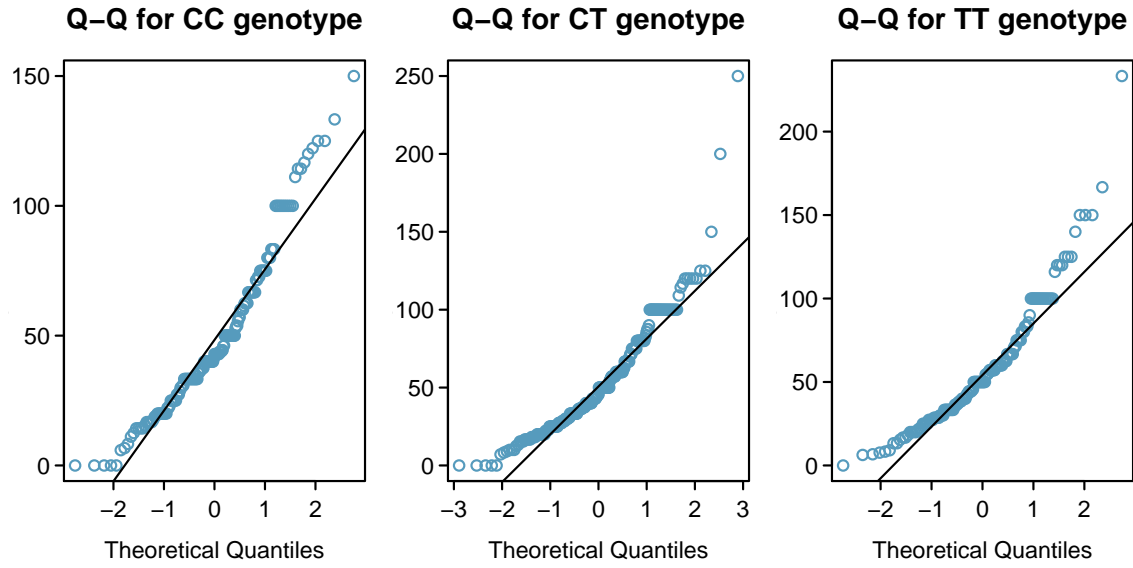


Figure 7.4: Q-Q plots of the change in non-dominant arm strength for 595 participants across three groups.

means should be equal to the variability observed within a single group. The **mean square error** (MSE) is a pooled variance estimate with associated degrees of freedom $df_E = n - k$ that provides a measure of variability within the groups. The mean square error is computed as:

$$MSE = \frac{1}{n - k} \sum_{i=1}^k (n_i - 1)s_i^2 = \frac{1}{df_E} SSE,$$

where the SSE is the **sum of squared errors**, n_i is the sample size of group i , and s_i is the standard deviation of group i .

Under the null hypothesis that all the group means are equal, any differences among the sample means are only due to chance; thus, the MSG and MSE should also be equal. ANOVA is based on comparing the MSG and MSE . The test statistic for ANOVA, the **F-statistic**, is the ratio of the between-group variability to the within-group variability:

$$F = \frac{MSG}{MSE}. \quad (7.4)$$

EXAMPLE 7.5

Calculate the F -statistic for the famuss data summarized in Figure 7.2. The overall mean \bar{x} across all observations is 53.29.

First, calculate the MSG and MSE .

$$\begin{aligned}
 MSG &= \frac{1}{k-1} \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2 \\
 &= \frac{1}{3-1} [(173)(48.89 - 53.29)^2 + (261)(53.25 - 53.29)^2 + (161)(58.08 - 53.29)^2] \\
 &= 3521.69 \\
 MSE &= \frac{1}{n-k} \sum_{i=1}^k (n_i - 1) s_i^2 \\
 &= \frac{1}{595-3} [(173-1)(29.96^2) + (261-1)(33.23^2) + (161-1)(35.69^2)] \\
 &= 1090.02
 \end{aligned}$$

The F -statistic is the ratio:

$$\frac{MSG}{MSE} = \frac{3521.69}{1090.02} = 3.23.$$

A p -value can be computed from the F -statistic using an F -distribution, which has two associated parameters: df_1 and df_2 . For the F -statistic in ANOVA, $df_1 = df_G$ and $df_2 = df_E$. An F distribution with 2 and 592 degrees of freedom, corresponding to the F -statistic for the genotype and muscle strength hypothesis test, is shown in Figure 7.5.

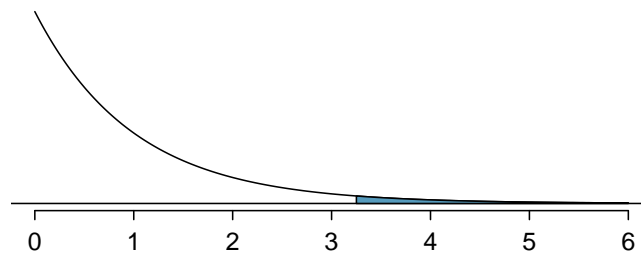


Figure 7.5: An F -distribution with $df_1 = 2$ and $df_2 = 592$. The tail area greater than $F = 3.23$ is shaded.

The larger the observed variability in the sample means (MSG) relative to the within-group variability (MSE), the larger F will be. Larger values of F represent stronger evidence against the null hypothesis. The upper tail of the distribution is used to compute a p -value, which is typically done using statistical software.

EXAMPLE 7.6

The p -value corresponding to the test statistic is equal to about 0.04. Does this provide strong evidence against the null hypothesis at significance level $\alpha = 0.05$?

E

The p -value is smaller than 0.05, indicating the evidence is strong enough to reject the null hypothesis at a significance level of 0.05. The data suggest that average change in strength in the non-dominant arm varies by participant genotype.

THE F -STATISTIC AND THE F -TEST

Analysis of variance (ANOVA) is used to test whether the mean outcome differs across two or more groups. ANOVA uses a test statistic F , which represents a standardized ratio of variability in the sample means relative to the variability within the groups. If H_0 is true and the model assumptions are satisfied, the statistic F follows an F distribution with parameters $df_1 = k - 1$ and $df_2 = n - k$. The upper tail of the F -distribution is used to calculate the p -value.

7.1.3 Reading an ANOVA table from software

The calculations required to perform an ANOVA by hand are tedious and prone to human error. Instead, it is common to use statistical software to calculate the F -statistic and associated p -value.

Figure 7.6 shows an ANOVA summary to test whether the mean change in non-dominant arm strength varies by genotype. Many of these values should look familiar; in particular, the F -statistic and p -value can be retrieved from the last two columns.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
famuss\$actn3.r577x	2	7043	3522	3.231	0.0402
Residuals	592	645293	1090		

Figure 7.6: ANOVA summary for testing whether the mean change in non-dominant arm strength varies by genotype at the actn3.r577x location on the ACTN3 gene.

7.1.4 An example with R-Commander**Example 7.7**

An ethnology research group wants to determine whether or not there exist significant differences from the height of horses depending on the breed. For this purpose, 45 horses have been chosen belonging to three different breeds (Berber, Buckskin and Budyonny) and they have been randomly selected in such a way there are 15 horses for each breed. In Figure 7.7 the height of the horses is displayed for each group.

Condition 1: Independence This hypothesis is fulfilled, since the assignment of the treatment has been randomized across the breeds. The text explains that the horses have been randomly selected among the horses of the same breed.

Condition 2: Normality The hypotheses of the normality of Berber horses are:

H_0 :	The height is normally distributed
H_A :	The height is not normally distributed

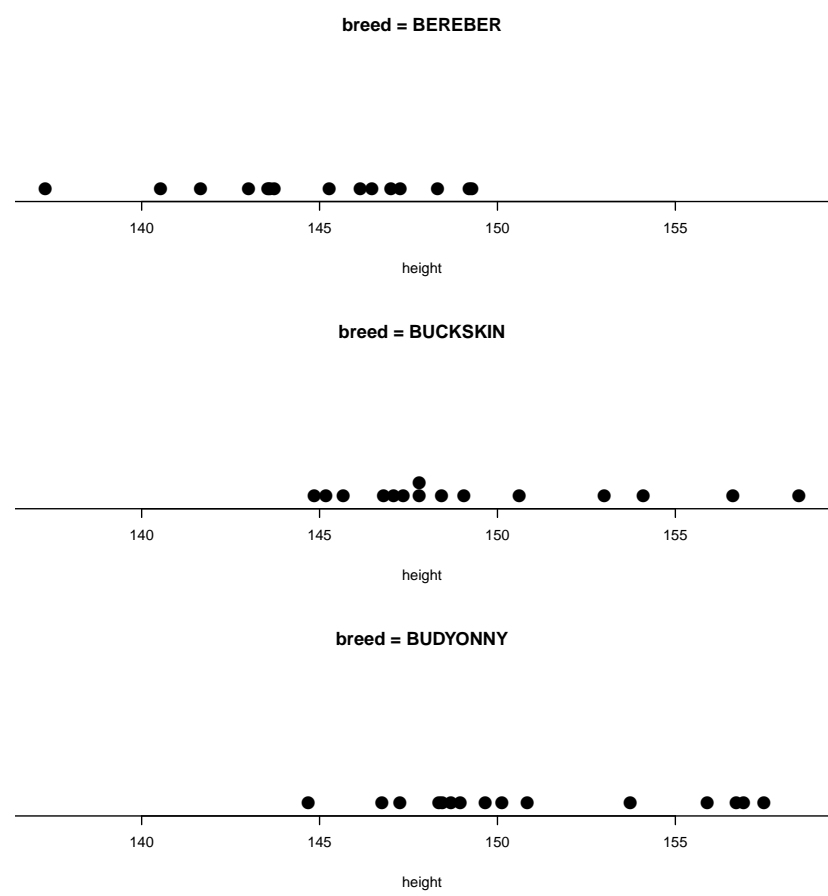


Figure 7.7: Distribution of height within each breed.

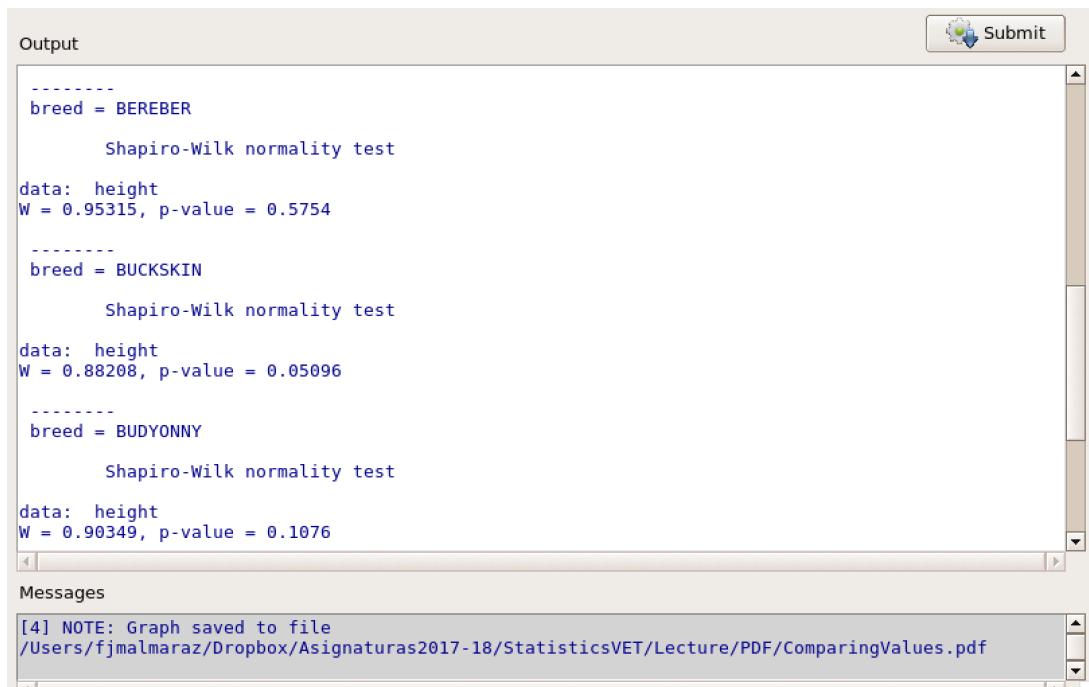


Figure 7.8: Screenshot of Shapiro-Wilk test.

According to the data shown in the figure 7.8, the p-value 0.5754 is greater than $\alpha = 0.05$. Hence, we fail to reject the null hypothesis and a normal distribution of height for Berber horses. The corresponding p-values for Buckskin and Budyonny breeds are 0.05096 and 0.1076, respectively.

Levene's test (identical variances) The hypotheses of the variances of the height of the horses are:

$$H_0 \quad \sigma_{Berber}^2 = \sigma_{Buckskin}^2 = \sigma_{Budyonny}^2$$

$$H_A \quad \sigma_{Berber}^2 \neq \sigma_{Buckskin}^2 \text{ or } \sigma_{Berber}^2 \neq \sigma_{Budyonny}^2 \text{ or } \sigma_{Buckskin}^2 \neq \sigma_{Budyonny}^2$$

where σ_{Berber}^2 is the population variance of the height of Berber horses.

According to the figure 7.9, T the p-value=0.8564 is greater than $\alpha = 0.05$ We fail to reject the null hypothesis and we do not have evidence against being any difference among the variances. We can assume that for the three breeds have identical variance.

Finally, we can perform the ANOVA test, since all the conditions are met.

The hypotheses of the ANOVA test are:

$$H_0 \quad \mu_{Berber} = \mu_{Buckskin} = \mu_{Budyonny}$$

$$H_A \quad \mu_{Berber} \neq \mu_{Buckskin} \text{ or } \mu_{Berber} \neq \mu_{Budyonny} \text{ or } \mu_{Buckskin} \neq \mu_{Budyonny}$$

where μ_{Berber} is the population mean of the height of Berber horses and analogously for the other groups.

The degrees of freedom are $k-1 = 3-1 = 2$ and $n-k = 45-3 = 42$. The significance level is $\alpha = 0.05$. Hence, we use 95 percentile for the Snedecor's F-table with degrees of freedom 2 and 50, which is the closest value to 42 in the table, the value is $F_{2,50,0.95} = 3.18$. The acceptance interval is $[0, 3.18)$.

The test statistic $F = 10.1$ is outside interval we reject null hypothesis. We have enough evidence that the height of horses depends on the breed. Please, observe that the computer (figure 7.10) provides a more accurate p-value and the test statistic.

```

Output

> with(horses, tapply(height, breed, var, na.rm=TRUE))
  BEREBER BUCKSKIN BUDYONNY
11.39555 17.54513 17.05423

> leveneTest(height ~ breed, data=horses, center="median")
Levene's Test for Homogeneity of Variance (center = "median")
      Df F value Pr(>F)
group  2  0.1555 0.8564
      42

```

Figure 7.9: Screenshot of Levene's test. .

```

Output
Submit

> summary(AnovaModel.1)
      Df Sum Sq Mean Sq F value    Pr(>F)
breed    2   309.6   154.79    10.1 0.000263 ***
Residuals 42   643.9    15.33
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> with(horses, numSummary(height, groups=breed, statistics=c("mean", "sd")))
      mean      sd data:n
BEREBER 144.8240 3.375730    15
BUCKSKIN 149.5213 4.188690    15
BUDYONNY 150.9687 4.129676    15

```

Figure 7.10: Screenshot of ANOVA test. .

7.2 Multiple comparisons and controlling Type I Error rate

Rejecting the null hypothesis in an ANOVA analysis only allows for a conclusion that there is evidence for a difference in group means. In order to identify the groups with different means, it is necessary to perform further testing. For example, in the famuss analysis, there are three comparisons to make: CC to CT, CC to TT, and CT to TT. While these comparisons can be made using two sample t -tests, it is important to control the Type I error rate. One of the simplest ways to reduce the overall probability of identifying a significant difference by chance in a multiple comparisons setting is to use the Bonferroni correction procedure.

In the Bonferroni correction procedure, the p -value from a two-sample t -test is compared to a modified significance level, α^* ; $\alpha^* = \alpha/K$, where K is the total number of comparisons being considered. For k groups, $K = \frac{k(k-1)}{2}$. When calculating the t -statistic, use the pooled estimate of standard deviation between groups (which equals \sqrt{MSE}); to calculate the p -value, use a t -distribution with df_2 . It is typically more convenient to do these calculations using software.

BONFERRONI CORRECTION

The **Bonferroni correction** suggests that a more stringent significance level is appropriate when conducting multiple tests:

$$\alpha^* = \alpha/K$$

where K is the number of comparisons being considered. For k groups, $K = \frac{k(k-1)}{2}$.

EXAMPLE 7.8

The ANOVA conducted on the famuss dataset showed strong evidence of differences in the mean strength change in the non-dominant arm between the three genotypes. Complete the three possible pairwise comparisons using the Bonferroni correction and report any differences.

Use a modified significance level of $\alpha^* = 0.05/3 = 0.0167$. The pooled estimate of the standard deviation is $\sqrt{MSE} = \sqrt{1090.02} = 33.02$.

Genotype CC versus Genotype CT:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\text{pooled}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{48.89 - 53.25}{33.02 \sqrt{\frac{1}{173} + \frac{1}{261}}} = -1.35.$$

This results in a p -value of 0.18 on $df = 592$. This p -value is larger than $\alpha^* = 0.0167$, so there is not evidence of a difference in the means of genotypes CC and CT.

Genotype CC versus Genotype TT:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\text{pooled}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{48.89 - 58.08}{33.02 \sqrt{\frac{1}{173} + \frac{1}{161}}} = -2.54.$$

This results in a p -value of 0.01 on $df = 592$. This p -value is smaller than $\alpha^* = 0.0167$, so there is evidence of a difference in the means of genotypes CC and TT.

Genotype CT versus Genotype TT:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\text{pooled}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{53.25 - 58.08}{33.02 \sqrt{\frac{1}{261} + \frac{1}{161}}} = -1.46.$$

This results in a p -value of 0.14 on $df = 592$. This p -value is larger than $\alpha^* = 0.0167$, so there is not evidence of a difference in the means of genotypes CT and TT.

In summary, the mean percent strength change in the non-dominant arm for genotype CT individuals is not statistically distinguishable from those of genotype CC and TT individuals. However, there is evidence that mean percent strength change in the non-dominant arm differs between individuals of genotype CC and TT are different.

7.2.1 Reading the results of pairwise t -tests from software

Statistical software can be used to calculate the p -values associated with each possible pairwise comparison of the groups in ANOVA. The results of the pairwise tests are summarized in a table that shows the p -value for each two-group test.

Figure 7.11 shows the p -values from the three possible two-group t -tests comparing change in non-dominant arm strengths between individuals with genotypes CC, CT, and TT. For example, the table indicates that when comparing mean change in non-dominant arm strength between TT and CC individuals, the p -value is 0.01. This coheres with the calculations above, and these unadjusted p -values should be compared to $\alpha^* = 0.0167$.

The use of statistical software makes it easier to apply corrections for multiple testing, such that it is not necessary to explicitly calculate the value of α^* . Figure 7.12 shows the Bonferroni-

	CC	CT
CT	0.18	-
TT	0.01	0.14

Figure 7.11: Unadjusted p -values for pairwise comparisons testing whether the mean change in non-dominant arm strength varies by genotype at the actn3.r577x location on ACTN3 gene.

adjusted p -values from the three possible tests. When statistical software applies the Bonferroni correction, the unadjusted p -value is multiplied by K , the number of comparisons, allowing for the values to be directly compared to α , not α^* . Comparing an unadjusted p -value to α/K is equivalent to comparing the quantity ($K \times p$ -value) to α .

	CC	CT
CT	0.54	-
TT	0.03	0.43

Figure 7.12: Bonferroni-adjusted p -values for pairwise comparisons testing whether the mean change in non-dominant arm strength varies by genotype at the actn3.r577x location on ACTN3 gene.

7.2.2 Tukey's range test

Another technique to compare which pairs of groups are significantly different is the Tukey's range test, where a confidence interval is given for a pairwise comparison. If zero is not included in the confidence interval, we have enough evidence of a significant difference between the two groups. An example with the breeds of horses is in the figure 7.13

Caution: Sometimes an ANOVA will reject the null but no groups will have statistically significant differences

It is possible to reject the null hypothesis using ANOVA and then to not subsequently identify differences in the pairwise comparisons. However, *this does not invalidate the ANOVA conclusion*. It only means we have not been able to successfully identify which groups differ in their means.

The ANOVA procedure examines the big picture: it considers all groups simultaneously to decipher whether there is evidence that some difference exists. Even if the test indicates that there is strong evidence of differences in group means, identifying with high confidence a specific difference as statistically significant is more difficult.

Consider the following analogy: we observe a Wall Street firm that makes large quantities of money based on predicting mergers. Mergers are generally difficult to predict, and if the prediction success rate is extremely high, that may be considered sufficiently strong evidence to warrant investigation by the Securities and Exchange Commission (SEC). While the SEC may be quite certain that there is insider trading taking place at the firm, the evidence against any single trader may not be very strong. It is only when the SEC considers all the data that they identify the pattern. This is effectively the strategy of ANOVA: stand back and consider all the groups simultaneously.

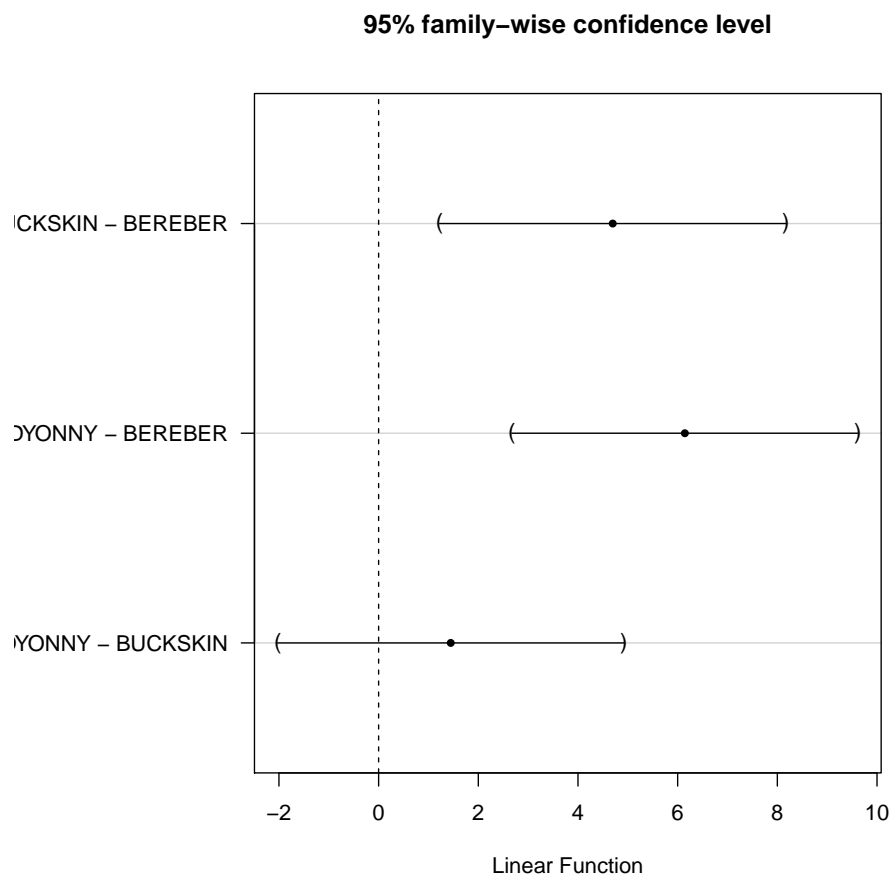


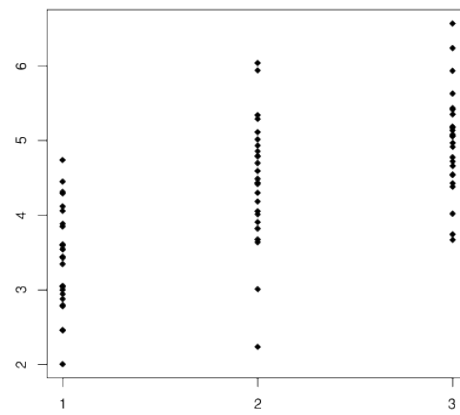
Figure 7.13: Tuckey's confidence interval for pairwise comparison. We have evidence of a difference between Berber and the other two breeds, since the zero value is not included in the interval. However, we do not have enough evidence that there exists a difference of mean between Budyonny and Buckskin.

7.3 Exercises

Exercise. 7.1

A epidemiologist wants to compare three types of vaccine against brucellosis in cattle. With that purpose 103 cows were randomized in three groups for each of the type of vaccines. The response of the antibodies was measured two weeks later for every individual.

1. This graphic shows the values got for each animal. Do you think that ANOVA test will detect any difference according to the figure? Do you think that they will meet the conditions to apply ANOVA?



%labelFig1

2. We got the following outcome with the computer:

- Levene's test (p-value=0.3758)
- Type 1: Shapiro-Wilk Test (p-value=0.4567)
- Type 2: Shapiro-Wilk Test (p-value=0.4538)
- Type 3: Shapiro-Wilk Test (p-value=0.0834)

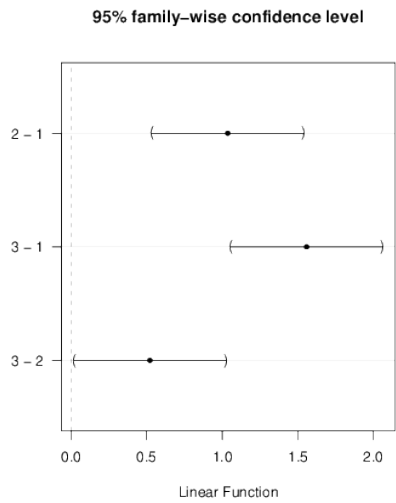
Set a hypothesis testing and explain your conclusion for each p-value. Are fulfilled the conditions for ANOVA?

3. The ANOVA test statistic is $F = 35.3$

- (a) What distribution is used to calculate the acceptance interval
- (b) Calculate the acceptance interval with a significance level $\alpha = 0.05$
- (c) Is ANOVA test enough to know which pairs of treatments are going to be different

4. The computer displays the following graph and data for a pairwise comparison:

Comparison	Diference Sample mean	95% interval Lower Limit	95% Interval Upper Limit
2-1	1.03	0.53	1.54
3-1	1.56	1.05	2.06
3-2	0.42	0.02	1.02



Exercise. 7.2

A medical department wants to determine whether there exists significant difference in the recovering time of a surgery to remove a tumor in the urinary bladder using three techniques: A (laparoscopy), B (conventional surgery) and C (minimally invasive procedure) . They selected 58 patients with this kind of tumor and they were randomized to one of techniques: 25 patients underwent laparoscopy, 13 conventional procedure and 20 were operated with the minimally invasive procedure.

The conditions of ANOVA were checked with the following tests:

Levene’s test (p-value= 0.6082)

Procedure A: Shapiro - Wilk Test (p-value= 0.3444)

Procedure B: Shapiro - Wilk Test (p-value= 0.5688)

Procedure C: Shapiro - Wilk Test (p-value= 0.3060)

ANOVA: F=80.13

Comparison	Difference Sample mean	95% Interval Lower Lim	95% Interval Upper Lim
B-A	8.43	6.06	10.80
C-A	10.28	8.20	12.36
C-B	1.85	-0.61	4.32

- (a) Are all the conditions met to apply ANOVA?
- (b) Set hypothesis testing of ANOVA and write your conclusions according to the data provided in the text.
- (c) What are your conclusions according to Tuckey’s pairwise comparison test?

Exercise. 7.3

A professor who teaches a large introductory statistics class (197 students) with eight discussion sections would like to test if student performance differs by discussion section, where each discussion section has a different teaching assistant. The summary table below shows the average final exam score for each discussion section as well as the standard deviation of scores and the number of students in each section.

	Sec 1	Sec 2	Sec 3	Sec 4	Sec 5	Sec 6	Sec 7	Sec 8
n_i	33	19	10	29	33	10	32	31
\bar{x}_i	92.94	91.11	91.80	92.45	89.30	88.30	90.12	93.35
s_i	4.21	5.58	3.43	5.92	9.32	7.27	6.93	4.57

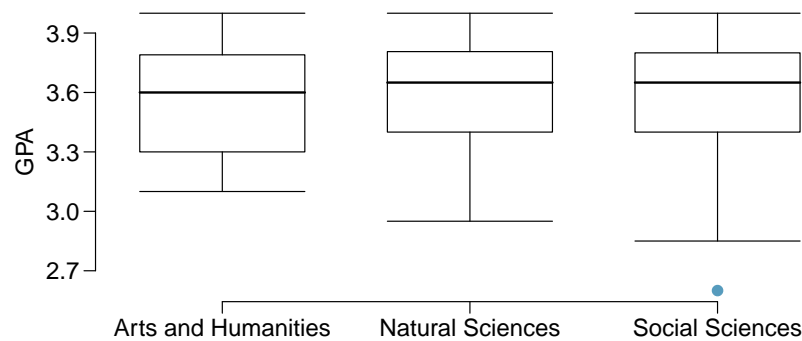
The ANOVA output below can be used to test for differences between the average scores from the different discussion sections.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
section	7	525.01	75.00	1.87	0.0767
Residuals	189	7584.11	40.13		

Conduct a hypothesis test to determine if these data provide convincing evidence that the average score varies across some (or all) groups. Check conditions and describe any assumptions you must make to proceed with the test.

Exercise. 7.4

Undergraduate students taking an introductory statistics course at Duke University conducted a survey about GPA and major. The side-by-side box plots show the distribution of GPA among three groups of majors. Also provided is the ANOVA output.



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
major	2	0.03	0.02	0.21	0.8068
Residuals	195	15.77	0.08		

- Write the hypotheses for testing for a difference between average GPA across majors.
- What is the conclusion of the hypothesis test?
- How many students answered these questions on the survey, i.e. what is the sample size?

Chapter 8

The chi-square test

8.1 Inference for two or more groups

8.2 Examples of the chi-square test

8.3 Exercises

Two-way tables of two qualitative variables are often used to summarize data from medical research studies, and entire texts have been written about methods of analysis for these tables. This chapter covers only the most basic of those methods.



For labs, slides, and other resources, please visit
www.openintro.org/book/biostat

8.1 Inference for two or more groups

The comparison of the proportion of breast cancer deaths between the two groups can also be approached using a two-way contingency table, which contains counts for combinations of outcomes for two variables. The results for the mammogram study in this format are shown in Figure 8.1. Each value in the table represents the number of times a particular combination of variable outcomes occurred. For example, the value 500 corresponds to the number of death from breast cancer (BC) in the group taking mammographs on a regular basis. Row and column totals are also included. The **row totals** provide the total counts across each row (e.g. $500 + 44,425 = 44,925$), and **column totals** are total counts down each column.

Previously, the main question of interest was stated as, "Is there evidence of a difference in the proportion of breast cancer deaths between the two screening groups?" If the probability of a death from breast cancer does not depend the method of screening, then screening method and outcome are independent. Thus, the question can be re-phrased: "Is there evidence that screening method is associated with outcome?"

Hypothesis testing in a two-way table assesses whether the two variables of interest are associated (i.e., not independent). The approach can be applied to settings with two or more groups and for responses that have two or more categories. The observed number of counts in each table cell are compared to the number of **expected counts**, where the expected counts are calculated under the assumption that the null hypothesis of no association is true. A χ^2 test of significance is based on the differences between observed and expected values in the cells.

Death from BC	Yes	No	Total
Mammogram	500	44,425	44,925
Control	505	44,405	44,910
Total	1,005	88,830	89,835

Figure 8.1: Results of the mammogram study, as a contingency table with marginal totals.



GUIDED PRACTICE 8.1

Formulate hypotheses for a contingency-table approach to analyzing the mammogram data.¹

8.1.1 Expected counts

If type of breast cancer screening had no effect on outcome in the mammogram data, what would the expected results be?

Recall that if two events A and B are independent, then $P(A \cap B) = P(A)P(B)$. Let A represent assignment to the mammogram group and B the event of death from breast cancer. Under independence, the number of individuals out of 89,835 that are expected to be in the mammogram screening group and die from breast cancer equals:

$$(89,835)P(A)P(B) = (89,835)\left(\frac{44,925}{89,835}\right)\left(\frac{1,005}{89,835}\right) = 502.6.$$

Note that the quantities 44,925 and 1,005 are the row and column totals corresponding to the

¹ H_0 : There is no association between type of breast cancer screening and death from breast cancer. H_A : There is an association between type of breast cancer screening and death from breast cancer.

upper left cell of Figure 8.1, and 89,835 is the total number n of observations in the table. A general formula for computing expected counts for any cell can be written from the marginal totals and the total number of observations.

COMPUTING EXPECTED COUNTS IN A TWO-WAY TABLE

To calculate the expected count for the i^{th} row and j^{th} column, compute

$$\text{Expected Count}_{\text{row } i, \text{col } j} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{table total}}.$$

EXAMPLE 8.2

Calculate expected counts for the data in Figure 8.1.

E

$$\begin{aligned} E_{1,1} &= \frac{44,925 \times 1,005}{89,835} = 502.6 & E_{1,2} &= \frac{44,925 \times 88,830}{89,835} = 44,422.4 \\ E_{2,1} &= \frac{2,922 \times 1,005}{89,835} = 502.4 & E_{2,2} &= \frac{7,078 \times 88,830}{89,835} = 44,407.6 \end{aligned}$$

Death from BC	Yes	No	Total
Mammogram	500 (502.6)	44,425 (44,422.4)	44,925
Control	505 (502.4)	44,405 (44,407.6)	44,910
Total	1,005	88,830	89,835

Figure 8.2: Results of the mammogram study, with (expected counts). The expected counts should also sum to the row and column totals; this can be a useful check for accuracy.

EXAMPLE 8.3

If a newborn is HIV⁺, should he or she be treated with nevirapine (NVP) or a more expensive drug, lopinavir (LPV)? In this setting, success means preventing virologic failure; i.e., growth of the virus. A randomized study was conducted to assess whether there is an association between treatment and outcome.² Of the 147 children administered NVP, about 41% experienced virologic failure; of the 140 children administered LPV, about 19% experienced virologic failure. Construct a table of observed counts and a table of expected counts.

E

Convert the proportions to count data: 41% of 147 is approximately 60, and 19% of 140 is approximately 27. The observed results are given in Figure 8.3.

Calculate the expected counts for each cell:

$$\begin{aligned} E_{1,1} &= \frac{87 \times 147}{287} = 44.6 & E_{1,2} &= \frac{87 \times 140}{287} = 42.4 \\ E_{2,1} &= \frac{200 \times 147}{287} = 102.4 & E_{2,2} &= \frac{200 \times 140}{287} = 97.6 \end{aligned}$$

The expected counts are summarized in Figure 8.4.

	NVP	LPV	Total
Virologic Failure	60	27	87
Stable Disease	87	113	200
Total	147	140	287

Figure 8.3: Observed counts for the HIV study.

	NVP	LPV	Total
Virologic Failure	44.6	42.4	87
Stable Disease	102.4	97.6	200
Total	147	140	287

Figure 8.4: Expected counts for the HIV study.

8.1.2 The χ^2 test statistic

Previously, test statistics have been constructed by calculating the difference between a point estimate and a null value, then dividing by the standard error of the point estimate to standardize the difference. The χ^2 statistic is based on a different idea. In each cell of a table, the difference *observed - expected* is a measure of the discrepancy between what was observed in the data and what should have been observed under the null hypothesis of no association. If the row and column variables are highly associated, that difference will be large. Two adjustments are made to the differences before the final statistic is calculated. First, since both positive and negative differences suggest a lack of independence, the differences are squared to remove the effect of the sign. Second, cells with larger counts may have larger discrepancies by chance alone, so the squared differences in each cell are scaled by the number expected in the cell under the hypothesis of independence. The final χ^2 statistic is the sum of these standardized squared differences, where the sum has one term for each cell in the table.

The χ^2 test statistic is calculated as:

χ^2
chi-square
test statistic

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}.$$

The theory behind the χ^2 test and its sampling distribution relies on the same normal approximation to the binomial distribution that was introduced earlier. The cases in the dataset must be independent and each expected cell count should be at least 10. The second condition can be relaxed in tables with more than 4 cells.

CONDITIONS FOR THE χ^2 TEST

Two conditions that must be checked before performing a χ^2 test:

- Independence.** Each case that contributes a count to the table must be independent of all the other cases in the table.
- Sample size.** Each expected cell count must be greater than or equal to 10. For tables larger than 2×2 , it is appropriate to use the test if no more than 1/5 of the expected counts are less than 5, and all expected counts are greater than 1.

²Violari A, et al. N Engl J Med 2012; 366:2380-2389 DOI: 10.1056/NEJMoa1113249

EXAMPLE 8.4

For the mammogram data, check the conditions for the χ^2 test and calculate the χ^2 test statistic.

Independence is a reasonable assumption, since individuals have been randomized to either the treatment or control group. Each expected cell count is greater than 10.

E

$$\begin{aligned}\chi^2 &= \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(500 - 502.6)^2}{502.6} + \frac{(44,425 - 44,422.4)^2}{44,422.4} + \frac{(505 - 502.4)^2}{502.4} + \frac{(44,405 - 44,407.6)^2}{44,407.6} \\ &= 0.02.\end{aligned}$$

G

GUIDED PRACTICE 8.5

For the HIV data, check the conditions for the χ^2 test and calculate the χ^2 test statistic.³

8.1.3 Calculating p -values for a χ^2 distribution

The **chi-square distribution** is often used with data and statistics that are positive and right-skewed. The distribution is characterized by a single parameter, the degrees of freedom. Figure 8.5 demonstrates three general properties of chi-square distributions as the degrees of freedom increases: the distribution becomes more symmetric, the center moves to the right, and the variability increases.

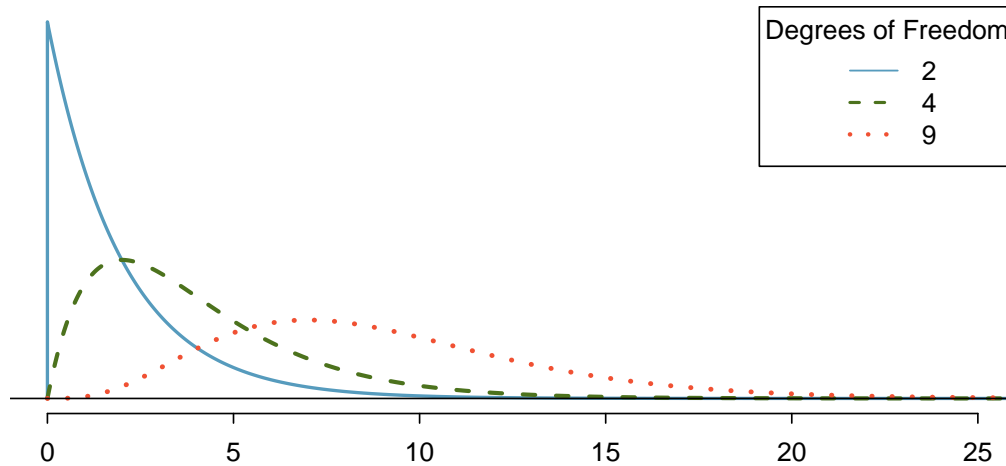


Figure 8.5: Three chi-square distributions with varying degrees of freedom.

The χ^2 statistic from a contingency table has a sampling distribution that approximately follows a χ^2 distribution with degrees of freedom $df = (r - 1)(c - 1)$, where r is the number of rows and c is the number of columns. Either statistical software or a table can be used to calculate p -values from the χ^2 distribution. The **chi-square table** is partially shown in Figure 8.6, and a more complete table is presented in Appendix B.3 on page 222. This table is very similar to the t -table:

³Independence holds, since this is a randomized study. The expected counts are greater than 10. $\chi^2 = \frac{(60-44.6)^2}{44.6} + \frac{(27-42.4)^2}{42.4} + \frac{(87-102.4)^2}{102.4} + \frac{(113-97.6)^2}{97.6} = 14.7$.

each row provides values for distributions with different degrees of freedom, and a cut-off value is provided for specified tail areas. One important difference from the t -table is that the χ^2 table only provides upper tail values.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

Figure 8.6: A section of the chi-square table. A complete table is in Appendix B.3 on page 222.

EXAMPLE 8.6

Calculate an approximate p -value for the mammogram data, given that the χ^2 statistic equals 0.02. Assess whether the data provides convincing evidence of an association between screening group and breast cancer death.

E

The degrees of freedom in a 2×2 table is 1, so refer to the values in the first column of the probability table. The value 0.02 is less than 1.07, so the p -value is greater than 0.3. The data do not provide convincing evidence of an association between screening group and breast cancer death. This supports the conclusions from Example 6.5, where the p -value was calculated to be 0.8650 and is visualized in Figure 8.7.

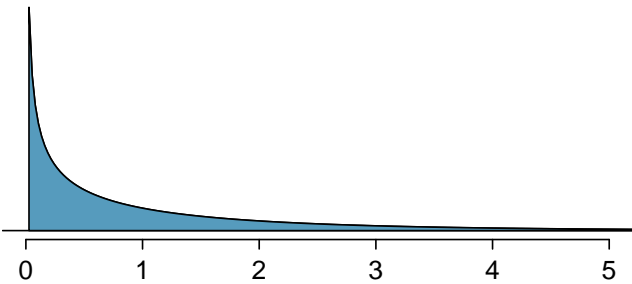


Figure 8.7: The p -value for the mammogram data is shaded on the χ^2 distribution with $df = 1$. The shaded area is to the right of $x = 0.02$.

GUIDED PRACTICE 8.7

G

Calculate an approximate p -value for the HIV data. Assess whether the data provides convincing evidence of an association between treatment and outcome at the $\alpha = 0.01$ significance level.⁴

⁴The χ^2 statistic is 14.7. For degrees of freedom 1, the tail area beyond 14.7 is smaller than 0.001. There is evidence to suggest that treatment is not independent of outcome.

8.1.4 Interpreting the results of a χ^2 test

If the p -value from a χ^2 test is small enough to provide evidence to reject the null hypothesis of no association, it is important to explore the results further to understand direction of the observed association. This is done by examining the residuals, the standardized differences of the *observed* - *expected*, for each cell. Instead of using squared differences, the residuals are based on the differences themselves, and the standardizing or scaling factor is $\sqrt{\text{expected}}$. Calculating residuals can be particularly helpful for understanding the results from large tables.

For each cell in a table, the residual equals:

$$\frac{\text{observed} - \text{expected}}{\sqrt{\text{expected}}}.$$

Residuals with a large magnitude contribute the most to the χ^2 statistic. If a residual is positive, the observed value is greater than the expected value, and vice versa for a negative residual.

8.2 Examples of the chi-square test

EXAMPLE 8.8

In the FAMuSS study introduced in Chapter 1, researchers measured a variety of demographic and genetic characteristics for about 1,300 participants, including data on race and genotype at a specific locus on the ACTN3 gene (Figure 8.8). Is there evidence of an association between genotype and race?

First, check the assumptions for applying a χ^2 test. It is reasonable to assume independence, since it is unlikely that any participants were related to each other. None of the expected counts, as shown in Figure 8.9, are less than 5.

H_0 : Race and genotype are independent.

H_A : Race and genotype are not independent.

Let $\alpha = 0.05$.

Calculate the χ^2 statistic:

$$\begin{aligned}\chi^2 &= \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(16 - 7.85)^2}{7.85} + \frac{(6 - 11.84)^2}{11.84} + \dots + \frac{(5 - 6.22)^2}{6.22} \\ &= 19.4.\end{aligned}$$

Calculate the p -value: for a table with 3 rows and 5 columns, the χ^2 statistic is distributed with $(3 - 1)(5 - 1) = 8$ degrees of freedom. From the table, a χ^2 value of 19.4 corresponds to a tail area between 0.01 and 0.02. Thus, there is sufficient evidence to reject the null hypothesis of independence between race and genotype.

The p -value can be obtained using the R function `pchisq` (`pchisq(19.4, df = 8, lower.tail = FALSE)`), which returns a value of 0.012861.

To further explore the differences in genotype distribution between races, calculate residuals for each cell (Figure 8.10). The largest residuals are in the first row; there are many more African Americans with the CC genotype than expected under independence, and fewer with the CT genotype than expected. The residuals in the second row indicate a similar trend for Asians, but with a less pronounced difference. These results suggest further directions for research; a future study could enroll a larger number of African American and Asian participants to examine whether the observed trend holds with a more representative sample. Geneticists might also be interested in exploring whether this genetic difference between populations has an observable phenotypic effect.

	CC	CT	TT	Sum
African American	16	6	5	27
Asian	21	18	16	55
Caucasian	125	216	126	467
Hispanic	4	10	9	23
Other	7	11	5	23
Sum	173	261	161	595

Figure 8.8: Observed counts for race and genotype data from the FAMuSS study.

	CC	CT	TT	Sum
African Am	7.85	11.84	7.31	27.00
Asian	15.99	24.13	14.88	55.00
Caucasian	135.78	204.85	126.36	467.00
Hispanic	6.69	10.09	6.22	23.00
Other	6.69	10.09	6.22	23.00
Sum	173.00	261.00	161.00	595.00

Figure 8.9: Expected counts for race and genotype data from the FAMuSS study.

EXAMPLE 8.9

In Guided Practice 8.7, the p -value was found to be smaller than 0.001, suggesting that treatment is not independent of outcome. Does the evidence suggest that infants should be given nevirapine or lopinarvir?

E

In a 2×2 table, it is relatively easy to directly compare observed and expected counts. For nevirapine, more infants than expected experienced virologic failure ($60 > 44.6$), while fewer than expected reached a stable disease state ($87 < 102.4$). For lopinarvir, fewer infants than expected experienced virologic failure ($27 < 42.4$), and more infants than expected reached a stable disease state ($113 > 97.6$) (Figure 8.11). The outcomes for infants on lopinarvir are better than for those on nevirapine; combined with the results of the significance test, the data suggest that lopinarvir is associated with better treatment outcomes.

G**GUIDED PRACTICE 8.10**

Confirm the conclusions reached in Example 8.9 by analyzing the residuals.⁵

GUIDED PRACTICE 8.11**G**

Chapter 1 started with the discussion of a study examining whether exposure to peanut products reduce the rate of a child developing peanut allergies. Children were randomized either to the peanut avoidance or the peanut consumption group; at 5 years of age, each child was tested for peanut allergy using an oral food challenge (OFC). The results of the OFC are reproduced in Figure 8.12; failing the food challenge indicates an allergic reaction. Assess whether there is evidence for exposure to peanut allergy reducing the chance of developing peanut allergies.⁶

⁵ $R_{1,1} = \frac{(44.6-60)}{\sqrt{44.6}} = 2.31$; $R_{1,2} = \frac{(42.4-27)}{\sqrt{27}} = -2.37$; $R_{2,1} = \frac{(87-102.4)}{\sqrt{102.4}} = -1.53$; $R_{2,2} = \frac{(113-97.6)}{\sqrt{97.6}} = 1.56$. The positive residuals for the upper left and lower right cells indicate that more infants than expected experienced virologic failure on NVP and stable disease on LPV; vice versa for the upper right and lower left cells. The larger magnitude of the residuals for the two NVP cells indicates that most of the discrepancy between observed and expected counts is for outcomes related to NVP.

⁶The assumptions for conducting a χ^2 test are satisfied. Calculate a χ^2 test statistic: 24.29. The associated p -value is 8.3×10^{-7} . There is evidence to suggest that treatment group is not independent of outcome. Specifically, a residual analysis

	CC	CT	TT	Sum
African Am	2.91	-1.70	-0.85	0.00
Asian	1.25	-1.25	0.29	0.00
Caucasian	-0.93	0.78	-0.03	0.00
Hispanic	-1.04	-0.03	1.11	0.00
Other	0.12	0.29	-0.49	0.00
Sum	0.00	0.00	0.00	0.00

Figure 8.10: Residuals for race and genotype data from the FAMuSS study.

	NVP	LPV	Total
Virologic Failure	60 44.6	27 42.4	87
Stable Disease	87 102.4	113 97.6	200
Total	147	140	287

Figure 8.11: Observed and (expected) counts for the HIV study.

shows that in the peanut avoidance group, more children than expected failed the OFC; in the peanut consumption group, more children than expected passed the OFC.

EXAMPLE 8.12

The objective of a study is to identify factors of postpartum endometritis in dairy cows. One of the factors of the study is the breed of the cow. The study has been performed between Holstein–Friesian, Jersey and the crossbreed between them. Two variables were considered: breed (3 possible values: Holstein–Friesian, Jersey and crossbreed) and whether they suffer from endometritis. For the study a sample of 300 dairy cattle have been selected. The result of the analysis have been the following 11 out of 50 Holstein–Friesian have endometritis, 7 out of 50 Jersey cows and 12 out of 200 crossbreed cows suffer from this symptom. These data are summarized in the contingency table of Figure 8.13

What the study is trying to assess is whether there exists a statistically significant evidence that the breed of the cattle is going to be an independent factor of endometritis or there exists a relation between the two variables. It's important to say that a relation between the two variables does not mean that one variable is the cause of the other variable. It's possible to find multiple relation between variable and none of this variable is the cause of the other. Therefore, the null and alternative hypothesis are the following:

$$\begin{cases} H_0: & \text{Breed and endometritis are independent (non-association)} \\ H_1: & \text{Breed and endometritis are dependent (association)} \end{cases}$$

In Table 8.14 is shown a column proportion table and it can be seen that the proportion of endometritis between the three groups are quite different. So, it can be thought that breeds and endometritis are dependent variables, but a more accurate method is using the chi-square statistic.

E

The expected value for every cell is calculated and a contingency table with this information is on the table below

$$\begin{aligned} E_{\text{Non-End,HF}} &= \frac{O_{\text{Non-End},\bullet} \cdot O_{\bullet,\text{HF}}}{n} = \frac{270 \cdot 50}{300} = 45.00 & E_{\text{End,HF}} &= \frac{O_{\text{End},\bullet} \cdot O_{\bullet,\text{HF}}}{n} = \frac{30 \cdot 50}{300} = 5.00 \\ E_{\text{Non-End,J}} &= \frac{O_{\text{Non-End},\bullet} \cdot O_{\bullet,\text{J}}}{n} = \frac{270 \cdot 50}{300} = 45.00 & E_{\text{End,J}} &= \frac{O_{\text{End},\bullet} \cdot O_{\bullet,\text{J}}}{n} = \frac{30 \cdot 50}{300} = 5.00 \\ E_{\text{Non-End,CB}} &= \frac{E_{\text{Non-End},\bullet} \cdot E_{\bullet,\text{CB}}}{n} = \frac{270 \cdot 200}{300} = 180.00 & E_{\text{End,HF}} &= \frac{E_{\text{End},\bullet} \cdot E_{\bullet,\text{HF}}}{n} = \frac{30 \cdot 200}{300} = 20.00 \end{aligned}$$

	Holstein–Friesian	Jersey	Crossbreed	Total
No endometritis	45	45	180	270
Endometritis	5	5	20	30
Total	50	50	200	

The χ^2 test statistic of the contingency table of the variables endometritis and breeds of the cows is calculated

$$\begin{aligned} \chi^2 &= \frac{(39 - 45)^2}{45} + \frac{(43 - 45)^2}{45} + \frac{(188 - 180)^2}{180} + \frac{(11 - 5)^2}{5} + \frac{(7 - 7)^2}{7} + \frac{(12 - 20)^2}{20} = \\ &= \frac{36}{45} + \frac{4}{45} + \frac{64}{180} + \frac{36}{5} + \frac{4}{5} + \frac{64}{20} = \frac{2240}{180} = 12.4444 \end{aligned}$$

The degree of freedom is $df = (R - 1) \cdot (C - 1) = 1 \cdot 2 = 2$ and the critical value is $\chi_{2,95} = 5.99$. Hence, the acceptance interval is (0, 5.99). The value 12.44 is outside of the interval. Therefore, the null hypothesis is rejected. It means that there exists enough evidence that breeds and postpartum endometritis are dependent variables and the proportions of endometritis is not the same between the breeds.

	FAIL OFC	PASS OFC	Sum
Peanut Avoidance	36	227	263
Peanut Consumption	5	262	267
Sum	41	489	530

Figure 8.12: LEAP Study Results.

	Holstein–Friesian	Jersey	Crossbreed	Total
No endometritis	39	43	188	270
Endometritis	11	7	12	30
Total	50	50	200	

Figure 8.13: Contingency table of the cows suffering from endometritis depending on the breed.

	Holstein–Friesian	Jersey	Crossbreed	Total
No endometritis	0.78	0.86	0.94	0.90
Endometritis	0.22	0.14	0.06	0.10

Figure 8.14: Table with the column proportions across groups

8.3 Exercises

Exercise. 8.1

On a farm that was using a determined chemical product for disinfection was detected that some workers started to ace some disorders of the respiratory system. The disinfectant was thought to be related to these disorders. 500 workers were selected to evaluate this hypothesis, being classified according to the level of exposure and whether or not the symptoms of these disorders. The results are presented

	Direct contact	Limited contact	No contact
Disorder symptoms	185	33	17
Asymptomatic	120	73	72

Do we have enough evidence to state that there exists a relation between the level of exposure and the presence of these symptoms among workers? Set and solve an adequate hypothesis testing, explain your conclusion from the study.

Exercise. 8.2

A study at a horse have the objective of a relation between body size of the horse and the presence of certain anomalies close to knees called “angular deformities”. to check this hypothesis, 500 horses were randomly selected and classified according to their body size within four categories and the presence of angular deformities. The outcome of this study was the following table:

	Low	Medium-low	Medium-high	High
Presence of angular deformities	8	24	32	27
Absence of angular deformities	42	121	138	108

Are these data compatible with the hypothesis that the presence of angular deformities is related to the body size of the horse? Use a significance level $\alpha = 0.05$. Set an adequate hypothesis testing to answer this question and draw your conclusions.

Exercise. 8.3

A study is being performed to compare two drugs for the acute pain provoke by bone fracture in horses. They have been labeled as D1 and D2. After treatment, every animal is classified within three categories: *pain removed*, *reduced intensity* and *changes not appreciated*. The drug D1 was given to 32 horses with bone fracture and 28 horses received drug D2 with the same situation. In 12 cases the pain was removed, being 7 horses with D1 treatment and others with treatment D2. The pain was reduced for 30 horses, 17 out of 30 had a intake of D1 and the remaining horses D2. Finally, from 18 horses where no changes were appreciated, 8 of them have received drug D1. Could we say that one drug is more effective than the other one to reduce the pain from bone fracture? Set an adequate hypothesis testing with significance level $\alpha = 0.05$ to answer this question and draw your conclusions.

Exercise. 8.4

It is well-known that the exposure of tobacco smoke is totally harmful to the health, destroying cells in the body. One of the possible effects is thought to be a progressive muscle atrophy. With the objective of confirming that smoking affects the mass of the muscle, an experiment was performed with lab rat. One group was exposed to tobacco smoke and the other group not. The muscle protein synthesis was measured by mean of the protein content in blood and the participants were classified depending on the values was between the 10 and 90-percentile, lower than 10-percentile or higher than 90-percentile of the population.

	< 10-perc.	Betw. 10 and 90-perc.	> 90-perc.
rats exposed to tobacco smoke	117	529	19
rats not exposed to tobacco smoke	124	1147	117

Is there enough evidence in favour of an association of the protein content and smoking? Set an adequate hypothesis testing using $\alpha = 0.1$, calculate test statistics, acceptance interval and draw your conclusions.

Exercise. 8.5

At the beginning of the last flu season, Orange County Community Hospital staff administered a flu vaccine. To determine if there is an association between the flu vaccine and contraction of the flu, 235 of the inoculated patients and 188 uninoculated subjects were randomly sampled. The investigators monitored the study subjects throughout the flu season and gathered the data shown in the table.⁷

Flu vaccine	Contracted Flu		
	Yes	No	
Yes	43	192	235
No	51	137	188
	94	329	423

1. State the appropriate null hypothesis to test for an association between two variables.
2. Find the expected frequency for each cell.
3. Find the value of the test statistic.
4. State the conclusion of the significance test.

⁷J.S. Kim and R.J. Dailey. *Biostatistics for Oral Healthcare*. Wiley, 2008. ISBN: 9780470388273. URL: <https://books.google.es/books?id=n1fR0LF1jhsC>.

Chapter 9

Simple linear regression

9.2 Examining scatterplots

9.3 Estimating a regression line using least squares

9.4 Interpreting a linear model

9.5 Statistical inference with regression

9.6 Interval estimates with regression

9.7 Notes

The relationship between two numerical variables can be visualized using a scatterplot in the xy -plane. The **predictor** or **explanatory variable** is plotted on the horizontal axis, while the **response variable** is plotted on the vertical axis.¹

This chapter explores simple linear regression, a technique for estimating a straight line that best fits data on a scatterplot.² A line of best fit functions as a linear model that can not only be used for prediction, but also for inference. Linear regression should only be used with data that exhibit linear or approximately linear relationships.

For example, scatterplots in Chapter 1 illustrated the linear relationship between height and weight in the NHANES data, with height as a predictor of weight. Adding a best-fitting line to these data using regression techniques would allow for prediction of an individual's weight based on their height. The linear model could also be used to investigate questions about the population-level relationship between height and weight, since the data are a random sample from the population of adults in the United States.

Not all relationships in data are linear. For example, the scatterplot in Figure 9.3 of Chapter 1 shows a highly non-linear relationship between annual per capita income and life expectancy for 165 countries in 2011. Relationships are called **strong relationships** if the pattern of the dependence between the predictor and response variables is clear, even if it is nonlinear as in Figure 9.3. A **weak relationship** is one in which the points in the scatterplot are so diffuse as to make it difficult to discern any relationship. Figure 9.4 in Chapter 1 showed relationships progressing from weak to strong moving from left to right in the top and bottom panels. Each of the relationships shown in the second panels from the left are **moderate relationships**. Finally, changing the scale of measurement of one or both variables, such as changing age from age in years to age in months, simply stretches or compresses one or both axes and does not change the nature of the relationship. If a relationship is linear it will remain so, and with a simple change of scale, a nonlinear relationship will remain nonlinear.

¹ Sometimes, the predictor variable is referred to as the independent variable, and the response variable referred to as the dependent variable.

² Although the response variable in linear regression is necessarily numerical, the predictor variable can be numerical or categorical.

The next chapter covers multiple regression, a statistical model used to estimate the relationship between a single numerical response variable and several predictor variables.



For labs, slides, and other resources, please visit
www.openintro.org/book/biostat

9.1 Summaries of two quantitative variables

Scatterplots

In the frog parental investment study, researchers used clutch volume as a primary variable of interest rather than egg size because clutch volume represents both the eggs and the protective gelatinous matrix surrounding the eggs. The larger the clutch volume, the higher the energy required to produce it; thus, higher clutch volume is indicative of increased maternal investment. Previous research has reported that larger body size allows females to produce larger clutches; is this idea supported by the frog data?

A **scatterplot** provides a case-by-case view of the relationship between two numerical variables. Figure 9.1 shows clutch volume plotted against body size, with clutch volume on the y -axis and body size on the x -axis. Each point represents a single case. For this example, each case is one egg clutch for which both volume and body size (of the female that produced the clutch) have been recorded.

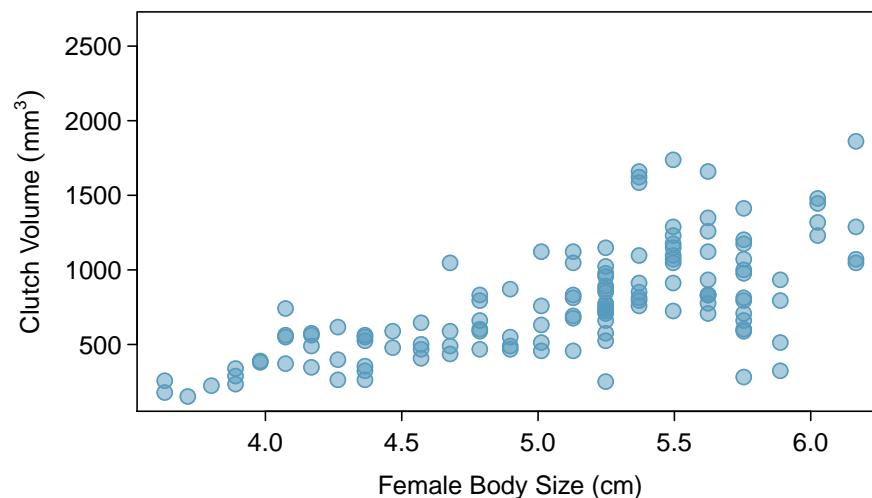


Figure 9.1: A scatterplot showing `clutch.volume` (vertical axis) vs. `body.size` (horizontal axis).

The plot shows a discernible pattern, which suggests an **association**, or relationship, between clutch volume and body size; the points tend to lie in a straight line, which is indicative of a **linear association**. Two variables are **positively associated** if increasing values of one tend to occur with increasing values of the other; two variables are **negatively associated** if increasing values of one variable occurs with decreasing values of the other. If there is no evident relationship between two variables, they are said to be **uncorrelated** or **independent**.

As expected, clutch volume and body size are positively associated; larger frogs tend to produce egg clutches with larger volumes. These observations suggest that larger females are capable of investing more energy into offspring production relative to smaller females.

The National Health and Nutrition Examination Survey (NHANES) consists of a set of surveys and measurements conducted by the US CDC to assess the health and nutritional status of adults and children in the United States. The following example uses data from a sample of 500 adults (individuals ages 21 and older) from the NHANES dataset.³

³The sample is available as `nhanes.samp.adult.500` in the R `oiobiostat` package.

EXAMPLE 9.1

Body mass index (BMI) is a measure of weight commonly used by health agencies to assess whether someone is overweight, and is calculated from height and weight.⁴ Describe the relationships shown in Figure 9.2. Why is it helpful to use BMI as a measure of obesity, rather than weight?

Figure 9.2(a) shows a positive association between height and weight; taller individuals tend to be heavier. Figure 9.2(b) shows that height and BMI do not seem to be associated; the range of BMI values observed is roughly consistent across height.

E

Weight itself is not a good measure of whether someone is overweight; instead, it is more reasonable to consider whether someone's weight is unusual relative to other individuals of a comparable height. An individual weighing 200 pounds who is 6 ft tall is not necessarily an unhealthy weight; however, someone who weighs 200 pounds and is 5 ft tall is likely overweight. It is not reasonable to classify individuals as overweight or obese based only on weight.

BMI acts as a relative measure of weight that accounts for height. Specifically, BMI is used as an estimate of body fat. According to US National Institutes of Health (US NIH) and the World Health Organization (WHO), a BMI between 25.0 - 29.9 is considered overweight and a BMI over 30 is considered obese.⁵

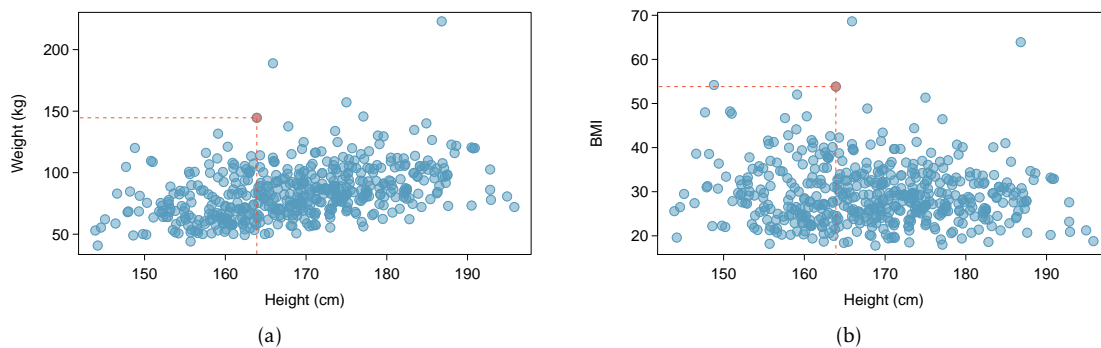


Figure 9.2: (a) A scatterplot showing height versus weight from the 500 individuals in the sample from NHANES. One participant 163.9 cm tall (about 5 ft, 4 in) and weighing 144.6 kg (about 319 lb) is highlighted. (b) A scatterplot showing height versus BMI from the 500 individuals in the sample from NHANES. The same individual highlighted in (a) is marked here, with BMI 53.83.

$$^4 BMI = \frac{weight_{kg}}{height_m^2} = \frac{weight_{lb}}{height_{in}^2} \times 703.$$

⁵https://www.nhlbi.nih.gov/health/educational/lose_wt/risk.htm

EXAMPLE 9.2

Figure 9.3 is a scatterplot of life expectancy versus annual per capita income for 165 countries in 2011. Life expectancy is measured as the expected lifespan for children born in 2011 and income is adjusted for purchasing power in a country. Describe the relationship between life expectancy and annual per capita income; do they seem to be linearly associated?

E

Life expectancy and annual per capita income are positively associated; higher per capita income is associated with longer life expectancy. However, the two variables are not linearly associated. When income is low, small increases in per capita income are associated with relatively large increases in life expectancy. However, once per capita income exceeds approximately \$20,000 per year, increases in income are associated with smaller gains in life expectancy.

In a linear association, change in the y -variable for every unit of the x -variable is consistent across the range of the x -variable; for example, a linear association would be present if an increase in income of \$10,000 corresponded to an increase in life expectancy of 5 years, across the range of income.

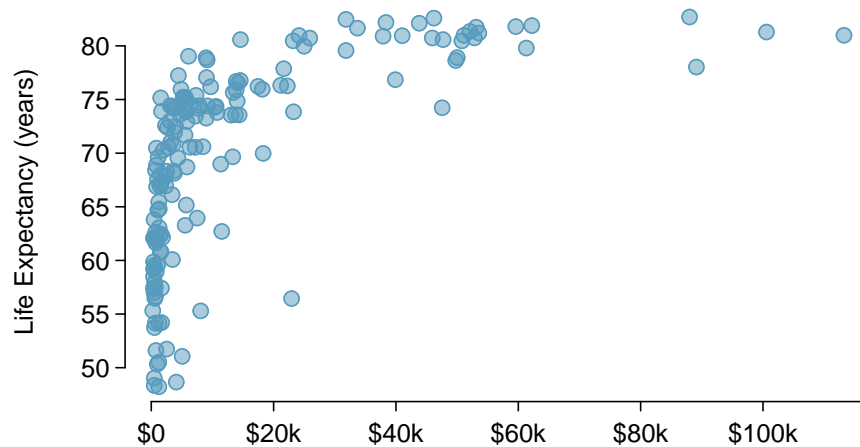


Figure 9.3: A scatterplot of life expectancy (years) versus annual per capita income (US dollars) in the wdi . 2011 dataset.

Correlation

Correlation is a numerical summary statistic that measures the strength of a linear relationship between two variables. It is denoted by r , the **correlation coefficient**, which takes on values between -1 and 1.

If the paired values of two variables lie exactly on a line, $r = \pm 1$; the closer the correlation coefficient is to ± 1 , the stronger the linear association. When two variables are positively associated, with paired values that tend to lie on a line with positive slope, $r > 0$. If two variables are negatively associated, $r < 0$. A value of r that is 0 or approximately 0 indicates no apparent association between two variables.⁶

The correlation coefficient quantifies the strength of a linear trend. Prior to calculating a correlation, it is advisable to confirm that the data exhibit a linear relationship. Although it is mathematically possible to calculate correlation for any set of paired observations, such as the life

⁶If paired values lie perfectly on either a horizontal or vertical line, there is no association and r is mathematically undefined.

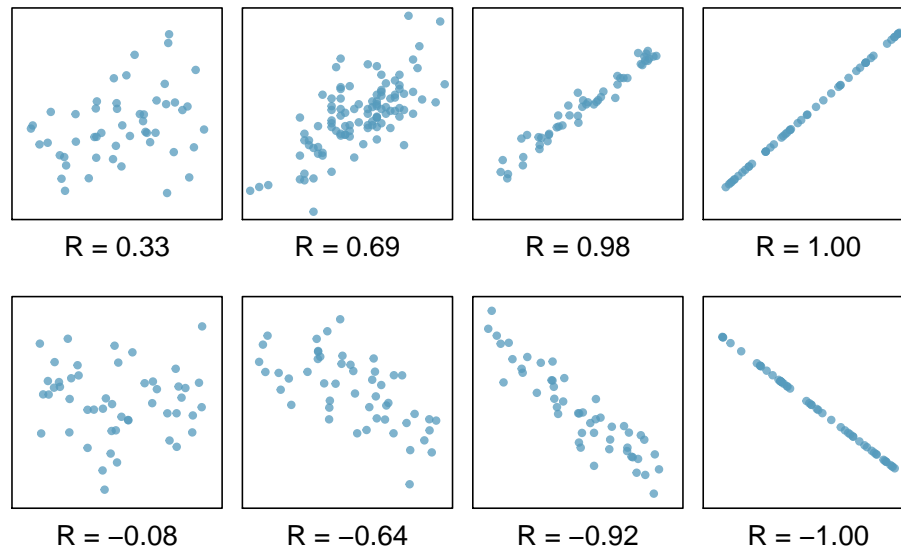


Figure 9.4: Scatterplots and their correlation coefficients. The first row shows positive associations and the second row shows negative associations. From left to right, strength of the linear association between x and y increases.

expectancy versus income data in Figure 9.3, correlation cannot be used to assess the strength of a nonlinear relationship.

CORRELATION

The correlation between two variables x and y is given by:

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right), \quad (9.3)$$

where $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ are the n paired values of x and y , and s_x and s_y are the sample standard deviations of the x and y variables, respectively.

EXAMPLE 9.4

Calculate the correlation coefficient of x and y , plotted in Figure 9.5.

Calculate the mean and standard deviation for x and y : $\bar{x} = 2$, $\bar{y} = 3$, $s_x = 1$, and $s_y = 2.65$.

E

$$\begin{aligned} r &= \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \\ &= \frac{1}{3-1} \left[\left(\frac{1-2}{1} \right) \left(\frac{5-3}{2.65} \right) + \left(\frac{2-2}{1} \right) \left(\frac{4-3}{2.65} \right) + \left(\frac{3-2}{1} \right) \left(\frac{0-3}{2.65} \right) \right] \\ &= -0.94. \end{aligned}$$

The correlation is -0.94, which reflects the negative association visible from the scatterplot in Figure 9.5.

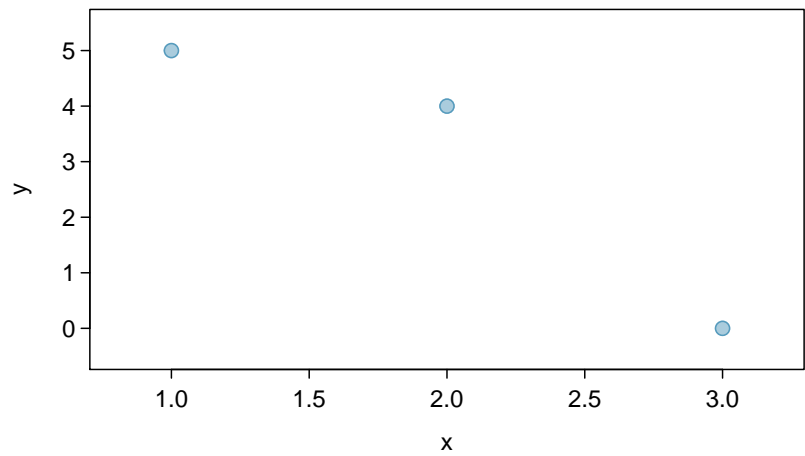


Figure 9.5: A scatterplot showing three points: (1, 5), (2, 4), and (3, 0).

EXAMPLE 9.5

Is it appropriate to use correlation as a numerical summary for the relationship between life expectancy and income after a log transformation is applied to both variables? Refer to Figure 9.6.

E

Figure 9.6 shows an approximately linear relationship; a correlation coefficient is a reasonable numerical summary of the relationship. As calculated from statistical software, $r = 0.79$, which is indicative of a strong linear relationship.

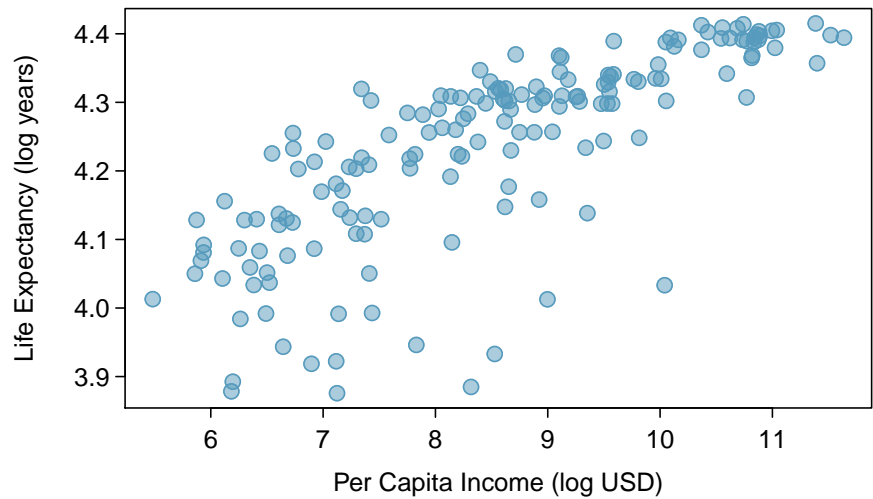


Figure 9.6: A scatterplot showing $\log(\text{income})$ (horizontal axis) vs. $\log(\text{life expectancy})$ (vertical axis).

9.2 Examining scatterplots

Various demographic and cardiovascular risk factors were collected as a part of the Prevention of REnal and Vascular END-stage Disease (PREVEND) study, which took place in the Netherlands. The initial study population began as 8,592 participants aged 28-75 years who took a first survey in 1997-1998.⁷ Participants were followed over time; 6,894 participants took a second survey in 2001-2003, and 5,862 completed the third survey in 2003-2006. In the third survey, measurement of cognitive function was added to the study protocol. Data from 4,095 individuals who completed cognitive testing are in the prevend dataset, available in the R package *oibiostat*.

As adults age, cognitive function changes over time, largely due to various cerebrovascular and neurodegenerative changes. It is thought that cognitive decline is a long-term process that may start as early as 45 years of age.⁸ The Ruff Figural Fluency Test (RFFT) is one measure of cognitive function that provides information about cognitive abilities such as planning and the ability to switch between different tasks. The test consists of drawing as many unique designs as possible from a pattern of dots, under timed conditions; scores range from 0 to 175 points (worst and best score, respectively).

RFFT scores for a random sample of 500 individuals are shown in Figure 9.7, plotted against age at enrollment, which is measured in years. The variables Age and RFFT are negatively associated; older participants tend to have lower cognitive function. There is an approximately linear trend observable in the data, which suggests that adding a line could be useful for summarizing the relationship between the two variables.

It is important to avoid adding straight lines to non-linear data, such as in Figure 9.3.

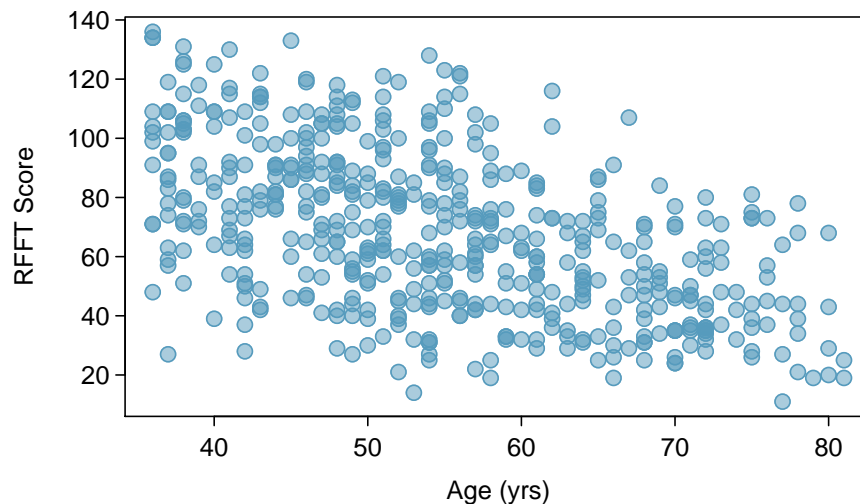


Figure 9.7: A scatterplot showing age vs. RFFT. Age is the predictor variable, while RFFT score is the response variable.

⁷Participants were selected from the city of Groningen on the basis of their urinary albumin excretion; urinary albumin excretion is known to be associated with abnormalities in renal function.

⁸Joosten H, et al. Cardiovascular risk profile and cognitive function in young, middle-aged, and elderly subjects. *Stroke*. 2013;44:1543-1549, <https://doi.org/10.1161/STROKEAHA.111.000496>

The following conditions should be true in a scatterplot for a line to be considered a reasonable approximation to the relationship in the plot and for the application of the methods of inference discussed later in the chapter:

- 1 Linearity.** The data shows a linear trend. If there is a nonlinear trend, an advanced regression method should be applied; such methods are not covered in this text. Occasionally, a transformation of the data will uncover a linear relationship in the transformed scale.
- 2 Constant variability.** The variability of the response variable about the line remains roughly constant as the predictor variable changes.
- 3 Independent observations.** The (x, y) pairs are independent; i.e., the value of one pair provides no information about other pairs. Be cautious about applying regression to sequential observations in time (**time series** data), such as height measurements taken over the course of several years. Time series data may have a complex underlying structure, and the relationship between the observations should be accounted for in a model.
- 4 Residuals that are approximately normally distributed.** This condition can be checked only after a line has been fit to the data and will be explained in Section 9.4.1, where the term residual is defined. In large datasets, it is sufficient for the residuals to be approximately symmetric with only a few outliers. This condition becomes particularly important when inferences are made about the line, as discussed in Section 9.5.

GUIDED PRACTICE 9.6



Figure 9.8 shows the relationship between `clutch.volume` and `body.size` in the frog data. The plot also appears as Figure 9.1 in Chapter 1. Are the first three conditions met for linear regression?⁹

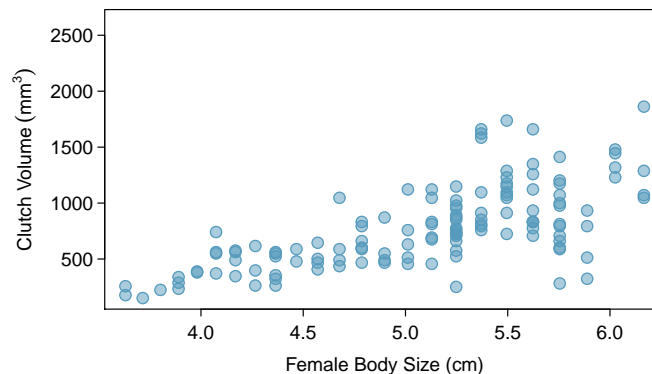


Figure 9.8: A plot of `clutch.volume` versus `body.size` in the frog data.

⁹No. While the relationship appears linear and it is reasonable to assume the observations are independent (based on information about the frogs given in Chapter 1), the variability in `clutch.volume` is noticeably less for smaller values of `body.size` than for larger values.

9.3 Estimating a regression line using least squares

Figure 9.9 shows the scatterplot of age versus RFFT score, with the **least squares regression line** added to the plot; this line can also be referred to as a **linear model** for the data. An RFFT score can be predicted for a given age from the equation of the regression line:

$$\widehat{\text{RFFT}} = 137.55 - 1.26(\text{age}).$$

The vertical distance between a point in the scatterplot and the predicted value on the regression line is the **residual** for the observation represented by the point; observations below the line have negative residuals, while observations above the line have positive residuals. The size of a residual is usually discussed in terms of its absolute value; for example, a residual of -13 is considered larger than a residual of 5 .

For example, consider the predicted RFFT score for an individual of age 56. According to the linear model, this individual has a predicted score of $137.550 - 1.261(56) = 66.934$ points. In the data, however, there is a participant of age 56 with an RFFT score of 72; their score is about 5 points higher than predicted by the model (this observation is shown on the plot with a “x”).

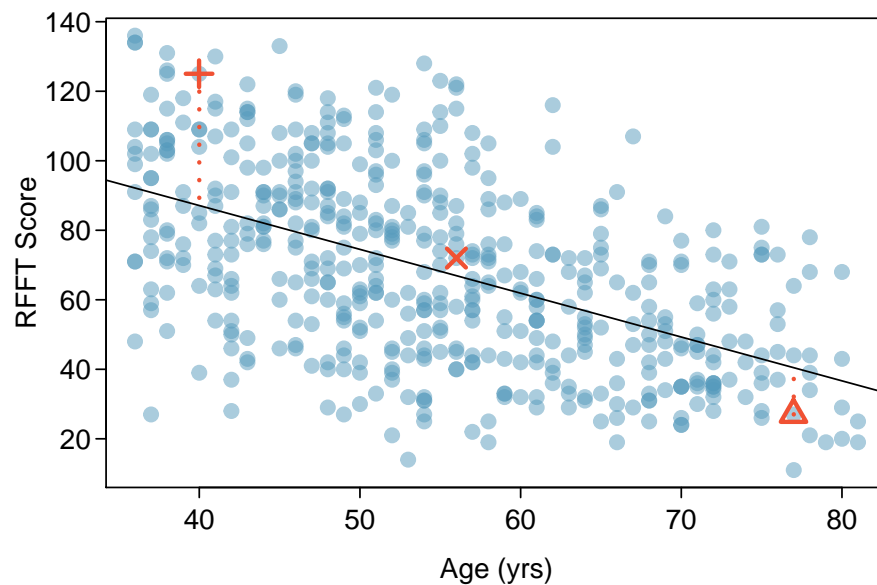


Figure 9.9: A scatterplot showing age (horizontal axis) vs. RFFT (vertical axis) with the regression line added to the plot. Three observations are marked in the figure; the one marked by a “+” has a large residual of about +38, the one marked by a “x” has a small residual of about +5, and the one marked by a “△” has a moderate residual of about -13. The vertical dotted lines extending from the observations to the regression line represent the residuals.

RESIDUAL: DIFFERENCE BETWEEN OBSERVED AND EXPECTED

The residual of the i^{th} observation (x_i, y_i) is the difference of the observed response (y_i) and the response predicted based on the model fit (\hat{y}_i) :

$$e_i = y_i - \hat{y}_i$$

The value \hat{y}_i is calculated by plugging x_i into the model equation.

The **least squares regression line** is the line which minimizes the sum of the squared residuals for all the points in the plot. Let \hat{y}_i be the predicted value for an observation with value x_i for the explanatory variable. The value $e_i = y_i - \hat{y}_i$ is the residual for a data point (x_i, y_i) in a scatterplot with n pairs of points. The least squares line is the line for which

$$e_1^2 + e_2^2 + \cdots + e_n^2 \quad (9.7)$$

is smallest.

For a general population of ordered pairs (x, y) , the **population regression model** is

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

The term ε is a normally distributed ‘error term’ that has mean 0 and standard deviation σ . Since $E(\varepsilon) = 0$, the model can also be written

$$E(Y|x) = \beta_0 + \beta_1 x,$$

where the notation $E(Y|x)$ denotes the expected value of Y when the predictor variable has value x .¹⁰ For the PREVEND data, the population regression line can be written as

$$\text{RFFT} = \beta_0 + \beta_1(\text{age}) + \varepsilon, \text{ or as } E(\text{RFFT}|\text{age}) = \beta_0 + \beta_1(\text{age}).$$

The term β_0 is the vertical intercept for the line (often referred to simply as the intercept) and β_1 is the slope. The notation b_0 and b_1 are used to represent the point estimates of the parameters β_0 and β_1 . The point estimates b_0 and b_1 are estimated from data; β_0 and β_1 are parameters from the population model for the regression line.

The regression line can be written as $\hat{y} = b_0 + b_1(x)$, where \hat{y} represents the predicted value of the response variable. The slope of the least squares line, b_1 , is estimated by

$$b_1 = \frac{s_y}{s_x} r, \quad (9.8)$$

where r is the correlation between the two variables, and s_x and s_y are the sample standard deviations of the explanatory and response variables, respectively. The intercept for the regression line is estimated by

$$b_0 = \bar{y} - b_1 \bar{x}. \quad (9.9)$$

Typically, regression lines are estimated using statistical software.

¹⁰The error term ε can be thought of as a population parameter for the residuals (e). While ε is a theoretical quantity that refers to the deviation between an observed value and $E(Y|x)$, a residual is calculated as the deviation between an observed value and the prediction from the linear model.

EXAMPLE 9.10

From the summary statistics displayed in Figure 9.10 for `prevend.samp`, calculate the equation of the least-squares regression line for the PREVEND data.

$$b_1 = \frac{s_y}{s_x} r = \frac{27.40}{11.60} (-0.534) = -1.26$$

$$b_0 = \bar{y} - b_1 \bar{x} = 68.40 - (-1.26)(54.82) = 137.55.$$

The results agree with the equation shown at the beginning of this section:

$$\widehat{\text{RFFT}} = 137.55 - 1.26(\text{age}).$$

	Age (yrs)	RFFT score
mean	$\bar{x} = 54.82$	$\bar{y} = 68.40$
standard deviation	$s_x = 11.60$	$s_y = 27.40$
		$r = -0.534$

Figure 9.10: Summary statistics for age and RFFT from `prevend.samp`.

GUIDED PRACTICE 9.11

Figure 9.11 shows the relationship between height and weight in a sample from the NHANES dataset introduced in Chapter 1. Calculate the equation of the regression line given the summary statistics: $\bar{x} = 168.78, \bar{y} = 83.83, s_x = 10.29, s_y = 21.04, r = 0.410$.¹¹

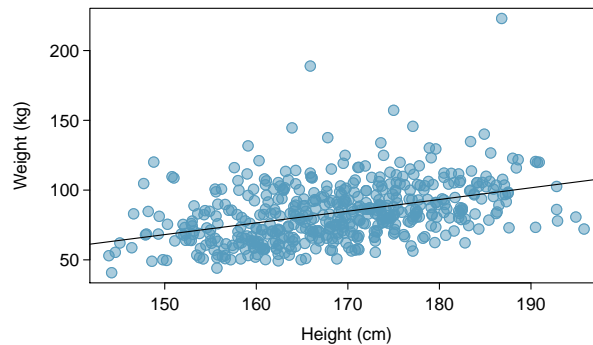


Figure 9.11: A plot of Height versus Weight in `nhanes.samp.adult.500`, with a least-squares regression line

GUIDED PRACTICE 9.12

Predict the weight in pounds for an adult who is 5 feet, 11 inches tall. 1 cm = .3937 in; 1 lb = 0.454 kg.¹²

¹¹ The equation of the line is $\widehat{\text{weight}} = -57.738 + 0.839(\text{height})$, where height is in centimeters and weight is in kilograms.

¹² 5 feet, 11 inches equals $71/.3937 = 180.34$ centimeters. From the regression equation, the predicted weight is $-57.738 + 0.839(180.34) = 93.567$ kilograms. In pounds, this weight is $93.567/0.454 = 206.280$.

9.4 Interpreting a linear model

A least squares regression line functions as a statistical model that can be used to estimate the relationship between an explanatory and response variable. While the calculations for constructing a regression line are relatively simple, interpreting the linear model is not always straightforward. In addition to discussing the mathematical interpretation of model parameters, this section also addresses methods for assessing whether a linear model is an appropriate choice, interpreting categorical predictors, and identifying outliers.

The slope parameter of the regression line specifies how much the line rises (positive slope) or declines (negative slope) for one unit of change in the explanatory variable. In the PREVENT data, the line decreases by 1.26 points for every increase of 1 year. However, it is important to clarify that RFFT score *tends* to decrease as age increases, with *average* RFFT score decreasing by 1.26 points for each additional year of age. As visible from the scatter of the data around the line, the line does not perfectly predict RFFT score from age; if this were the case, all the data would fall exactly on the line.

When interpreting the slope parameter, it is also necessary to avoid phrasing indicative of a causal relationship, since the line describes an association from data collected in an observational study. From these data, it is not possible to conclude that increased age causes a decline in cognitive function.¹³

Mathematically, the intercept on the vertical axis is a predicted value on the line when the explanatory variable has value 0. In biological or medical examples, 0 is rarely a meaningful value of the explanatory variable. For example, in the PREVENT data, the linear model predicts a score of 137.55 when age is 0—however, it is nonsensical to predict an RFFT score for a newborn infant.

In fact, least squares lines should never be used to extrapolate values outside the range of observed values. Since the PREVENT data only includes participants between ages 36 and 81, it should not be used to predict RFFT scores for people outside that age range. The nature of a relationship may change for very small or very large values of the explanatory variable; for example, if participants between ages 15 and 25 were studied, a different relationship between age and RFFT scores might be observed. Even making predictions for values of the explanatory variable slightly larger than the minimum or slightly smaller than the maximum can be dangerous, since in many datasets, observations near the minimum or maximum values (of the explanatory variable) are sparse.

Linear models are useful tools for summarizing a relationship between two variables, but it is important to be cautious about making potentially misleading claims based on a regression line. The following subsection discusses two commonly used approaches for examining whether a linear model can reasonably be applied to a dataset.

9.4.1 Checking residuals from a linear model

Recall that there are four assumptions that must be met for a linear model to be considered reasonable: linearity, constant variability, independent observations, normally distributed residuals. In the PREVENT data, the relationship between RFFT score and age appears approximately linear, and it is reasonable to assume that the data points are independent. To check the assumptions of constant variability around the line and normality of the residuals, it is helpful to consult residual plots and normal probability plots.¹⁴

¹³Similarly, avoid language such as increased age *leads to* or *produces* lower RFFT scores.

¹⁴While simple arithmetic can be used to calculate the residuals, the size of most datasets makes hand calculations impractical. The plots here are based on calculations done in R.

Examining patterns in residuals

There are a variety of residual plots used to check the fit of a least squares line. The plots shown in this text are scatterplots in which the residuals are plotted on the vertical axis against predicted values from the model on the horizontal axis. Other residual plots may instead show values of the explanatory variable or the observed response variable on the horizontal axis. When a least squares line fits data very well, the residuals should scatter about the horizontal line $y = 0$ with no apparent pattern.

Figure 9.12 shows three residual plots from simulated data; the plots on the right show data plotted with the least squares regression line, and the plots on the left show residuals on the y -axis and predicted values on the x -axis. A linear model is a particularly good fit for the data in the first row, where the residual plot shows random scatter above and below the horizontal line. In the second row, the original data cycles below and above the regression line; this nonlinear pattern is more evident in the residual plot. In the last row, the variability of the residuals is not constant; the residuals are slightly more variable for larger predicted values.

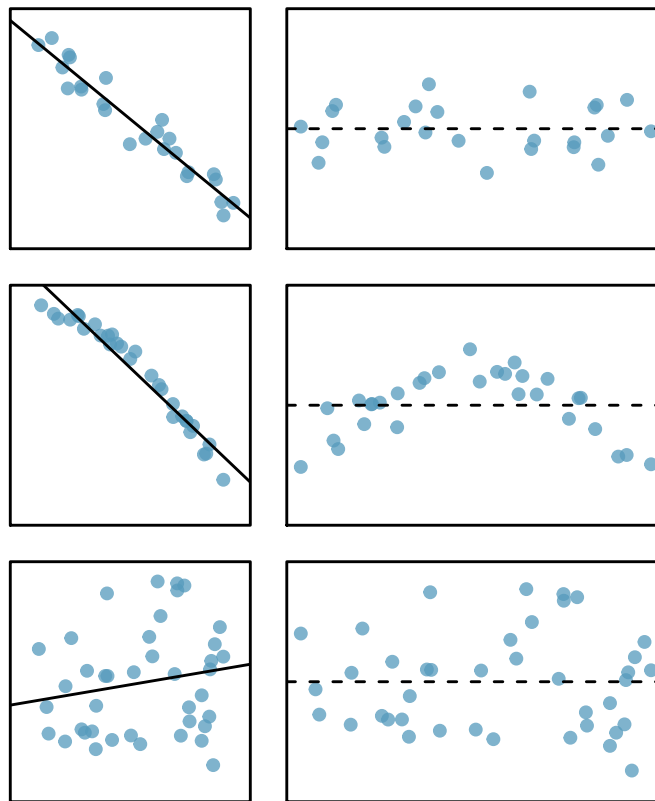


Figure 9.12: Sample data with their best fitting lines (left) and their corresponding residual plots (right).

Figure 9.13 shows a residual plot from the estimated linear model $\widehat{\text{RFFT}} = 137.55 - 1.26(\text{age})$. While the residuals show scatter around the line, there is less variability for lower predicted RFFT scores. A data analyst might still decide to use the linear model, with the knowledge that predictions of high RFFT scores may not be as accurate as for lower scores. Reading a residual plot critically can reveal weaknesses about a linear model that should be taken into account when interpreting model results. More advanced regression methods beyond the scope of this text may be more suitable for these data.

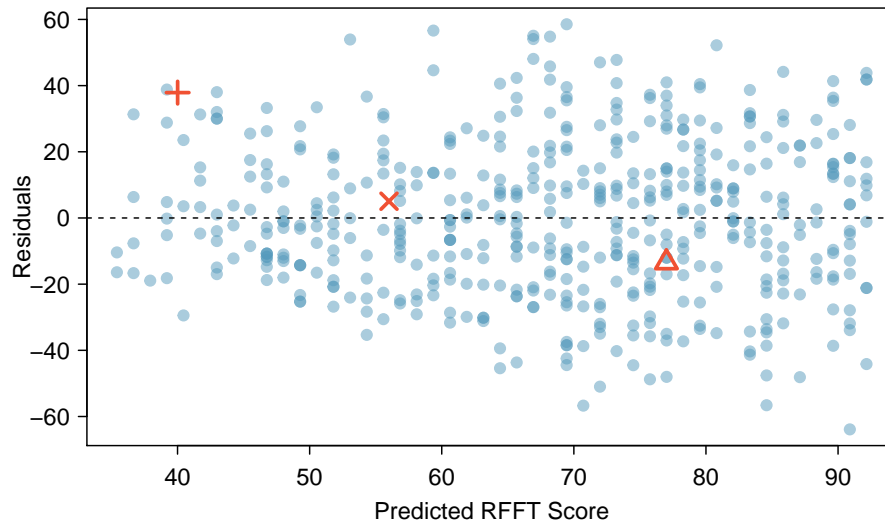


Figure 9.13: Residual plot for the model in Figure 9.9 using `prevend.samp`.

EXAMPLE 9.13

Figure 9.14 shows a residual plot for the model predicting weight from height using the sample of 500 adults from the NHANES data, `nhanes.samp.adult.500`. Assess whether the constant variability assumption holds for the linear model.

E

The residuals above the line are more variable, taking on more extreme values than those below the line. Larger than expected residuals imply that there are many large weights that are under-predicted; in other words, the model is less accurate at predicting relatively large weights.

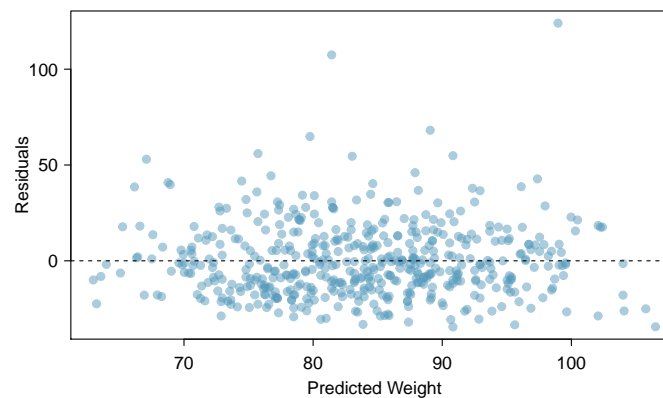


Figure 9.14: A residual plot from the linear model for height versus weight in `nhanes.samp.adult.500`.

9.4.2 Using R^2 to describe the strength of a fit

The correlation coefficient r measures the strength of the linear relationship between two variables. However, it is more common to measure the strength of a linear fit using r^2 , which is commonly written as R^2 in the context of regression.¹⁵

The quantity R^2 describes the amount of variation in the response that is explained by the least squares line. While R^2 can be easily calculated by simply squaring the correlation coefficient, it is easier to understand the interpretation of R^2 by using an alternative formula:

$$R^2 = \frac{\text{variance of predicted } y\text{-values}}{\text{variance of observed } y\text{-values}}.$$

It is possible to show that R^2 can also be written

$$R^2 = \frac{s_y^2 - s_{\text{residuals}}^2}{s_y^2}.$$

In the linear model predicting RFFT scores from age, the predicted values on the least squares line are the values of RFFT that are 'explained' by the linear model. The variability of the residuals about the line represents the remaining variability after the prediction; i.e., the variability unexplained by the model. For example, if a linear model perfectly captured all the data, then the variance of the predicted y -values would be equal to the variance of the observed y -values, resulting in $R^2 = 1$. In the linear model for \widehat{RFFT} , the proportion of variability explained is

$$R^2 = \frac{s_{\widehat{RFFT}}^2 - s_{\text{residuals}}^2}{s_{\widehat{RFFT}}^2} = \frac{750.52 - 536.62}{750.52} = \frac{213.90}{750.52} = 0.285,$$

about 29%. This is equal to the square of the correlation coefficient, $r^2 = (-0.534)^2 = 0.285$.

Since R^2 in simple linear regression is simply the square of the correlation coefficient between the predictor and the response, it does not add a new tool to regression. It becomes much more useful in models with several predictors, where it has the same interpretation as the proportion of variability explained by a model but is no longer the square of any one of the correlation coefficients between the individual responses and the predictor. Those models are not discussed in this book.

GUIDED PRACTICE 9.14

(G)

In the NHANES data, the variance of Weight is 442.53 kg² and the variance of the residuals is 368.1. What proportion of the variability in the data is explained by the model?¹⁶

GUIDED PRACTICE 9.15

(G)

If a linear model has a very strong negative relationship with a correlation of -0.97, how much of the variation in the response is explained by the explanatory variable?¹⁷

¹⁵In software output, R^2 is usually labeled **R-squared**.

¹⁶About 16.8%: $\frac{s_{\text{weight}}^2 - s_{\text{residuals}}^2}{s_{\text{weight}}^2} = \frac{442.53 - 368.1}{442.53} = \frac{74.43}{442.53} = 0.168$

¹⁷About $R^2 = (-0.97)^2 = 0.94$ or 94% of the variation is explained by the linear model.

9.4.3 Categorical predictors with two levels

Although the response variable in linear regression is necessarily numerical, the predictor variable may be either numerical or categorical. This section explores the association between a country's infant mortality rate and whether or not 50% of the population has access to adequate sanitation facilities.

The World Development Indicators (WDI) is a database of country-level variables (i.e., indicators) recording outcomes for a variety of topics, including economics, health, mortality, fertility, and education.¹⁸ The dataset `wdi.2011` contains a subset of variables on 165 countries from the year 2011.¹⁹ The infant mortality rate in a country is recorded as the number of deaths in the first year of life per 1,000 live births. Access to sanitation is recorded as the percentage of the population with adequate disposal facilities for human waste. Due to the availability of death certificates, infant mortality is measured reasonably accurately throughout the world. However, it is more difficult to obtain precise measurements of the percentage of a population with access to adequate sanitation facilities; instead, considering whether half the population has such access may be a more reliable measure. The analysis presented here is based on 163 of the 165 countries; the values for access to sanitation are missing for New Zealand and Turkmenistan.

Figure 9.15(a) shows that infant mortality rates are highly right-skewed, with a relatively small number of countries having high infant mortality rates. In 13 countries, infant mortality rates are higher than 70 deaths per thousand live births. Figure 9.15(b) shows infant mortality after a log transformation; the following analysis will use the more nearly symmetric transformed version of `inf.mortality`.

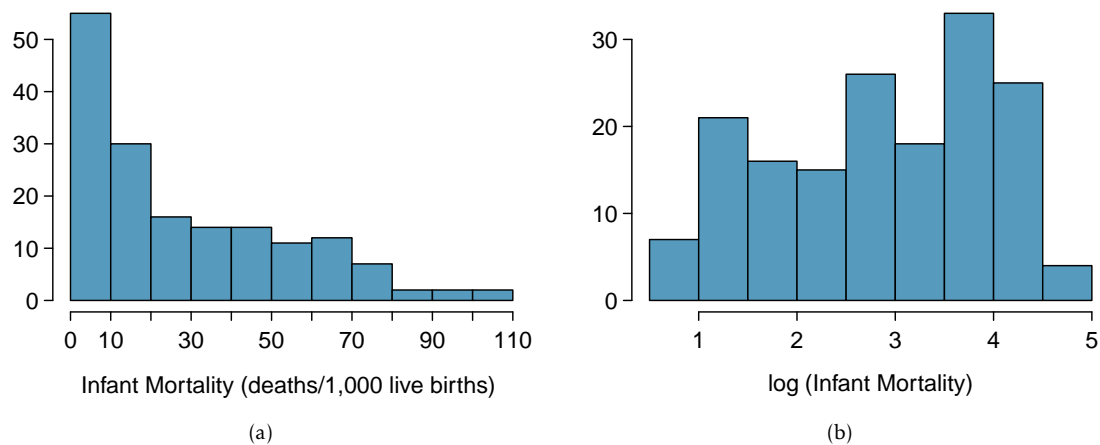


Figure 9.15: (a) Histogram of infant mortality, measured in deaths per 1,000 live births in the first year of life. (b) Histogram of the log-transformed infant mortality.

Figure 9.16 shows a scatterplot of $\log(\text{inf.mortality})$ against the categorical variable for sanitation access, coded 1 if at least 50% of the population has access to adequate sanitation, and 0 otherwise. Since there are only two values of the predictor, the values of infant mortality are stacked above the two predictor values 0 and 1.²⁰

¹⁸<http://data.worldbank.org/data-catalog/world-development-indicators>

¹⁹The data were collected by a Harvard undergraduate in the Statistics department, and are accessible via the `oibiostat` package.

²⁰Typically, side-by-side boxplots are used to display the relationship between a numerical variable and a categorical variable. In a regression context, it can be useful to use a scatterplot instead, in order to see the variability around the regression line.

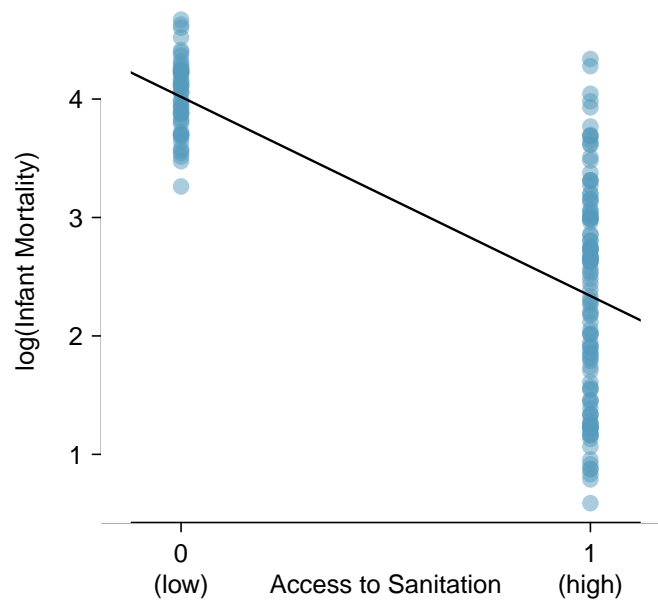


Figure 9.16: Country-level infant mortality rates, divided into low access ($x = 0$) and high access ($x = 1$) to sanitation. The least squares regression line is also shown.

The least squares regression line has the form

$$\widehat{\log(\text{inf.mortality})} = b_0 + b_1(\text{sanit.access}). \quad (9.16)$$

The estimated least squares regression line has intercept and slope parameters of 4.018 and -1.681, respectively. While the scatterplot appears unlike those for two numerical variables, the interpretation of the parameters remains unchanged. The slope, -1.681, is the estimated change in the logarithm of infant mortality when the categorical predictor changes from low access to sanitation facilities to high access. The intercept term 4.018 is the estimated log infant mortality for the set of countries where less than 50% of the population has access to adequate sanitation facilities ($\text{sanit.access} = 0$).

Using the model in Equation 9.16, the prediction equation can be written

$$\widehat{\log(\text{inf.mortality})} = 4.018 - 1.681(\text{sanit.access}).$$

Exponentiating both sides of the equation yields

$$\widehat{\text{inf.mortality}} = e^{4.018 - 1.681(\text{sanit.access})}.$$

When $\text{sanit.access} = 0$, the equation simplifies to $e^{4.018} = 55.590$ deaths among 1,000 live births; this is the estimated infant mortality rate in the countries with low access to sanitation facilities. When $\text{sanit.access} = 1$, the estimated infant mortality rate is $e^{4.018 - 1.681(1)} = e^{2.337} = 10.350$ deaths per 1,000 live births. The infant mortality rate drops by a factor of 0.186; i.e., the mortality rate in the high access countries is approximately 20% of that in the low access countries.²¹

EXAMPLE 9.17

Check the assumptions of constant variability around the regression line and normality of the residuals in the model for the relationship between the transformed infant mortality variable and access to sanitation variable. Residual plots are shown in Figure 9.17.

E

While the normal probability plot does show that the residuals are approximately normally distributed, the residual plot reveals that variability is far from constant around the two predictors. Another method for assessing the relationship between the two groups is advisable; this is discussed further in Section 9.5.

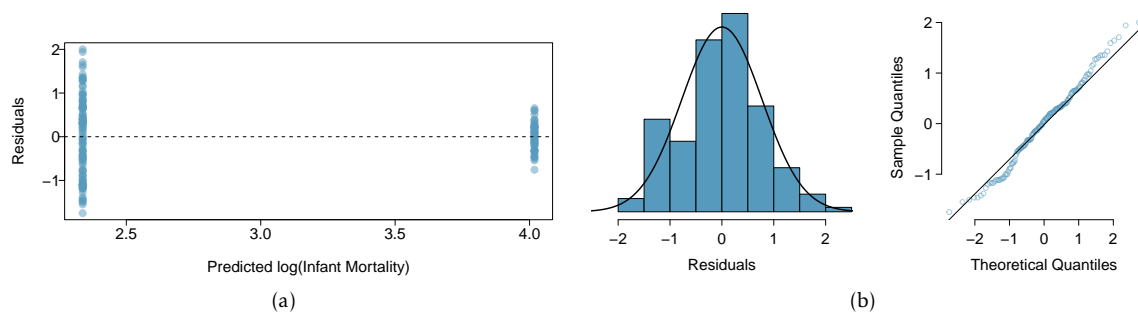


Figure 9.17: (a) Residual plot of $\log(\text{inf.mortality})$ and sanit.access . (b) Histogram and normal probability plot of the residuals.

9.4.4 Outliers in regression

Depending on their position, data points in a scatterplot have varying degrees of contribution to the estimated parameters of a regression line. Points that are at particularly low or high values of the predictor (x) variable are said to have **high leverage**, and have a large influence on the estimated intercept and slope of the regression line; observations with x values closer to the center of the distribution of x do not have a large effect on the slope.

A data point in a scatterplot is considered an **outlier in regression** if its value for the response (y) variable does not follow the general linear trend in the data. Outliers that sit at extreme values of the predictor variable (i.e., have high leverage) have the potential to contribute disproportion-

²¹When examining event rates in public health, associations are typically measured using rate ratios rather than rate differences.

ately to the estimated parameters of a regression line. If an observation does have a strong effect on the estimates of the line, such that estimates change substantially when the point is omitted, the observation is **influential**. These terms are formally defined in advanced regression courses.

This section examines the relationship between infant mortality and number of doctors, using data for each state and the District of Columbia.²² Infant mortality is measured as the number of infant deaths in the first year of life per 1,000 live births, and number of doctors is recorded as number of doctors per 100,000 members of the population. Figure 9.18 shows scatterplots with infant mortality on the y -axis and number of doctors on the x -axis.

One point in Figure 9.18(a), marked in red, is clearly distant from the main cluster of points. This point corresponds to the District of Columbia, where there were approximately 807.2 doctors per 100,000 members of the population, and the infant mortality rate was 11.3 per 1,000 live births. Since 807.2 is a high value for the predictor variable, this observation has high leverage. It is also an outlier; the other points exhibit a downward sloping trend as the number of doctors increases, but this point, with an unusually high y -value paired with a high x -value, does not follow the trend.

Figure 9.18(b) illustrates that the DC observation is influential. Not only does the observation simply change the numerical value of the slope parameter, it reverses the direction of the linear trend; the regression line fitted with the complete dataset has a positive slope, but the line re-fitted without the DC observation has a negative slope. The large number of doctors per population is due to the presence of several large medical centers in an area with a population that is much smaller than a typical state.

It seems natural to ask whether or not an influential point should be removed from a dataset, but that may not be the right question. Instead, it is usually more important to assess whether the influential point might be an error in the data, or whether it belongs in the dataset. In this case, the District of Columbia has certain characteristics that may make comparisons with other states inappropriate; this is one argument in favor of excluding the DC observation from the data.

Generally speaking, if an influential point arises from random sampling from a large population and is not a data error, it should be left in the dataset, since it probably represents a small subset of the population from which the data were sampled.

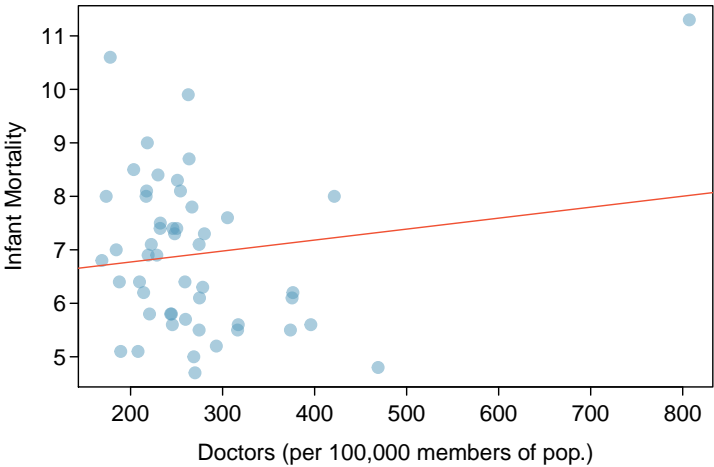
GUIDED PRACTICE 9.18



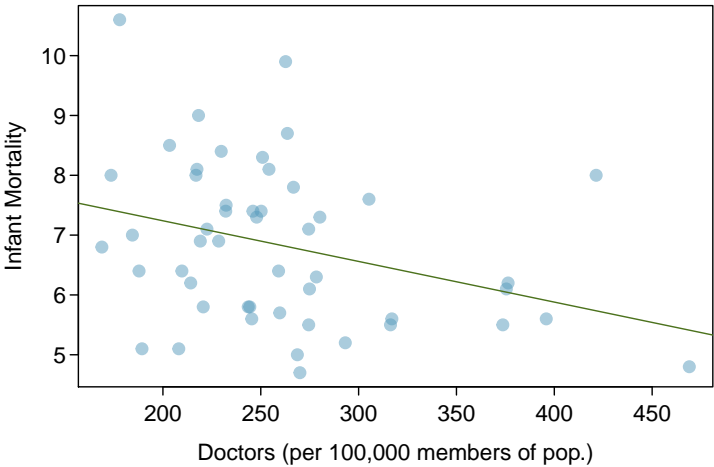
Once the influential DC point is removed, assess whether it is appropriate to use linear regression on these data by checking the four assumptions behind least squares regression: linearity, constant variability, independent observations, and approximate normality of the residuals. Refer to the residual plots shown in Figure 9.19.²³

²²Data are from the Statistical Abstract of the United States, published by the US Census Bureau. Data are for 2010, and available as census.2010 in the oibioestat package.

²³The scatterplot in Figure 9.18(b) does not show any nonlinear trends. Similarly, Figure 9.19(a) does not indicate any nonlinear trends or noticeable difference in the variability of the residuals, although it does show that there are relatively few observations for low values of predicted infant mortality. From Figure 9.19(b), the residuals are approximately normally distributed. Infant mortality across the states reflects a complex mix of different levels of income, access to health care, and individual state initiatives in health care; these and other state-specific features probably act independently across the states, although there is some dependence from federal influence such as funding for pre-natal care. Overall, independence seems like a reasonable assumption.



(a)



(b)

Figure 9.18: (a) Plot including District of Columbia data point. (b) Plot without influential District of Columbia data point.

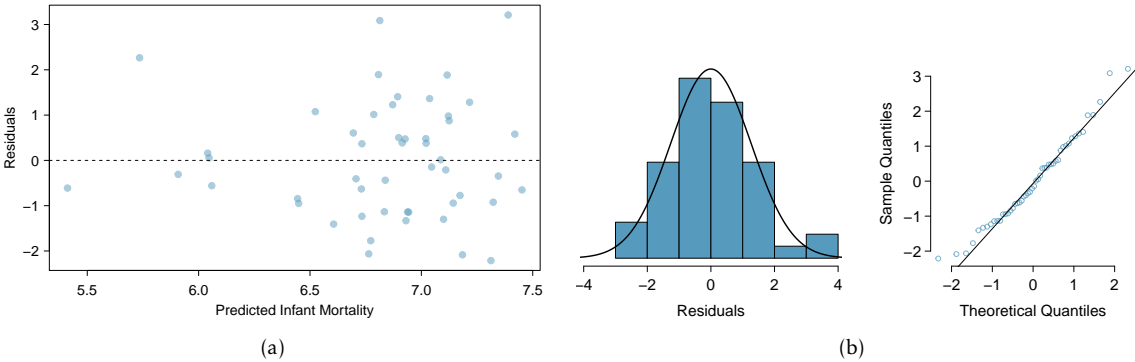


Figure 9.19: (a) Residual plot of inf.mortality and doctors. (b) Histogram and normal probability plot of the residuals.

9.5 Statistical inference with regression

The previous sections in this chapter have focused on linear regression as a tool for summarizing trends in data and making predictions. These numerical summaries are analogous to the methods discussed in Chapter 1 for displaying and summarizing data. Regression is also used to make inferences about a population.

The same ideas covered in Chapters 3 and 4 about using data from a sample to draw inferences about population parameters apply with regression. Previously, the goal was to draw inference about the population parameter μ ; in regression, the population parameter of interest is typically the slope parameter β_1 . Inference about the intercept term is rare, and limited to the few problems where the vertical intercept has scientific meaning.²⁴

Inference in regression relies on the population linear model for the relationship between an explanatory variable X and a response variable Y given by

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad (9.19)$$

where ε is assumed to have a normal distribution with mean 0 and standard deviation σ ($\varepsilon \sim N(0, \sigma)$). This population model specifies that a response Y has value $\beta_0 + \beta_1 X$ plus a random term that pushes Y symmetrically above or below the value specified by the line.²⁵

The set of ordered pairs (x_i, y_i) used when fitting a least squares regression line are assumed to have been sampled from a population in which the relationship between the explanatory and response variables follows Equation 9.19. Under this assumption, the slope and intercept values of the least squares regression line, b_0 and b_1 , are estimates of the population parameters β_0 and β_1 ; b_0 and b_1 have sampling distributions, just as \bar{X} does when thought of as an estimate of a population mean μ . A more advanced treatment of regression would demonstrate that the sampling distribution of b_1 is normal with mean $E(b_1) = \beta_1$ and standard deviation

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}}.$$

The sampling distribution of b_0 has mean $E(b_0) = \beta_0$ and standard deviation

$$\sigma_{b_0} = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}.$$

In both of these expressions, σ is the standard deviation of ε .

Hypothesis tests and confidence intervals for regression parameters have the same basic form as tests and intervals about population means. The test statistic for a null hypothesis $H_0 : \beta_1 = \beta_1^0$ about a slope parameter is

$$t = \frac{b_1 - \beta_1^0}{\text{s.e.}(b_1)},$$

where the formula for $\text{s.e.}(b_1)$ is given below. In this setting, t has a t -distribution with $n-2$ degrees of freedom, where n is the number of ordered pairs used to estimate the least squares line.

Typically, hypothesis testing in regression involves tests of whether the x and y variables are associated; in other words, whether the slope is significantly different from 0. In these settings, the null hypothesis is that there is no association between the explanatory and response variables, or

²⁴In some applications of regression, the predictor x is replaced by $x^* = x - \bar{x}$. In that case, the vertical intercept is the value of the line when $x^* = 0$, or $x = \bar{x}$.

²⁵Since $E(\varepsilon) = 0$, this model can also be written as $Y \sim N(\mu_x)$, with $\mu_x = E(Y) = \beta_0 + \beta_1 X$. The term ε is the population model for the observed residuals e_i in regression.

$H_0 : \beta_1 = 0 = \beta_1^0$, in which case

$$t = \frac{b_1}{\text{s.e.}(b_1)}.$$

The hypothesis is rejected in favor of the two-sided alternative $H_A : \beta_1 \neq 0$ with significance level α when $|t| \geq t_{df}^*$, where t_{df}^* is the point on a t -distribution with $n - 2$ degrees of freedom that has $\alpha/2$ area to its right (i.e., when $p \leq \alpha$).

A two-sided confidence interval for β_1 is given by

$$b_1 \pm \text{s.e.}(b_1) \times t_{df}^*.$$

Tests for one-sided alternatives and one-sided confidence intervals make the usual adjustments to the rejection rule and confidence interval, and p -values are interpreted just as in Chapters 4 and 5.

Formulas for calculating standard errors

Statistical software is typically used to obtain t -statistics and p -values for inference with regression, since using the formulas for calculating standard error can be cumbersome.

The standard errors of b_0 and b_1 used in confidence intervals and hypothesis tests replace σ with s , the standard deviation of the residuals from a fitted line. Formally,

$$s = \sqrt{\frac{\sum e_i^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}. \quad (9.20)$$

The term s^2 is often called the mean squared error from the regression, and s the root mean squared error.

The two standard errors are

$$\text{s.e.}(b_1) = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} \quad \text{and} \quad \text{s.e.}(b_0) = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}.$$

EXAMPLE 9.21

Is there evidence of a significant association between number of doctors per 100,000 members of the population in a state and infant mortality rate?

The numerical output that R returns is shown in Figure 9.20.²⁶

The question implies that the District of Columbia should not be included in the analysis. The assumptions for applying a least squares regression have been verified in Exercise 9.18. Whenever possible, formal inference should be preceded by a check of the assumptions for regression.

The null and alternative hypotheses are $H_0 : \beta_1 = 0$ and $H_A : \beta_1 \neq 0$.

The estimated slope of the least squares line is -0.0068, with standard error 0.0028. The t -statistic equals -2.40, and the probability that the absolute value of a t -statistic with $50 - 2 = 48$ degrees of freedom is smaller than -2.40 or larger than 2.40 is 0.021.

Since $p = 0.021 < 0.05$, the data support the alternative hypothesis that the number of physicians is associated with infant mortality at the 0.05 significance level. The sign of the slope implies that the association is negative; states with more doctors tend to have lower rates of infant mortality.

Care should be taken in interpreting the above results. The R^2 for the model is 0.107; the model explains only about 10% of the state-to-state variability in infant mortality, which suggests

²⁶Other software packages, such as Stata or Minitab, provide similar information but with slightly different labeling.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	8.5991	0.7603	11.31	0.0000
Doctors Per 100,000	-0.0068	0.0028	-2.40	0.0206

Figure 9.20: Summary of regression output from R for the model predicting infant mortality from number of doctors, using the census .2010 dataset.

there are several other factors affecting infant mortality that are not accounted for in the model.²⁷ Additionally, an important implicit assumption being made in this example is that data from the year 2010 are representative; in other words, that the relationship between number of physicians and infant mortality is constant over time, and that the data from 2010 can be used to make inference about other years.

Note that it would be incorrect to make claims of causality from these data, such as stating that an additional 100 physicians (per 100,000 residents) would lead to a decrease of 0.68 in the infant mortality rate.

GUIDED PRACTICE 9.22



Calculate a 95% two-sided confidence interval for the slope parameter β_1 in the state-level infant mortality data.²⁸

²⁷Calculations of the R^2 value are not shown here.

²⁸The t^* value for a t -distribution with 48 degrees of freedom is 2.01, and the standard error of b_1 is 0.0028. The 95% confidence interval is $-0.0068 \pm 2.01(0.0028) = (-0.0124, -0.0012)$.

Connection to two-group hypothesis testing

Conducting a regression analysis with a numerical response variable and a categorical predictor with two levels is analogous to conducting a two-group hypothesis test.

For example, Section 9.4.3 shows a regression model that compares the average infant mortality rate in countries with low access to sanitation facilities versus high access.²⁹ In other words, the purpose of the analysis is to compare mean infant mortality rate between the two groups: countries with low access versus countries with high access. Recall that the slope parameter b_1 is the difference between the means of $\log(\text{mortality rate})$. A test of the null hypothesis $H_0 : \beta_1 = 0$ in the context of a categorical predictor with two levels is a test of whether the two means are different, just as for the two-group null hypothesis, $H_0 : \mu_1 = \mu_2$.

When the pooled standard deviation assumption (Section 6.4.1) is used, the t -statistic and p -value from a two-group hypothesis test are equivalent to that returned from a regression model.

Figure 9.21 shows the R output from a regression model in the `wdi.2011` data, in which `sanit.access = 1` for countries where at least 50% of the population has access to adequate sanitation and 0 otherwise. The abbreviated R output from two-group t -tests are shown in Figure 9.22. The version of the t -test that does not assume equal standard deviations and uses non-integer degrees of freedom is often referred to as the Welch test.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.0184	0.1100	36.52	< 0.001
High Access	-1.6806	0.1322	-12.72	0.001

Figure 9.21: Regression of $\log(\text{infant mortality})$ versus sanitation access.

Test	df	t value	Pr(> t)
Two-group t -test	161	12.72	< 0.001
Welch two-group t -test	155.82	17.36	< 0.001

Figure 9.22: Results from the independent two-group t -test, under differing assumptions about standard deviations between groups, for mean $\log(\text{infant mortality})$ between sanitation access groups.

The sign of the t -statistic differs because for the two-group test, the difference in mean $\log(\text{infant mortality})$ was calculated by subtracting the mean in the high access group from the mean in the low access group; in the regression model, the negative sign reflects the reduction in mean $\log(\text{infant mortality})$ when changing from low access to high access. Since the t -distribution is symmetric, the two-sided p -value is equal. In this case, p is a small number less than 0.001, as calculated from a t -distribution with $163 - 2 = 161$ degrees of freedom (recall that 163 countries are represented in the dataset). The degrees of freedom for the pooled two-group test and linear regression are equivalent.

Example 9.17 showed that the constant variability assumption does not hold for these data. As a result, it might be advisable for a researcher interested in comparing the infant mortality rates between these two groups to conduct a two-group hypothesis test without using the pooled standard deviation assumption. Since this test uses a different formula for calculating the standard error of the difference in means, the t -statistic is different; additionally, the degrees of freedom are not equivalent. In this particular example, there is not a noticeable effect on the p -value.

²⁹Recall that a log transformation was used on the infant mortality rate.

9.6 Interval estimates with regression

Section 9.5 introduced interval estimates for regression parameters, such as the population slope β_1 . An estimated regression line can also be used to construct interval estimates for the regression line itself and to calculate prediction intervals for a new observation.

9.6.1 Confidence intervals

As initially discussed in Section 9.3, the estimated regression line for the association between RFFT score and age from the 500 individuals in `prevend.samp` is

$$\widehat{\text{RFFT}} = 137.55 - 1.26(\text{age}).$$

Figure 9.23 shows the summary output from R when the regression model is fit. R also provides the value of R^2 as 0.285 and the value of s , the estimated standard deviation of the residuals, as 23.2.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	137.55	5.02	27.42	0.000
Age	-1.26	0.09	-14.09	0.000
$df = 498$				

Figure 9.23: Summary of regression output from R for the model predicting RFFT score from age, using the `prevend.samp` dataset.

A confidence interval for the slope parameter β_1 is centered at the point estimate b_1 , with a width based on the standard error for the slope. For this model, the 95% confidence interval for age is $-1.26 \pm (1.96)(0.09) = (-1.44, -1.09)$ years.³⁰ With 95% confidence, each additional year of age is associated with between a 1.1 and 1.4 point lower RFFT score.

A confidence interval can also be calculated for a specific point on a least squares line. Consider a specific value of the predictor variable, x^* , such as 60 years of age. At age 60 years, the predicted value of RFFT score is $137.55 - 1.26(60) = 61.95$ points. The fitted line suggests that individuals from this population who are 60 years of age score, on average, about 62 points on the RFFT. Each point on the estimated regression line represents the predicted average RFFT score for a certain age.

More generally, the population model for a regression line is $E(Y|x) = \beta_0 + \beta_1 x$, and at a value x^* of the predictor x , the fitted regression line

$$\widehat{E(Y|x^*)} = b_0 + b_1 x^*$$

estimates the mean of Y for members of the population with predictor value x^* .

Thus, each point on a fitted regression line represents a point estimate for $E(Y|x^*)$. The corresponding interval estimate for $E(Y|x^*)$ measures the uncertainty in the estimated mean of Y at predictor value x^* , just as how an interval estimate for the population slope β_1 represents the uncertainty around b_1 .

The confidence interval for $E(Y|x^*)$ is computed using the standard error of the estimated

³⁰The critical value 1.96 is used here because at degrees of freedom 498, the t -distribution is very close to a normal distribution. From software, $t_{0.975, df=498}^* = 1.9647$.

mean of the regression model at a value of the predictor:

$$\text{s.e.}(\widehat{E(Y|x^*)}) = \sqrt{s^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)} = s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}.$$

In this expression, s is given by Equation 9.20, the usual estimate of σ , the standard deviation of the error term ϵ in the population linear model $Y = \beta_0 + \beta_1 X + \epsilon$.

The standard error of an estimated mean in regression is rarely calculated by hand; with all but the smallest datasets, the calculations are long and best left to software. When necessary, it can be calculated from basic features of the data and summary statistics.

Consider computing a 95% confidence interval for $E(\text{RFFT}|\text{age} = 60)$.

- The sample size is $n = 500$.
- $s = 23.2$ appears in the regression output.
- The sample mean \bar{x} of the predictor is $\overline{\text{age}} = 54.8$ years.
- $(x^* - \bar{x})^2$ is the squared distance between the predictor value of interest and the sample mean of the predictors: $(60 - 54.8)^2 = 27.04$.
- The sum $\sum (x_i - \bar{x})^2$ is the numerator in the calculation of the variance of the predictor, and equals $(n-1)\text{Var}(x) = (499)(134.4445) = 67,088$.

Using these values, the standard error of the estimated mean RFFT score at age 60 is

$$\text{s.e.}(\widehat{E(\text{RFFT}|\text{age} = 60)}) = 23.2 \sqrt{\frac{1}{500} + \frac{27.04}{67,088}} = 1.14.$$

Thus, a 95% confidence interval for the estimated mean is $61.95 \pm (1.96)(1.14) = (59.72, 64.18)$ points. With 95% confidence, the interval (59.72, 64.18) points contains the average RFFT score of a 60-year-old individual.

It is also possible to calculate approximate confidence intervals for the estimated mean at a specific value of a predictor. When $x^* = \bar{x}$, the second term in the square root will be 0, and the standard error of the estimated mean at the average value \bar{x} will have the simple form s/\sqrt{n} . For values close to \bar{x} , approximating the standard error as s/\sqrt{n} is often sufficient. In the PREVENT data, 60 years is reasonably close to the average age 54.8 years, and the approximate value of the standard error is $23.2/\sqrt{500} = 1.03$. For values x^* that are more distant from the mean, the second term in the square root cannot be reasonably ignored.

The approximate form of the standard error for the mean at a predictor value, s/\sqrt{n} , makes it easier to see that for large n , the standard error approaches 0; thus, the confidence interval narrows as sample size increases, allowing the estimates to become more precise. This behavior is identical to the confidence interval for a simple mean, as one would expect. It is possible to show algebraically that the confidence intervals at any value of the predictor become increasingly narrow as the sample size increases.

9.7 Notes

This chapter provides only an introduction to simple linear regression.

When fitting a simple regression, be sure to visually assess whether the model is appropriate. Nonlinear trends or outliers are often obvious in a scatterplot with the least squares line plotted. If outliers are evident, the data source should be consulted when possible, since outliers may be indicative of errors in data collection. It is also important to consider whether observed outliers belong to the target population of inference, and assess whether the outliers should be included in the analysis.

There are several variants of residual plots used for model diagnostics. The ones shown in Section 9.4.1, which plot the predicted values on the horizontal axis, easily generalize to settings with multiple predictors, since there is always a single predicted value even when there is more than one predictor. If the only model used is a simple regression, plotting residuals against predictor values may make it easier to identify a case with a notable residual. Additionally, data analysts will sometimes plot residuals against case number of the predictor, since runs of large or small residuals may indicate that adjacent cases are correlated.

The R^2 statistic is widely used in the social sciences, where the unexplained variability in the data is typically much larger than the variability captured or explained by a model. It is important to be aware of what information R^2 does and does not provide. Even though a model may have a low proportion of explained variability, regression coefficients in the model can still be highly statistically significant. The R^2 should not be interpreted as a measure of the quality of the fit of the model. It is possible for R^2 to be large even when the data do not show a linear relationship.

Linear regression models are often estimated after an investigator has noticed a linear relationship in data, and experienced investigators can often guess correctly that regression coefficients will be significant before calculating a p -value. Unlike with two-sample hypothesis tests, regression models are rarely specified in advance at the design stage. In practice, it is best to be skeptical about a small p -value in a regression setting, and wait to see whether the observed statistically significant relationship can be confirmed in an independent dataset. The issue of model validation and assessing whether results of a regression analysis will generalize to other datasets is often discussed at length in advanced courses.

In more advanced texts, substantial attention is devoted to the subtleties of fitting straight line models. For instance, there are strategies for adjusting an analysis when one or more of the assumptions for regression do not hold. There are also specific methods to numerically assess the leverage or influence that each observation has on a fitted model.

Appendix A

Solutions

1 Introduction to data

1.1 (a) Ordinal (b) Ordinal or Continuous (c) Nominal (d) Continuous (e) Discrete (f) Nominal (g) Nominal (h) Discrete (i) Nominal (j) Nominal (k) Continuous (l) Continuous (m) Continuous (n) Nominal (o) Continuous (p) Nominal (q) Discrete (r) Nominal (s) Nominal

1.2 (a) 10-percentile=41.10 ; 35-percentile=44 ; 62-percentiles= 51.22 (b) $Q_1 = 43.00$; $Q_2 = 48.00$; $Q_3 = 53.25$ (c) $IQR = 10.25$. Lower limit= 27.625; Upper limit=68.625. There is not outlying value. (d) Fig.A.1

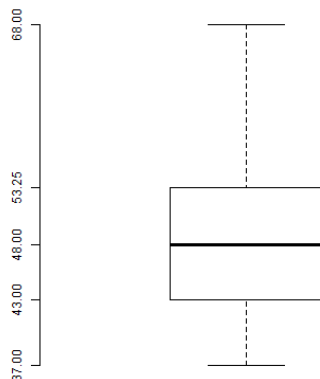


Figure A.1: Boxplot Exercise 1.2

1.3 (a) Quantitative (discrete)

(b)

0	1	2	3	4	5	6
7	4	6	7	4	4	8

(c) The value 6.

1.4

(a) The study variable is the moment of the day when emergency occurs. Qualitative nominal. (b) No, because it is a qualitative variable.

(c)

Afternoon	Morning	Night	Nonworking
9	11	7	5
0.2813	0.3438	0.2188	0.1563

(d) Fig. A.2

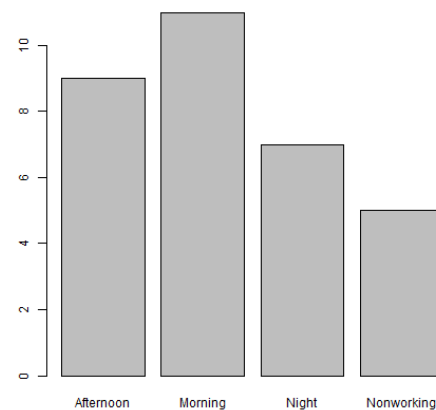


Figure A.2: Bar plot exercise

1.5 (a) Copper levels in urine. Quantitative continuous (b) Median=0.71; Range=1.24-0.1=1.14 (c) $Q_1 = 0.5425$; $Q_3 = 0.8350$ (d) 10-percentile=0.414 ; 95-percentile= 1.122 (e) $IQR = 0.2925$; Lower limit= 0.10375; Upper Limit= 1.27375; The value 0.1 is an outlying value (f) Fig. A.3 and A.4

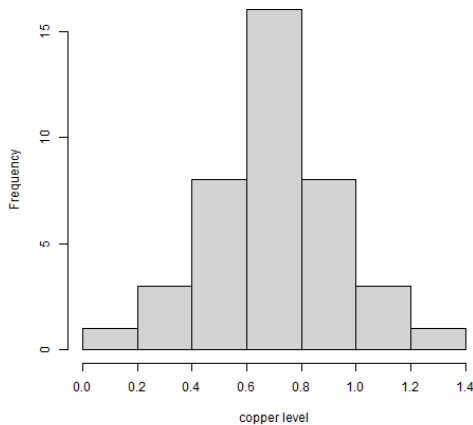


Figure A.3: Histogram Exercise 5

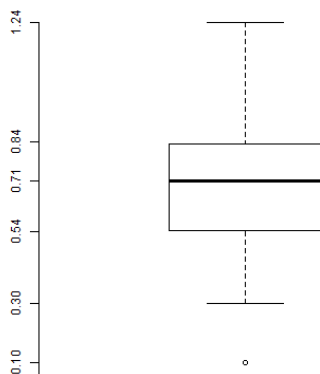


Figure A.4: Boxplot

1.6 (a) Qualitative continuous (b) minimum and maximum; 3281 and 4422
 10 and 90 percentiles: 3682.5 and 4288.0 ;
 quartiles, median: $Q_1 = 3955.75$; $Q_2 = 4125.00$ $Q_3 = 4173.00$
 mean = 4033.438
 mode (It is not appropriate for a qualitative continuous variable)
 range = 1141
 variance = 82981.2,
 standard deviation = 288.0646

2 Probability and Distributions of Random Variables

2.1 $0.34 + 0.54 - 0.25 = 0.63$

2.2 $(568 - 28) / 568 = 540 / 568 = 0.9507$

2.3 $P(B \cap D) = 0.15 \cdot 0.24 = 0.036$ and

$P(B \cup D) = P(B) + P(D) - P(B \cap D) = 0.15 + 0.24 - 0.036 = 0.354$

2.4 Yes, because $P(A \cap B) = P(A) \cdot P(B)$

2.5 i) 0.4332; ii) 0.3445; iii) 0.9875; iv) 0.4920; v) 0.3174; vi) 0.2728; vii) 0.0456; viii) 0.5886

2.6 i) 0.35; ii) 1.54; iii) 2.16 iv) -1.24

2.7 i) 0.9970; ii) 0.0668 ; iii) 0.0873

2.8 31.92%

2.9 95.44%

2.10 The 80-percentile is 73.4. So, the answer is 74

2.11 The length 6 feet is equivalent to 72 inches.

$P(X < 72) = P(Z < 4/3) = 0.9082$

2.12 i) 1.70%; ii) 85.55

2.13 i) (3.08, 10.92); ii) (2.34, 11.66); iii) (3.72, $+\infty$);
 iv) 4%; v) 89.44%; vi) 38.30%;
 vii) 9.19% viii) 96 percentile

2.14 i) (126.48, 173.52); ii) (130.32, 169.68); iii) (0, 169.68)
 iv) 74.86%; v) 77.34%; vi) 74.92% ; vii) 98.12 percentile

3 Chapter 3

3.1 (a) $\bar{x} = 0.6052$.

(b) $s = 0.0131$.

(c) $Z_{0.63} = \frac{0.63 - 0.6052}{0.0131} = 1.893$. No, this level of BGC is within 2 SD of the mean.

(d) The standard error of the sample mean is given by $\frac{s}{\sqrt{n}} = \frac{0.0131}{\sqrt{70}} = 0.00157$.

4 Chapter 4

4.1 $80.5 \pm 14.2 \cdot 1.96 / \sqrt{24}$. The 95- CI is (74.8188, 86.1812)

4.2 $85 = 110(1 - \alpha)$; $\alpha = 0.15$; $1 - \alpha/2 = 0.925$ and $z_{0.925} = 1.44$.

The error $z_{1-\alpha/2} \sigma / \sqrt{n}$ is less than 2.75. Hence

$$n > z_{1-\alpha/2}^2 \frac{\sigma^2}{2.75^2} = 1.44^2 \frac{86.4}{2.75^2} = 23.69$$

The sample size should be larger than 24.

4.3 $90 \pm 1.96 \cdot 10 / \sqrt{49}$. The 95-CI is (87.2, 92.8)

4.4 $\bar{x} = 80.1333$ and $s = 5.4099$. $t_{14, 0.975} = 2.145$.
 $80.13 \pm 2.145 \cdot 5.4099 / \sqrt{15}$ The 95-CI is (77.1371, 83.1285).

4.5 $n = 100 > 30$; $\bar{x} = 23$; $s = 10$; $100(1 - \alpha) = 90$;
 $\alpha = 0.1$; $1 - \alpha/2 = 0.95$; $z_{0.95} = 1.64$
 $23 \pm 1.64 \cdot 10 / \sqrt{100}$. The 90%-CI is (21.36, 24.64)

4.6 Assuming that the weight of ten year-old male gorilla is a normal variable.

$$\alpha = 0.1 \text{ and } t_{24,.95} = 1.711$$

$36.5 \pm 1.711 \cdot 5/\sqrt{25}$. The 90-CI is (34.789, 38.211).

4.7 $\bar{x} = 12$; $n = 20$; $s = 7.7392$; $\alpha = 0.05$; $t_{19,.975} = 2.093$

$12 \pm 2.093 \cdot 7.7392/\sqrt{20}$. The 95%-CI is (8.3780, 15.6220)

4.8 i) $140 \pm 1.96 \cdot 25/\sqrt{200}$. The 95%-CI is (136.5352, 143.4648).

ii) There is an association between glaucoma and blood pressure, since the average blood pressure for glaucoma patients is higher than the general population.

4.9 $\bar{x} = 1.435$; $s = 0.1461$, $t_{19,.975} = 2.093$

$1.435 \pm 2.093 \cdot 0.1461/\sqrt{20}$. The 95%-CI is (1.3666, 1.5034)

4.10 The question is about the confidence $100(1-\alpha)$.

The error $z_{1-\alpha/2}\sigma/\sqrt{n}$ is equal to 10 is

$$z_{1-\alpha/2} = \frac{10}{\sigma} \sqrt{n} = 10 \frac{\sqrt{24}}{\sqrt{424}} = 2.3792$$

$P(Z < 2.38) = 0.9913 = 1 - \frac{\alpha}{2}$; $\alpha = 0.0174$;
 $100(1-\alpha) = 98.26$

4.11 $\hat{p} = 6/46$; $\frac{6}{46} \pm 1.96 \frac{\sqrt{(6/46) \cdot (40/46)}}{\sqrt{46}}$. The 95-CI is (0.0331, 0.2278)

4.12 $\hat{p} = 311/375$; $\frac{311}{375} \pm 1.64 \frac{\sqrt{(311/375) \cdot (64/375)}}{\sqrt{375}}$. The 90-CI is (0.7975, 0.8612).

5 Chapter 5

5.1 $t = \frac{9.77 - 7.65}{\sqrt{11.14}} \sqrt{15} = 2.4600$. Acceptance interval: (-2.145, 2.145) Null Hypothesis is rejected.

5.2 $t = \frac{155.223 - 196.49}{88.04} \sqrt{27} = -2.4352$. Acceptance Interval: (-2.056, 2.056) Null hypothesis is rejected.

5.3 (a) (24.29, 25.71) (b) The test statistic is $t = 2.8207$. The p-value is $p = 0.006583164$. (This value is calculated with R-Commander) (c) Null-Hypothesis is rejected. There is a statistically significant difference between the mean of body mass indices of the diabetic and non diabetic population. (d) We know that the null hypothesis was going to be rejected because the value of the sample mean does not belong to the 95% Confidence Interval.

5.4 If we assume that blood pressure is normally distributed:

1. The test statistic is $t = \frac{143 - 136}{24.4} \sqrt{86} = 2.6605$
The acceptance region: $(-t_{85,.95}, t_{85,.95}) = (-1.663, 1.663)$. Null-hypothesis is rejected.

2. The test statistic is $t = \frac{87 - 84}{16.0} \sqrt{86} = 1.738803$
The acceptance region is the same than before. Now Null-hypothesis is rejected.

3. There are enough statistical evidence in order to state that workers who have experienced a major coronary event have higher systolic and diastolic blood pressure.

If normal distribution is not assumed, the statistic calculated is approximately normal distributed because the sample is large enough (> 30). Therefore, for the acceptance region normal distribution and $(-1.645, 1.645)$ and the conclusions are the same.

5.5 Let p be the proportion of solo practice for veterinary physician

$$H_0 : p = 0.43$$

$$H_A : p \neq 0.43$$

The data are sample proportion $\hat{p} = 0.32$ and sample size $n = 223$.

The test statistic is

$$z = \frac{0.32 - 0.43}{\sqrt{0.43(1 - 0.43)/223}} = -3.3180$$

The critical values for $\alpha = 0.05$ are $z_{1-\alpha/2} = z_{0.975} = 1.96$ and $z_{\alpha/2} = -1.96$. The acceptance region is $(-1.96, 1.96)$.

The test statistic value does not belong to the acceptance region, the null-hypothesis is rejected.

The p-value is 0.2726051

5.6 $n = 11$, sample mean $\bar{x} = 69.3636$, sample variance $s^2 = 39.6546$, sample standard deviation $s = 6.2972$. Let μ be the mean value of the strength of the lateral digital extensor muscle.

$$H_0 : \mu = 65$$

$$H_A : \mu \neq 65$$

The acceptance interval is $(-t_{10,.975}, t_{10,.975}) = (-2.228, 2.228)$ and the test statistic is

$$t = \frac{69.3636 - 65}{6.2972/\sqrt{11}} = 2.2982$$

The null hypothesis is rejected. There is enough evidence to state that the mean mean value of the strength of the lateral digital extensor muscle is different from 65.

5.7 $p_0 = 0.25$, $n = 125$, sample proportion $\hat{p} = 35/125 = 0.28$.

$$H_0 : p = 0.25$$

$$H_A : p \neq 0.25$$

¹ The number of positives outcomes is not greater than 10, so the interval calculated is not a very good approximation.

The acceptance interval is $(-1.96, 1.96)$ and the test statistic is

$$z = \frac{(0.28 - 0.25)}{\sqrt{0.25 \cdot 0.75 / \sqrt{125}}} = 0.7746$$

Null hypothesis is accepted. We don't have enough evidence to state that the proportion is different from 25%.

p-value is 0.44.

5.8 Acceptance interval $(-t_{14,.975}, t_{14,.975}) = (-2.145, 2.145)$ and test statistic is

$$t = \frac{162.5 - 160}{5/\sqrt{15}} = 1.9365$$

Null hypothesis is accepted.

5.9 Sample size: $n = 1500$, sample proportion $\hat{p} = 125/1500 = 0.0833$

Two-sided:

$$H_0 : p = 0.06 \quad H_A : p \neq 0.06$$

The acceptance interval is $(-1.96, 1.96)$ and the test statistic is

$$z = \frac{0.0833 - 0.06}{\sqrt{0.06 \cdot 0.94 / 1500}} = 3.7998$$

Null hypothesis is rejected.

p-value = $2(1 - Pr(Z < 3.80)) = 2(1 - 0.9999) = 0.0002$

One-sided:

$$H_0 : p = 0.06 \quad H_A : p > 0.06$$

The acceptance interval is $(-\infty, 1.64)$ and the test statistic is the same value

$$z = 3.7998$$

Null hypothesis is rejected. p-value = $1 - Pr(Z < 3.80) = 1 - 0.9999 = 0.0001$

5.10 Sample size: $n = 150 > 30$. With $\alpha = 0.05$ the acceptance interval is $(-1.96, 1.96)$ and $z = \frac{325-332}{52/\sqrt{150}} = -1.6487$

6 Chapter 6

6.1 Sample proportions: $\hat{p}_1 = 0.2333$ $\hat{p}_2 = 0.32$. Pool proportion: $\hat{p} = 0.2727$.

Test statistic: $z = 2.7831$. Acceptance interval: $(-1.96, 1.96)$

Null hypothesis is rejected.

6.2 Sample proportions: $\hat{p}_1 = 0.58$ $\hat{p}_2 = 0.5294$. Pool proportion: $\hat{p} = 0.5531$.

Test statistic: $z = -0.9083$. p-value = $2 \cdot (1 - Pr(Z < 0.91)) = 2 \cdot (1 - 0.8186) = 0.3628$

Null hypothesis is accepted.

6.3 Sample proportions: $\hat{p}_1 = 0.0704$ $\hat{p}_2 = 0.0562$.

Pool proportion: $\hat{p} = 0.0619$.

Test statistic: $z = -1.6688$. p-value $2 \cdot (1 - Pr(Z < 1.67)) = 2 \cdot (1 - 0.9525) = 0.095$.

Null hypothesis is accepted

6.4 The same calculation than problem 3.

6.5 i) $F = 1.0808$ Acceptance interval $(F_{11,8,.025}, F_{11,8,.975})$. We don't have the value $df_1 = 11$. So, we use $df_1 = 10$ (or $df_1 = 12$). $(F_{10,8,.025}, F_{10,8,.975}) = (1/3.85, 4.30) = (0.26, 4.30)$. Null hypothesis is accepted.

ii) $S_p^2 = (11 \cdot 1.05^2 + 8 \cdot 1.01^2) / 19 = 1.0678$. Test statistic $t = 4.8500$. Acceptance interval $(-t_{19,.975}, t_{19,.975}) = (-2.093, 2.093)$. Null hypothesis is rejected.

6.6 $\bar{x}_1 - \bar{x}_2 = 3.4125$ y $s_{x_1-x_2} = 3.6443$. Test statistic $t = 2.6485$. Acceptance interval: $(-t_{7,.975}, t_{7,.975}) = (-2.365, 2.365)$.

Null hypothesis is rejected.

6.7 i) Test statistic: $F = 1.1475$. Acceptance interval: $(0.36, 2.57)$. Assume equal variances.

ii) $S_p^2 = 12.9714$. Test statistic: $t = 2.0917$. Acceptance interval: $(-2.030, 2.030)$. Reject null hypothesis

iii) p-value is 0.0437 using R-Commander.

6.8 Test equal Variances. Test statistic: $F = 35.5^2 / 37.6^2 = 0.8914$. Acceptance interval $(F_{24,46,.025}, F_{24,46,.975})$. We use degree of freedom 25 and 50, $(1/2.08, 1.92) = (0.48, 1.92)$. Hence, null hypothesis is accepted.

Test means: $S_p^2 = 1312.882$. Test statistic: -0.02869 . Acceptance interval $(-t_{70,.975}, t_{70,.975}) = (-1.994, 1.994)$. Null hypothesis is accepted.

7 Chapter 7

7.1 1. There is not outlying values, so we could assume the distribution to be normal. Variability is similar across groups. 2. All the p-values are greater to 0.05, conditions to apply ANOVA could be assume according to these tests. 3. (a) Snedecor's F distribution. (b) $[0, 3.09)$ (c) No, ANOVA is not used to identify which pairs are different. 4. There are significant difference for any pair of comparisons since 0 is not included-

7.2 (a) Yes, (b) $H_0 : \mu_A = \mu_B = \mu_C$ and $H_1 : \mu_A \neq \mu_B$ or $\mu_A \neq \mu_C$ or $\mu_B \neq \mu_C$ 4. We have evidence that A is different from B and C.

7.3 With $\alpha = 0.05$ we do not have evidence of difference across sections.

7.4 (a) H_0 : Average GPA is the same for all majors. H_A : At least one pair of means are different. (b) Since p-value > 0.05 , fail to reject H_0 . The data do not pro-

vide convincing evidence of a difference between the average GPAs across three groups of majors. (c) The total degrees of freedom is $195 + 2 = 197$, so the sample size is $197 + 1 = 198$.

8 Chapter 8

8.1 $\chi^2 = 61.356$; Rejection interval: $(5.99, +\infty)$

8.2 $\chi^2 = 0.76536$; Rejection interval: $(7.81, +\infty)$

8.3 $\chi^2 = 0.82589$; ; Rejection interval: $(5.99, +\infty)$

8.4 $\chi^2 = 50.323$; Rejection interval: $(9.21, +\infty)$

8.5 $\chi^2 = 4.71$; Rejection interval: $(3.841, +\infty)$

Appendix B

Distribution tables

P(Z ≤ z) where Z ~ N(0, 1)										
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994</					

B.2 t Distribution Table

$df \backslash p$	0.650	0.700	0.750	0.800	0.850	0.900	0.950	0.975	0.990	0.995
1	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841
4	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604
5	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032
6	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707
7	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499
8	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355
9	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250
10	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169
11	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106
12	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055
13	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012
14	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977
15	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947
16	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921
17	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898
18	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878
19	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861
20	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845
21	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831
22	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819
23	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807
24	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797
25	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787
26	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779
27	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771
28	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763
29	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756
30	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750
31	0.389	0.530	0.682	0.853	1.054	1.309	1.696	2.040	2.453	2.744
32	0.389	0.530	0.682	0.853	1.054	1.309	1.694	2.037	2.449	2.738
33	0.389	0.530	0.682	0.853	1.053	1.308	1.692	2.035	2.445	2.733
34	0.389	0.529	0.682	0.852	1.052	1.307	1.691	2.032	2.441	2.728
35	0.388	0.529	0.682	0.852	1.052	1.306	1.690	2.030	2.438	2.724
40	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704
50	0.388	0.528	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678
60	0.387	0.527	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660
70	0.387	0.527	0.678	0.847	1.044	1.294	1.667	1.994	2.381	2.648
80	0.387	0.526	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639
90	0.387	0.526	0.677	0.846	1.042	1.291	1.662	1.987	2.368	2.632
100	0.386	0.526	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626
110	0.386	0.526	0.677	0.845	1.041	1.289	1.659	1.982	2.361	2.621
120	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617
∞	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576

B.3 Chi-Square Probability Table

$df \backslash p$	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.458	15.655	17.539	19.281	21.434	41.422	44.985	48.232	52.191	55.003
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.815	17.074	19.047	20.867	23.110	43.745	47.400	50.725	54.776	57.648
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.192	18.509	20.569	22.465	24.797	46.059	49.802	53.203	57.342	60.275
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169
110	75.550	78.458	82.867	86.792	91.471	129.385	135.480	140.917	147.414	151.948
120	83.852	86.923	91.573	95.705	100.624	140.233	146.567	152.211	158.950	163.648

B.4 Snedecor's F distribution

$n_2 \backslash n_1$		95-Percentile of Snedecor's F-distribution																	
		1	2	3	4	5	6	7	8	9	10	12	15	20	25	50	100		
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	243.95	248.01	249.26	251.77	253.04	100		
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.46	19.48	19.49	100		
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.63	8.58	8.55	100		
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.70	5.66	100		
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.52	4.44	4.41	100		
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.83	3.75	3.71	100		
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.40	3.32	3.27	100		
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.11	3.02	2.97	100		
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.89	2.80	2.76	100		
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.73	2.64	2.59	100		
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.50	2.40	2.35	100		
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.28	2.18	2.12	100		
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.07	1.97	1.91	100		
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.84	1.78	100		
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.87	1.78	1.73	1.60	1.52	100		
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.85	1.77	1.68	1.62	1.48	1.39	100		

$n_2 \backslash n_1$		97.5-Percentile of Snedecor's F-distribution																	
		1	2	3	4	5	6	7	8	9	10	12	15	20	25	50	100		
1	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.63	976.71	984.87	993.10	998.08	1008.12	1013.17	100		
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.48	39.49	100		
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.01	13.96	100		
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.50	8.38	8.32	100		
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.27	6.14	6.08	100		
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.11	4.98	4.92	100		
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.40	4.28	4.21	100		
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.94	3.81	3.74	100		
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.60	3.47	3.40	100		
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.35	3.22	3.15	100		
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.01	2.87	2.80	100		
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.69	2.55	2.47	100		
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.40	2.25	2.17	100		
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.23	2.08	2.00	100		
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.22	2.11	1.99	1.92	1.75	1.66	100		
100	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	2.08	1.97	1.85	1.77	1.59	1.48	100		

Index

- A^c , 37
- alternative hypothesis (H_A), 106
- analysis of variance (ANOVA), 154, 154–164
- association, 188
- bar plot, 15
- blocking, 72
- blocks, 72
- Bonferroni correction, 163
- boxplot, 21
- case, 10
- categorical variable, 12
 - levels, 12
 - nominal, 12
 - ordinal, 12
- Central Limit Theorem, 78
- chi-square distribution, 175
- chi-square statistic, 174
- chi-square table, 175
- cohort, 72
- collections, 34
- column totals, 172
- complement, 37
- confidence interval, 84, 85, 91
 - confidence level, 87–88
 - difference of two means, 143–145
 - difference of two proportions, 129
 - interpretation, 88–91
 - regression coefficient, 208
 - single proportion, 98
- confident, 85
- confounder, 75
- confounding factor, 75
- confounding variable, 75
- contingency table
 - column totals, 172
 - row totals, 172
- continuous random variable, 45
- control, 72
- correlation, 190
- correlation coefficient, 190
- data, 7
 - births, 145–147
 - breast cancer, 131–135
 - cancer in dogs, herbicide, 132
 - cdc, 63
 - Congress approval rating, 100
 - dolphins and mercury, 95–96
 - famuss, 11, 15–24, 177–178
 - FCID, 46
 - frog, 10–11, 17–24, 188
 - Great Danes, 106–107, 114–115
 - health care, 130
 - hiv, 173–174, 178
 - LEAP, 8–9, 72–73, 179
 - life expectancy, 189–190
 - mammography, 131–135
 - nhanes, 117, 188–189, 197
 - PREVEND, 193–195
 - stem cells, heart function, 143–145
 - swim suit velocities, 136
 - white fish and mercury, 96
- data density, 20
- data fishing, 153, 156
- data matrix, 11
- data snooping, 153
- deck of cards, 35
- descriptive statistics, 7
- deviation, 18
- discrete random variable, 45
- disjoint events, 33
- distribution, 17
 - t , 92–94

- normal, 51, 51
- dot plot, 20
- estimate, 63
- event, 29, 34, 34–35
- expectation, 46–47
- expected counts, 172
- expected value, 46
- experiment, 72
- explanatory variable, 14, 186
- F-statistic, 157
- factor variables, 12
- frequency, 15
- frequency table, 15
- General Addition Rule, 36
- Greek
 - mu (μ), 47
 - sigma (σ), 48
- high leverage, 204
- histogram, 20, 24
- hypothesis testing, 106–120
 - decision errors, 119
 - significance level, 120
 - single proportion, 116
- independent, 38, 188
- independent samples, 141
- inferential statistics, 7
- influential, 205
- interquartile range, 19
- Kolmogorov–Smirnov, 155
- Law of Large Numbers, 32
- least squares regression, 195–197
 - R-squared (R^2), 201
- least squares regression line, 195, 196
- levels, 11
- Levene’s test, 155
- linear association, 188
- linear model, 195
- lurking variable, 75
- margin of error, 85, 100
- mean, 17
 - average, 17
- mean square between groups (*MSG*), 156
- mean square error (*MSE*), 157
- median, 17
- modality
 - bimodal, 21
 - multimodal, 21
 - unimodal, 21
- mode, 15, 21
- moderate relationships, 186
- Multiplication Rule, 40
- mutually exclusive events, 33
- negatively associated, 188
- non-response, 68
- non-response bias, 68
- normal probability table, 54
- null hypothesis (H_0), 106
- numerical variable, 11
 - continuous, 12
 - discrete, 12
- observational study, 72
- Observational units, 10
- outcome, 29
- outlier, 22
- outlier in regression, 204
- p-value, 109
- paired data, 136, 136
- percentile, 17, 54, 55
- pie chart, 15
- placebo, 72
- point estimate, 77, 77–80
 - difference of two means, 143
 - difference of two proportions, 129
 - population mean, 77
 - single proportion, 98
- pooled proportion, 131
- pooled standard deviation, 141
- population, 65, 65–68
- population parameter, 63
- population regression model, 196
- positively associated, 188
- precise, 81
- predictor, 186
- probability, 32, 29–32
- probability density function, 45
- prosecutor’s fallacy, 153
- prospective study, 76
- random phenomena, 33
- random process, 29
- random variable, 43, 43
- reference value, 111
- rejection region, 115
- relative frequency, 15

- relative frequency table, 15
- replication, 73
- residual, 195
- residuals, 198
 - contingency table, 177
 - regression, 195
- response variable, 14, 108, 186
- retrospective study, 76
- robust estimates, 20
- row totals, 172
- S, 37
- s, 19
- sample, 65
 - cluster, 69
 - cluster sample, 69
 - cluster sampling, 71
 - convenience sample, 67
 - multistage sample, 69
 - multistage sampling, 71
 - non-response, 68
 - non-response bias, 68
 - random sample, 67–68
 - representative sample, 67
 - simple random, 69
 - simple random sampling, 70
 - strata, 69
 - stratified sampling, 69, 70
- sample proportion, 97
- sample size
 - estimating a proportion, 100
- sample space, 37
- sampling distribution
 - difference of two proportions, 129
 - regression coefficient, 207
 - sample mean, 78
 - sample proportion, 98
- sampling variation, 77
- scatter plots, 194
- scatterplot, 188
- scatterplots, 193
- SE, 79
- sets, 34
- shape, 20
- Shapiro–Wilk, 155
- side-by-side boxplots, 24
- significance level, 108, 120
 - multiple comparisons, 163–164
- simple linear regression, 186
 - assumptions, 194
 - categorical predictors, 202
 - interpretation, 198
 - outliers, 204
 - R-squared (R^2), 201
- simple random sample, 67
- skew
 - example: strong, 146
 - left skewed, 20
 - right skewed, 20
- standard deviation, 19, 48
- standard error (SE), 79
 - difference in means, 143
 - difference in proportions, 129
 - regression coefficient, 208
 - single proportion, 98
- standard normal distribution, 51
- strata, 69
- stratification, 72
- strong relationships, 186
- success-failure condition, 98
- sum of squared errors, 157
- sum of squares between groups, 156
- symmetric, 20
- t-distribution, 92–94
- t-table, 93
- t-test
 - one-sample, 95
 - paired data, 136–137
 - two independent groups, 145–147
- time series, 194
- two-sided alternative, 108
- two-sided confidence intervals, 85
- Type I error, 119
- Type II error, 119
- unbiased, 81
- uncorrelated, 188
- unit of observation, 10
- variables, 10
- variance, 18, 48
- Venn diagrams, 35
- weak relationship, 186
- whiskers, 22
- Z, 52
- Z-score, 52