Introduction to R and Statistics Revision

The University of Western Australia
Agricultural Resource Economics

2018

Authors

Fogarty, J.J.
Rensing, K.
Stuckey, A.

Sections 2, 6, 7, 8, 9, 10.5 and 12.3 are derivative of OpenIntro Statistics Second Edition

Original Authors

David M Diez
Christopher D Barr
Mine Çetinkaya-Rundel

Topics Covered

List of Figures

List of Tables

1 Introduction to R

R is a language and environment for statistical analysis and data science. R was created by and (initially) for statisticians around 1997 from within the Statistics Department of the University of Auckland. R is freely available to download from www.r-project.org/.

As an R user you can write code to use inbuilt functions to import, manipulate and analyse data. R also has very flexible data visualisation functionality. As a language R allows people to create their own functions which has meant that pretty much any statistical procedure you read about - someone, somewhere, has written it up as an R function and made it available for everyone to use. When someone has written some functions that they want to share, they create a *package* which is then available to any R user to install and use. As of July 2017 there are 11,001 packages available. To get an idea of the different domains in which people have created R packages, some more popular packages have been grouped into topics here cran.r-project.org/web/views/.

This entire document was created using R! The package **knitr** is one of several packages that allow us to create documents from within R. The real beauty of this is that you can create a report based on your data manipulation, analysis, plots and tables. If the data change (say you change the data or want the same analysis with a different data set) then you can re-produce the entire report by changing a couple of lines of code and then hitting the 'run' button. As you can imagine this saves a lot of pointing, clicking, cutting, pasting, typing and hours - and if you do it well can reduce potential for errors.

More recently a lot of work has gone into making R more user-friendly, so that scientists and others who maybe don't have strong programming skills can make use of R's extensive capability. One tool that helps make R even more accessible is RStudio, which is what we will be using in this course.

Due to the R's popularity there are many, many manuals, tutorials and other learning resources online for users at all levels. Rather than give you a long list of references, let's use the resources identified by RStudio as the most useful, available at www.rstudio.com/online-learning/. For more face-to-face learning, UWA periodically offers courses on R programming in general and specific statistical techniques. See www.cas.maths.uwa.edu.au/courses.

2 Introduction to Data

Scientists seek to answer questions using rigorous methods and careful observations. These observations – collected from the likes of field notes, surveys, and experiments – form the backbone of a statistical investigation and are called **data**. Statistics is the study of how best to collect, analyse, and draw conclusions from data. It is helpful to put statistics in the context of a general process of investigation:

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyse the data.
4. Form a conclusion.

Statistics as a subject focuses on making stages 2-4 objective, rigorous, and efficient. That is, statistics has three primary components: How best can we collect data? How should it be analysed? And what can we infer from the analysis?

The topics scientists investigate are as diverse as the questions they ask. However, many of these investigations can be addressed with a small number of data collection techniques, analytic tools, and fundamental concepts in statistical inference. This chapter provides a glimpse into these and other themes we will encounter throughout the rest of the book. We introduce the basic principles of each branch and learn some tools along the way. We will encounter applications from other fields, some of which are not typically associated with science but nonetheless can benefit from statistical study.

2.1 Case Study: Using stents to prevent strokes

Section 2.1 introduces a classic challenge in statistics: evaluating the efficacy of a medical treatment. Terms in this section, and indeed much of this chapter, will all be revisited later in the text. The plan for now is simply to get a sense of the role statistics can play in practice.

In this section we will consider an experiment that studies effectiveness of stents in treating patients at risk of stroke.[1] Stents are devices put inside blood vessels that assist in patient recovery after cardiac events and reduce the risk of an additional heart attack or death. Many doctors have hoped that there would be similar benefits for patients at risk of stroke. We start by writing the principal question the researchers hope to answer:

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

The researchers who asked this question collected data on 451 at-risk patients. Each volunteer patient was randomly assigned to one of two groups:

[1] Chimowitz MI, Lynn MJ, Derdeyn CP, et al. 2011. Stenting versus Aggressive Medical Therapy for Intracranial Arterial Stenosis. New England Journal of Medicine 365:993-1003. www.nejm.org/doi/full/10.1056/NEJMoa1105335. NY Times article reporting on the study: www.nytimes.com/2011/09/08/health/research/08stent.html.

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

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

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

Researchers studied the effect of stents at two time points: 30 days after enrolment and 365 days after enrolment. The results of 5 patients are summarized in Table 2.1. Patient outcomes are recorded as "stroke" or "no event", representing whether or not the patient had a stroke at the end of a time period.

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

Table 2.1: Results for five patients from the stent study.

Considering data from each patient individually would be a long, cumbersome path towards answering the original research question. Instead, performing a statistical data analysis allows us to consider all of the data at once. Table 2.2 summarizes the raw data in a more helpful way. In this table, we can quickly see what happened over the entire study. For instance, to identify the number of patients in the treatment group who had a stroke within 30 days, we look on the left-side of the table at the intersection of the treatment and stroke: 33.

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

Table 2.2: Descriptive statistics for the stent study.

⊙ **Guided Practice 2.1** Of the 224 patients in the treatment group, 45 had a stroke by the end of the first year. Using these two numbers, compute the proportion of patients in the treatment group who had a stroke by the end of their first year. (Please note: answers to all Guided Practice exercises are provided using footnotes.)[2]

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

We can compute summary statistics from the table. A **summary statistic** is a single number summarizing a large amount of data.[3] For instance, the primary results of the study after 1 year could be described by two summary statistics: the proportion of people who had a stroke in the treatment and control groups.

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

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

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

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

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

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

2.2 Data Basics

Effective presentation and description of data is a first step in most analyses. This section introduces one structure for organizing data as well as some terminology that will be used throughout this book.

Observations, variables, and data matrices

Table 2.3 displays rows 1, 2, 3, and 50 of a data set concerning 50 emails received during early 2012. These observations will be referred to as the `email50` data set, and they are a random sample from a larger data set.

Each row in the table represents a single email or **case**.[4] The columns represent characteristics, called **variables**, for each of the emails. For example, the first row represents

[3] Formally, a summary statistic is a value computed from the data. Some summary statistics are more useful than others.

[4] A case is also sometimes called a **unit of observation** or an **observational unit**.

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	no	15,829	242	html	small

Table 2.3: Four rows from the email50 data matrix.

variable	description
spam	Specifies whether the message was spam
num_char	The number of characters in the email
line_breaks	The number of line breaks in the email (not including text wrapping)
format	Indicates if the email contained special formatting, such as bolding, tables, or links, which would indicate the message is in HTML format
number	Indicates whether the email contained no number, a small number (under 1 million), or a large number

Table 2.4: Variables and their descriptions for the email50 data set.

email 1, which is a not spam, contains 21,705 characters, 551 line breaks, is written in HTML format, and contains only small numbers.

In practice, it is especially important to ask clarifying questions to ensure important aspects of the data are understood. For instance, it is always important to be sure we know what each variable means and the units of measurement. Descriptions of all five email variables are given in Table 2.4.

The data in Table 2.3 represent a **data matrix**, which is a common way to organize data. Each row of a data matrix corresponds to a unique case, and each column corresponds to a variable. A data matrix for the stroke study introduced in Section 2.1 is shown in Table 2.1, where the cases were patients and there were three variables recorded for each patient.

Data matrices are a convenient way to record and store data. If another individual or case is added to the data set, an additional row can be easily added. Similarly, another column can be added for a new variable.

⊙ **Guided Practice 2.2** We consider a publicly available data set that summarizes information about the 2,240 areas (known as "Statistical Area 2") in Australia, and we call this the Census2016_wide_by_SA2_year data set[5]. This data set includes information about each SA2: its name, the state where it resides, its population in 2006,2011 and 2016 and many other characteristics. How might these data be organized in a data

[5]The data come from the Australian Bureau of Statistics and are available in the Census2016 R package Hugh Parsonage (2017). Census2016: Data from the Australian Census 2016. R package version 0.2.0. https://CRAN.R-project.org/package=Census2016

matrix? Reminder: look in the footnotes for answers to in-text exercises.[6]

sa2_code	sa2_name	persons	prop_units	mortgage	income
501011001	Augusta	5431	Mid	21600	61672
501011002	Busselton	26334	Low	21600	62296
501011003	Busselton Region	10280	Low	24000	83876
501011004	Margaret River	8830	Low	20796	70408
501021005	Australind - Leschenault	17592	Mid	22644	87620
501021007	Capel	5195	Mid	20796	71864
501021008	College Grove - Carey Park	6746	Low	18204	55796
501021009	Collie	8798	Low	18204	60164
501021010	Dardanup	3142	Low	24000	85020
501021011	Davenport	9	Low	0	136500
501021012	Eaton - Pelican Point	11756	Low	21840	80652

Table 2.5: First 11 rows of 2016 Census area data set for selected variables

Types of variables

Examine the `mortgage`, `persons`, `sa2_name`, and `year` variables in the `Census2016_wide_by_SA2_year` data set. Each of these variables is inherently different from the other three yet many of them share certain characteristics.

First consider `mortgage`, which is said to be a **numerical** variable since it can take a wide range of numerical values, and it is sensible to add, subtract, or take averages with those values. On the other hand, we would not classify a variable reporting telephone area codes as numerical since their average, sum, and difference have no clear meaning.

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

The variable `SA2_name` can take up to 2240 values as these are one level of the statistical areas into which Australia can be divided. Because the responses themselves are categories, `sa2_name` is called a **categorical** variable, and the possible values are called the variable's **levels**.

Finally, consider the `prop_units` variable, which describes the proportion of dwellings in the area that are units or flats and takes values `Low`, `Mid`, or `High` in area. This variable seems to be a hybrid: it is a categorical variable but the levels have a natural ordering. A variable with these properties is called an **ordinal** variable, while a regular categorical variable without this type of special ordering is called a **nominal** variable. To simplify analyses, any ordinal variables in this book will be treated as categorical variables.

[6]Each SA2 may be viewed as a case, and there are 43 pieces of information recorded for each case.

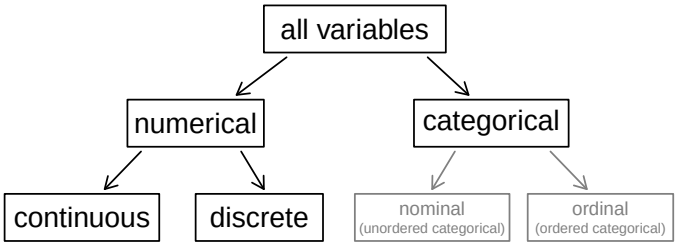


Figure 2.1: Breakdown of variables into their respective types.

Example 2.3 Data were collected about students in a statistics course. Three variables were recorded for each student: number of siblings, student height, and whether the student had previously taken a statistics course. Classify each of the variables as continuous numerical, discrete numerical, or categorical.

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

Guided Practice 2.4 Consider the variables `group` and `outcome` (at 30 days) from the stent study in Section 2.1. Are these numerical or categorical variables?[7]

Relationships between variables

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

(1) Are median rents related to proportion of dwellings that are units or flats?

(2) If homeownership is lower than the national average in one area, will the percent of units or flats in that area likely be above or below the national average?

(3) Which counties have a higher average income: those with more people born overseas or not?

To answer these questions, data must be collected, such as the `Census2016_wide_by_SA2_year` data set shown in Table 2.5. Examining summary statistics could provide insights for each of the three questions about areas. Additionally, graphs can be used to visually summarize data and are useful for answering such questions as well.

Scatterplots are one type of graph used to study the relationship between two numerical variables. Figure 2.2 compares the variables `mortgage` and `income`. Each point on the plot represents a single area at the SA2 level. For instance, the highlighted dot corresponds to SA2 503011035 in the `Census2016_wide_by_SA2_year` data set: "Nedlands - Dalkeith -

[7]There are only two possible values for each variable, and in both cases they describe categories. Thus, each is a categorical variable.

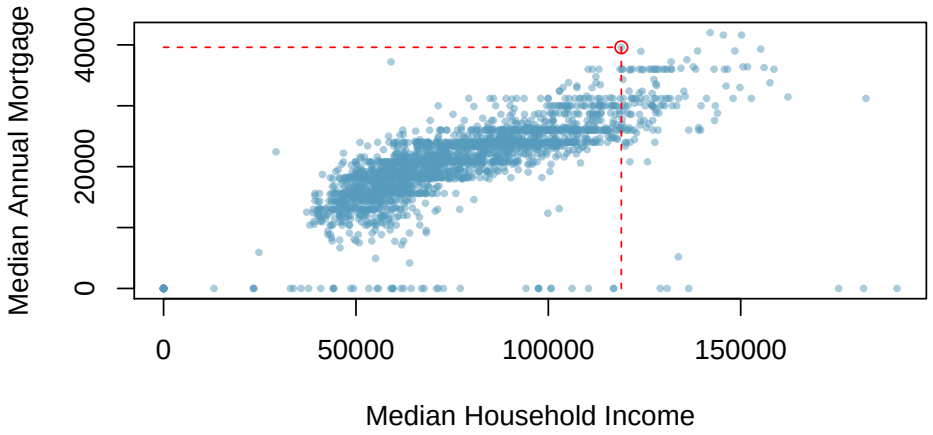


Figure 2.2: A scatterplot showing `mortgage` against `income`. The statistical area surrounding UWA, with median household income of $118,976 and median annual mortgage of $39,600, is highlighted.

Crawley", which had a median household income of $118,976 and median annual mortgage of $39,600. The scatterplot suggests a relationship between the two variables: SA2s with higher incomes tend to have higher mortgages.

⊙ **Guided Practice 2.5** Examine the variables in the `email50` data set, which are described in Table 2.4. Create two questions about the relationships between these variables that are of interest to you.[8]

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

Associated or independent, not both

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

2.3 Overview of Data Collection Principles

The first step in conducting research is to identify topics or questions that are to be investigated. A clearly laid out research question is helpful in identifying what subjects or cases

[8]Two sample questions: (1) Intuition suggests that if there are many line breaks in an email then there also would tend to be many characters: does this hold true? (2) Is there a connection between whether an email format is plain text (versus HTML) and whether it is a spam message?

should be studied and what variables are important. It is also important to consider *how* data are collected so that they are reliable and help achieve the research goals.

Populations and samples

Consider the following three research questions:

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

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

Guided Practice 2.6 For the second and third questions above, identify the target population and what represents an individual case.[9]

Sampling from a population

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

Why pick a sample randomly? Why not just pick a sample by hand? Consider the following scenario.

Example 2.7 Suppose we ask a student who happens to be majoring in nutrition to select several graduates for the study. What kind of students do you think she might collect? Do you think her sample would be representative of all graduates?

Perhaps she would pick a disproportionate number of graduates from health-related fields. Or perhaps her selection would be well-representative of the population. When selecting samples by hand, we run the risk of picking a *biased* sample, even if that bias is unintentional or difficult to discern.

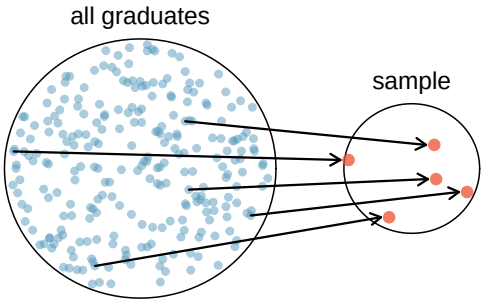


Figure 2.3: In this graphic, five graduates are randomly selected from the population to be included in the sample.

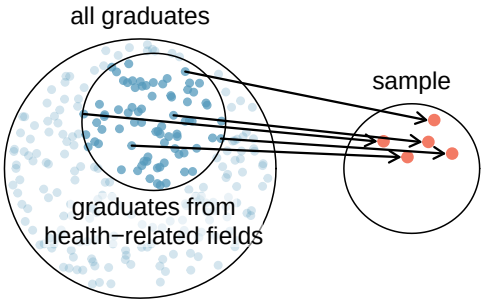


Figure 2.4: .
Instead of sampling from all graduates equally, a nutrition major might inadvertently pick graduates with health-related majors disproportionately often.

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

Sometimes a simple random sample is difficult to implement and an alternative method is helpful. One such substitute is a **systematic sample**, where one case is sampled after letting a fixed number of others, say 10 other cases, pass by. Since this approach uses a mechanism that is not easily subject to personal biases, it often yields a reasonably representative sample. This book will focus on random samples since the use of systematic samples is less common and requires additional considerations of the context.

The act of taking a simple random sample helps minimize bias, however, bias can crop up in other ways. Even when people are picked at random, e.g. for surveys, caution must be

[9](2) Notice that the first question is only relevant to students who complete their degree; the average cannot be computed using a student who never finished her degree. Thus, only UWA undergraduate students who have graduated in the last five years represent cases in the population under consideration. Each such student would represent an individual case. (3) A person with severe heart disease represents a case. The population includes all people with severe heart disease.

exercised if the **non-response** is high. For instance, if only 30% of the people randomly sampled for a survey actually respond, then it is unclear whether the results are **representative** of the entire population. This **non-response bias** can skew results.

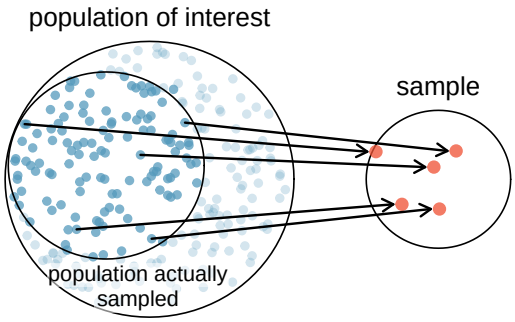


Figure 2.5: Due to the possibility of non-response, surveys studies may only reach a certain group within the population. It is difficult, and often times impossible, to completely fix this problem.

Introducing observational studies and experiments

There are two primary types of data collection: observational studies and experiments.

Researchers perform an **observational study** when they collect data in a way that does not directly interfere with how the data arise. For instance, researchers may collect information via surveys, review medical or company records, or follow a **cohort** of many similar individuals to study why certain diseases might develop. In each of these situations, researchers merely observe the data that arise. In general, observational studies can provide evidence of a naturally occurring association between variables. The methods used to establish causal connections in observational study data are complex, and continue to be debated.

When researchers want to investigate the possibility of a causal connection, they conduct an **experiment**. Usually there will be both an explanatory and a response variable. For instance, we may suspect administering a drug will reduce mortality in heart attack patients over the following year. To check if there really is a causal connection between the explanatory variable and the response, researchers will collect a sample of individuals and split them into groups. The individuals in each group are *assigned* a treatment. When individuals are randomly assigned to a group, the experiment is called a **randomized experiment**. For example, each heart attack patient in the drug trial could be randomly assigned, perhaps by flipping a coin, into one of two groups: the first group receives a **placebo** (fake treatment) and the second group receives the drug. See the case study in Section 2.1 for another example of an experiment, though that study did not employ a placebo.

TIP: association $\neq$ causation
In general, association does not imply causation, and causation can only be inferred from a randomized experiment.

2.4 Observational Studies and Sampling Strategies

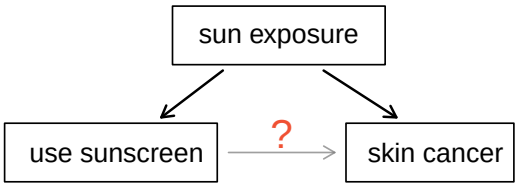
Observational studies

Generally, data in observational studies are collected only by monitoring what occurs, while experiments require the primary explanatory variable in a study be assigned for each subject by the researchers.

Making causal conclusions based on experiments is often reasonable. However, making the same causal conclusions based on observational data can be treacherous and is not recommended. Thus, observational studies are generally only sufficient to show associations.

⊙ **Guided Practice 2.8** Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen *causes* skin cancer?[10]

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



Sun exposure is what is called a **confounding variable**,[11] which is a variable that is correlated with both the explanatory and response variables. While one method to justify making causal conclusions from observational studies is to exhaust the search for confounding variables, there is no guarantee that all confounding variables can be examined or measured.

2.5 Experiments

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

Principles of experimental design

Randomized experiments are generally built on four principles.

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

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

Controlling. Researchers assign treatments to cases, and they do their best to **control** any other differences in the groups. For example, when patients take a drug in pill form, some patients take the pill with only a sip of water while others may have it with an entire glass of water. To control for the effect of water consumption, a doctor may ask all patients to drink a 300mL glass of water with the pill.

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

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

Blocking. Researchers sometimes know or suspect that variables, other than the treatment, influence the response. Under these circumstances, they may first group individuals based on this variable into **blocks** and then randomize cases within each block to the treatment groups. This strategy is often referred to as **blocking**. For instance, if we are looking at the effect of a drug on heart attacks, we might first split patients in the study into low-risk and high-risk blocks, then randomly assign half the patients from each block to the control group and the other half to the treatment group, as shown in Figure 2.6. This strategy ensures each treatment group has an equal number of low-risk and high-risk patients.

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

Reducing bias in human experiments

Randomized experiments are the gold standard for data collection, but they do not ensure an unbiased perspective into the cause and effect relationships in all cases. Human studies are perfect examples where bias can unintentionally arise. Here we reconsider a study where a new drug was used to treat heart attack patients.[12] In particular, researchers wanted to know if the drug reduced deaths in patients.

These researchers designed a randomized experiment because they wanted to draw causal conclusions about the drug's effect. Study volunteers[13] were randomly placed into two study groups. One group, the **treatment group**, received the drug. The other group, called the **control group**, did not receive any drug treatment.

[12] Anturane Reinfarction Trial Research Group. 1980. Sulfinpyrazone in the prevention of sudden death after myocardial infarction. New England Journal of Medicine 302(5):250-256.

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

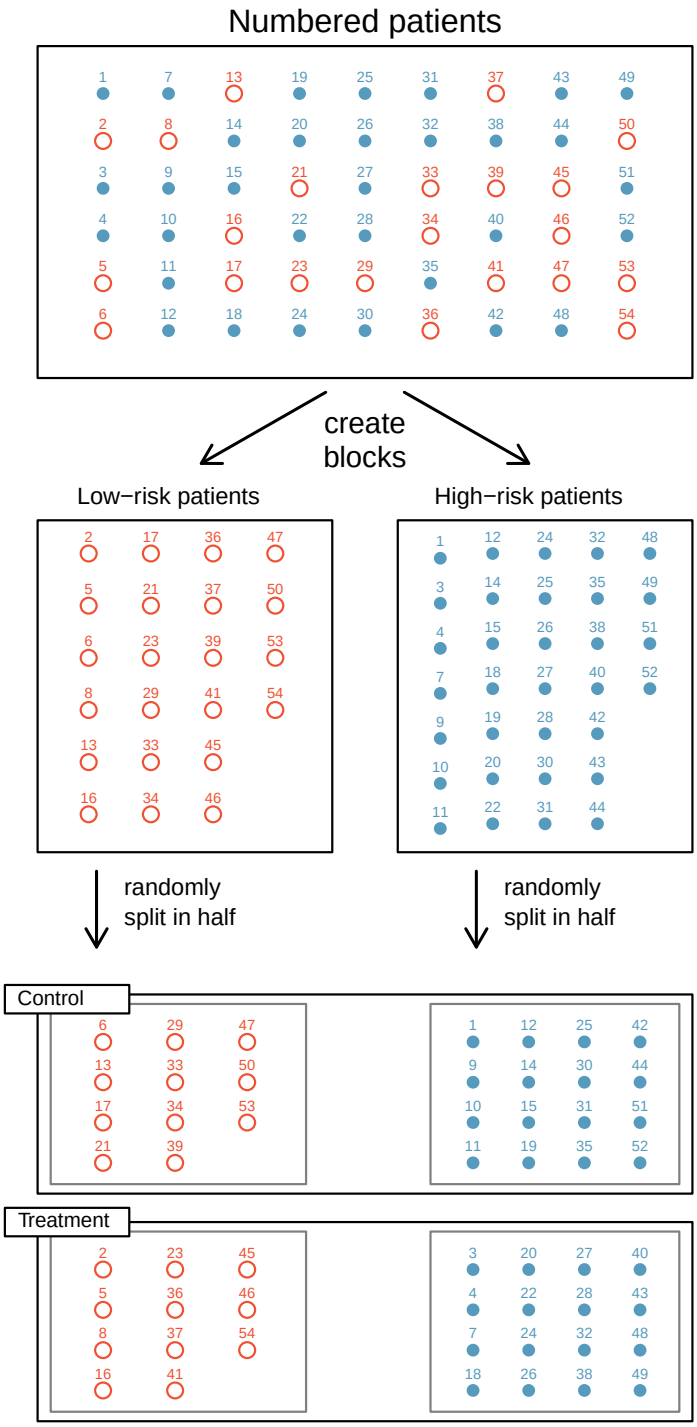


Figure 2.6: Blocking using a variable depicting patient risk. Patients are first divided into low-risk and high-risk blocks, then each block is evenly separated into the treatment groups using randomization. This strategy ensures an equal representation of patients in each treatment group from both the low-risk and high-risk categories.

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

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

The patients are not the only ones who should be blinded: doctors and researchers can accidentally bias a study. When a doctor knows a patient has been given the real treatment, she might inadvertently give that patient more attention or care than a patient that she knows is on the placebo. To guard against this bias, which again has been found to have a measurable effect in some instances, most modern studies employ a **double-blind** setup where doctors or researchers who interact with patients are, just like the patients, unaware of who is or is not receiving the treatment.[14]

Variance and standard deviation

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

We call the distance of an observation from its mean its **deviation**. Below are the deviations for the 1^{st}, 2^{nd}, 3^{rd}, and 50^{th} observations in the `num_char` variable in the `email` data set. For computational convenience, the number of characters is listed in the thousands

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

and rounded to the first decimal, and the $\bar{x}$ denotes the sample mean.

$$\begin{aligned} x_1 - \bar{x} &= 21.7 - 11.6 = 10.1 \\ x_2 - \bar{x} &= 7.0 - 11.6 = -4.6 \\ x_3 - \bar{x} &= 0.6 - 11.6 = -11.0 \\ &\vdots \\ x_{50} - \bar{x} &= 15.8 - 11.6 = 4.2 \end{aligned}$$

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

s^2 sample variance

$$\begin{aligned} s^2 &= \frac{10.1^2 + (-4.6)^2 + (-11.0)^2 + \cdots + 4.2^2}{50 - 1} \\ &= \frac{102.01 + 21.16 + 121.00 + \cdots + 17.64}{49} \\ &= 172.44 \end{aligned}$$

We divide by $n - 1$, rather than dividing by n, when computing the variance; you need not worry about this mathematical nuance for the material in this textbook. Notice that squaring the deviations does two things. First, it makes large values much larger, seen by comparing 10.1^2, $(-4.6)^2$, $(-11.0)^2$, and 4.2^2. Second, it gets rid of any negative signs.

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

$$s = \sqrt{172.44} = 13.13$$

s sample standard deviation

The standard deviation of the number of characters in an email is about 13.13 thousand. A subscript of $_x$ may be added to the variance and standard deviation, i.e. s_x^2 and s_x, as a reminder that these are the variance and standard deviation of the observations represented by x_1, x_2, ..., x_n. The $_x$ subscript is usually omitted when it is clear which data the variance or standard deviation is referencing.

Variance and standard deviation

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

Formulas and methods used to compute the variance and standard deviation for a population are similar to those used for a sample.[15] However, like the mean, the population values have special symbols: σ^2 for the variance and σ for the standard deviation. The symbol σ is the Greek letter *sigma*.

σ^2 population variance

σ population standard deviation

[15] The only difference is that the population variance has a division by n instead of $n - 1$.

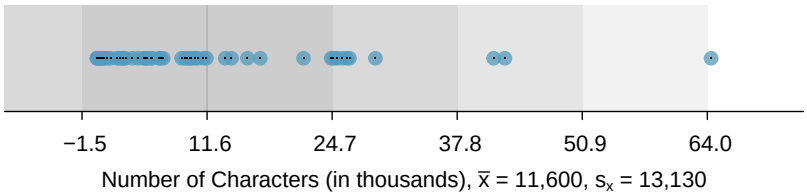


Figure 2.7: In the `num_char` data, 41 of the 50 emails (82%) are within 1 standard deviation of the mean, and 47 of the 50 emails (94%) are within 2 standard deviations. Usually about 70% of the data are within 1 standard deviation of the mean and 95% are within 2 standard deviations, though this rule of thumb is less accurate for skewed data, as shown in this example.

TIP: standard deviation describes variability

Focus on the conceptual meaning of the standard deviation as a descriptor of variability rather than the formulas. Usually 70% of the data will be within one standard deviation of the mean and about 95% will be within two standard deviations. However, as seen in Figures 2.7 and 2.8, these percentages are not strict rules.

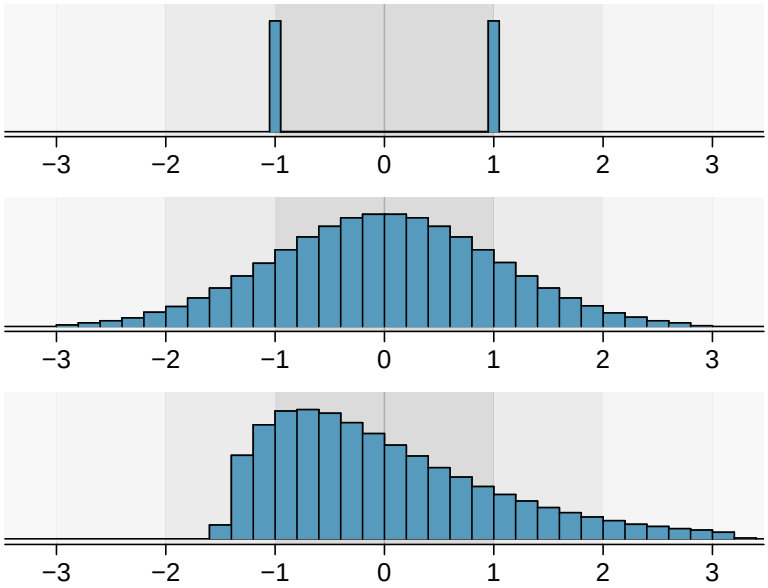


Figure 2.8: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

⊙ **Guided Practice 2.9** A good description of the shape of a distribution should include modality and whether the distribution is symmetric or skewed to one side. Using Figure 2.8 as an example, explain why such a description is important.[16]

[16]Figure 2.8 shows three distributions that look quite different, but all have the same mean, variance, and standard deviation. Using modality, we can distinguish between the first plot (bimodal) and the last two (unimodal). Using skewness, we can distinguish between the last plot (right skewed) and the first two. While

In practice, the variance and standard deviation are sometimes used as a means to an end, where the "end" is being able to accurately estimate the uncertainty associated with a sample statistic. For example, we can use the variance and standard deviation to assess how close the sample mean is to the population mean.

Robust statistics

How are the sample statistics of the `num_char` data set affected by the observation, 64,401? What would have happened if this email wasn't observed? What would happen to these summary statistics if the observation at 64,401 had been even larger, say 150,000? These scenarios are plotted alongside the original data in Figure 2.9, and sample statistics are computed under each scenario in Table 2.6.

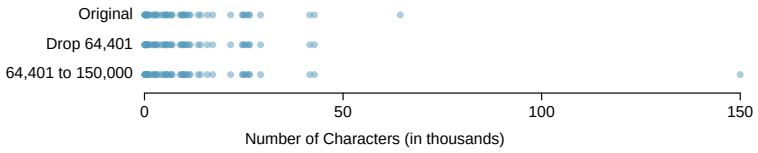


Figure 2.9: Dot plots of the original character count data and two modified data sets.

	robust		**not robust**	
scenario	median	IQR	$\bar{x}$	s
original `num_char` data	6,890	12,875	11,600	13,130
drop 64,401 observation	6,768	11,702	10,521	10,798
move 64,401 to 150,000	6,890	12,875	13,310	22,434

Table 2.6: A comparison of how the median, IQR, mean ($\bar{x}$), and standard deviation (s) change when extreme observations are present.

Guided Practice 2.10 (a) Which is more affected by extreme observations, the mean or median? Table 2.6 may be helpful. (b) Is the standard deviation or IQR more affected by extreme observations?[17]

The median and IQR are called **robust estimates** because extreme observations have little effect on their values. The mean and standard deviation are much more affected by changes in extreme observations.

Example 2.11 The median and IQR do not change much under the three scenarios in Table 2.6. Why might this be the case?

a picture, like a histogram, tells a more complete story, we can use modality and shape (symmetry/skew) to characterize basic information about a distribution.

[17] (a) Mean is affected more. (b) Standard deviation is affected more. Complete explanations are provided in the material following Guided Practice 10.

The median and IQR are only sensitive to numbers near Q_1, the median, and Q_3. Since values in these regions are relatively stable – there aren't large jumps between observations – the median and IQR estimates are also quite stable.

⊙ **Guided Practice 2.12** The distribution of vehicle prices tends to be right skewed, with a few luxury and sports cars lingering out into the right tail. If you were searching for a new car and cared about price, should you be more interested in the mean or median price of vehicles sold, assuming you are in the market for a regular car?[18]

[18] Buyers of a "regular car" should be concerned about the median price. High-end car sales can drastically inflate the mean price while the median will be more robust to the influence of those sales.

3 Histograms

The histogram is one of the most important plots used in Science. Histograms allow us to visualize the distribution of a data set. From a histogram it is relatively easy to see where most of the observations lie and get some idea about the variability of observations around the mean. Histograms also make it easy to see the maximum and minimum values in a data set; determine whether the distribution is symmetric or skewed in one direction or another; and detect outliers or unusual observations.

In a histogram the vertical axis indicates the frequency (or relative frequency) of the observations, and the horizontal axis describes the variable we are interested in. The key to creating a histogram is to group the observations into fixed intervals, and plot the number of observations in each interval. For example, consider the set: `2, 13, 15, 23, 24, 25, 26, 29, 36, 37, 38, 42, 49`. If we create intervals of: `0-10, 10-20, 20-30, 30-40, and 40-50`, the summary data we need to plot is:

Table 3.1: Frequency table

Intervals (bins)	0-10	10-20	20-30	30-40	40-50
Frequency	1	2	5	3	2

There is no one fixed rule for creating histogram bin sizes, but **R** has a number of pre-programmed options that simplify the process. The default decision rule is based on the method of Sturges, but as you will see, there are others. These short cuts are not available in generic spreadsheet software such as MS Excel.

3.1 Creating a Basic Histogram

First, we need some data. Let's use the **PlantGrowth** data set in the base **R** package **datasets**. Normally we will need to load a data set into **R**, but here we are using one of the in-built sample data sets. To look at the data structure we use the function *str()*. When you apply this function the output says that there are 30 observations and two variables. One of the variables is a continuous variable: weight; the other variable is a categorical variable: treatment. The categorical variable has three levels. When using **R** the first step is to always check the data structure to see what has been read in.

Our next step is to create a visual representation of the data. Specifically, we want to create a histogram. For our initial plot we will ignoring the treatment variable for the moment and just create a histogram of plant weights. The easiest way to create a histogram is to use the default **R** parameter settings. Creating a basic histogram is easy. You just tell **R** the name of the data set you are working with – *PlantGrowth* – which you do via the *with()* command; and then you specify the name of the column of data you want to apply the *hist()* function to, which in this instance is the 'weight' column.

```r
str(PlantGrowth)  # function to look at the data 'structure'

## 'data.frame': 30 obs. of  2 variables:
##  $ weight: num  4.17 5.58 5.18 6.11 4.5 4.61 5.17 4.53 5.33 5.14 ...
##  $ group : Factor w/ 3 levels "ctrl","trt1",..: 1 1 1 1 1 1 1 1 1 1 ...

with(PlantGrowth, hist(weight))  # read as: apply the hist() function to the
# weight column of the Plant growth data set
```

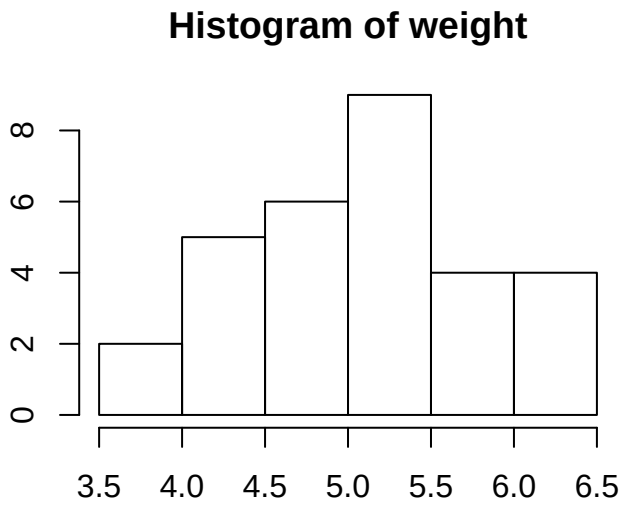


Figure 3.1: Illustration of a basic histogram using the default parameter options in R.

Note that throughout these guides, code blocks do not show the **R** command-line prompt '>', and console outputs are denoted with a double ##. This allows you to easily copy-paste the commands to see the figures and outputs in your own **R** software application, and create your own reference scripts. You are encouraged to start the learning process with this copy and paste approach.

3.2 Optional Parameters

Now, let's explore some of the optional parameters of the function *hist()* and compare the two plots. We will use `xlim` and `ylim` to specify the axes ranges, `col` and `border` to set the column fill colour and the border colour for the bins, and `main`, `xlab` and `ylab` to label the histogram and axes. Within the code, function parameters are separated by commas and contained within the (parentheses) following the function name. Notice how specifying these parameters can allow you to create a publication quality figure that best illustrates your data.

```r
with(PlantGrowth,hist(weight))
with(PlantGrowth,hist(weight,
      xlim= c(3,7),                          # set the x-axis range
      ylim= c(0,10),                         # set the y-axis range
      col= "lightgrey",                      # fill the columns
      border= "black",                       # column border
      main= "Histogram of Plant Weights",    # figure title
      xlab= "Dried weight (g)",              # x-axis label
      ylab= "Frequency"))                    # y-axis label
```

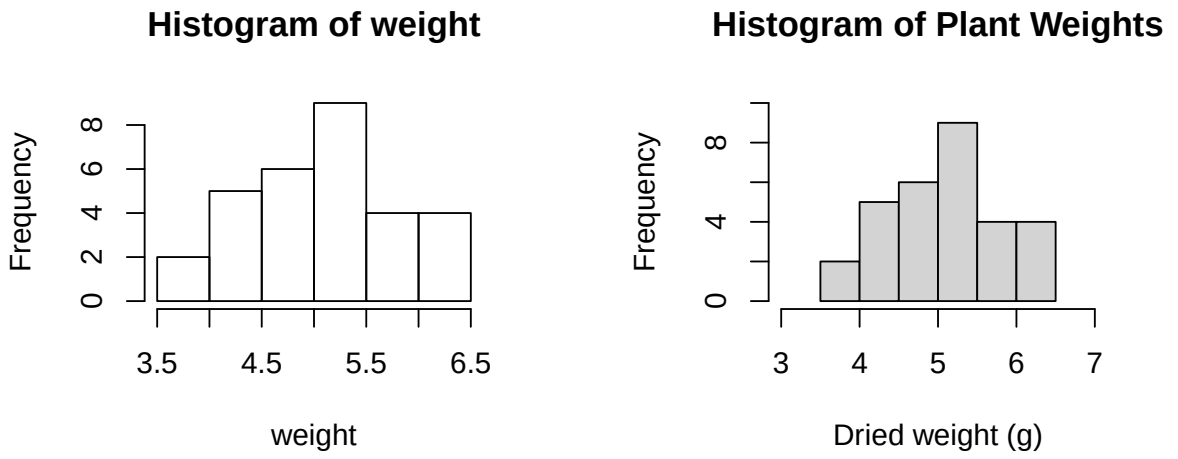


Figure 3.2: Comparison of a histogram with the default settings and one customised by setting the function parameters.

3.3 Changing Bin Sizes

In addition to specifying the axes values, we can also set the bin sizes in a histogram using the optional parameter `breaks`. Breaks can be specified using algorithms included in the function, the default is `"Sturges"`, or by creating custom bins where you set the sequence with *seq()*.

For this example we will simulate (make up) our data so we have a greater number of observations drawn from a wider range of values. We will use two new functions to create a data frame *data.frame()* that holds our data. A data.frame is the term used for any data set you read into *R*. The simulated data will be data that matches a normal distribution *rnorm()*. To set the number of observation in the data frame we use (`n`); and we set the mean and the standard deviation with (`mean, sd`). Data simulation is a neat trick to know about. Welcome to the data simulation club!

We will then use the function *names()* to change our simulated variable's name. Notice that with this function we use [brackets] to specify the column number to change, rather than (parentheses) which are used to enclose parameters of a function, e.g. *names(*`simulated_data`*)*. This may be confusing at first, but over time it becomes more clear.

Try executing `str(simulated_data)` before and after the *names()* function is used to better understand what it is doing. It is important to always have variable names which are long enough to easily understand, but short enough to keep your code clean and reduce unnecessary typing. Also, avoid using spaces in variable names as **R** is not able to easily recognize the words together. This is why we have used an underscore, '_', between 'var' and 'name' to make 'var_name' below.

```
# 1. Simulate Data:

set.seed(1234)  # allow data to be reproducible (you will get exactly the same!)
# we want 1000 obs, a mean of 500 and sd of 50
simulated_data <- data.frame(rnorm(n = 1000, mean = 500, sd = 50))
# assign variable name [column #1] as 'var_name'
names(simulated_data)[1] <- "var_name"
```

So, now we are actually at the normal starting point of analysis. Whenever we have a data frame in **R** the first thing we do is apply a function like the *str()* function to check the structure of the data object.

```
str(simulated_data)

## 'data.frame': 1000 obs. of  1 variable:
##  $ var_name: num  440 514 554 383 521 ...
```

Can you match the data structure reported back by **R**?

Now let's look at a series of histograms where we use different decision rules to decide the bin widths. There is no wrong or right way to decide on the bin widths. The default option is always OK, but sometimes you can improve on the default **R** settings. You really just need to use a trial and error approach.

```
with(simulated_data, hist(var_name, breaks = "Sturges", main = "Default bins"))
with(simulated_data, hist(var_name, breaks = "Scott", main = "Scott bins"))
with(simulated_data, hist(var_name, breaks = "FD", main = "FD bins"))
with(simulated_data, hist(var_name, breaks = (seq(from = 100, to = 700, by = 10)),
    main = "Custom bins"))
```

Note the colour-coding in the code examples here - it can really help to understand how **R** works. Red depicts a function, green a parameter, blue and pink show your input for

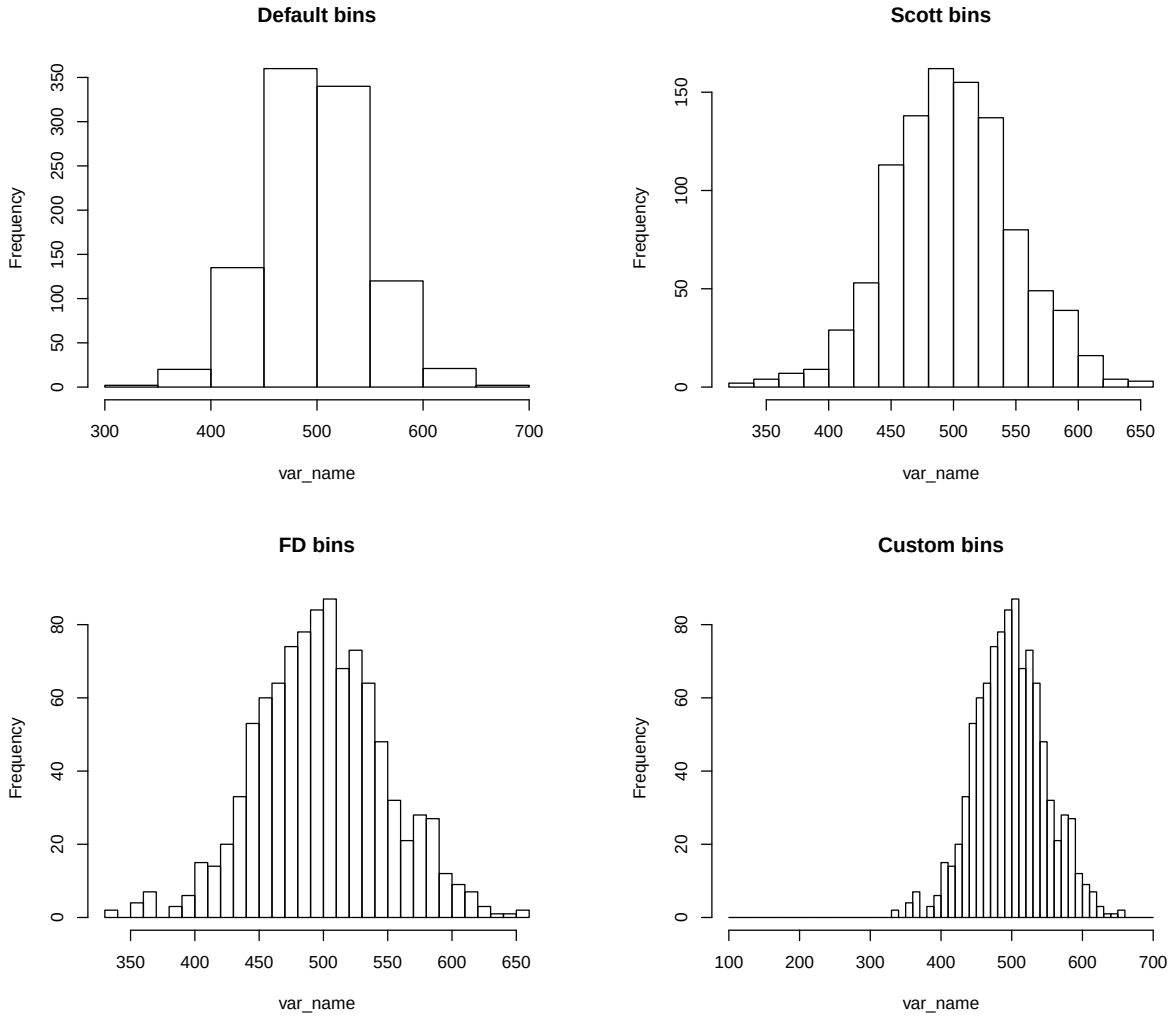


Figure 3.3: Example of histogram break options in R (Scott, FD and Sturges - default) and customised breaks.

character and numerical parameters, and black signifies your data (data set and variable name).

3.4 Advanced Histogram Features

This subsection covers how you would add extra information to your histogram using more complicated coding. These features are only relevant for students in the latter part of their degree and are not needed for students in an introductory course such as SCIE1104.

The below code may seem confusing at first, but it is actually easier than attempting to manually format a histogram (or any figure) in MS Excel to publication standard. We will use the same **PlantGrowth** data as in the earlier examples, but we will also add a vertical line for the mean and two vertical lines to indicate one standard deviation on each side of the

mean using the function *abline()*. Additionally, we will also include a basic legend showing what these lines signify with *legend()*, and tweak the axes formats slightly.

The parameters used in these functions are as follows:

- `lty` = Line TYpe (a number relating to a line type, e.g. solid or dashed)
- `lwd` = Line WiDth (larger number indicates thicker line)
- `col` = COLour of line (can be a name or number relating to a colour)
- `xaxs` = 'i' forces the x-axis to fit the 'internal' data range (3 to 8)
- `yaxs` = 'i' forces the y-axis to fit the 'internal' data range (0 to 10)
- `v` = variable name specifying the values on the x-axis for Vertical lines. To create horizontal lines you would use 'h' and specify the y-axis variable (only for abline)
- `legend` = a vector, specified using function *c()*, short for concatenate, of the names of the items to include in the legend (only for legend)
- "topright" = this specifies the location of the legend and can also be written as x,y coordinates.

Before we create out plot let's find the mean and standard deviation and save them as 'objects'.

```
# calculate the mean & SD and save them as objects so they can be called
# upon when creating our vertical lines
mean.weight <- with(PlantGrowth, mean(weight))
sd.weight <- with(PlantGrowth, sd(weight))
mean.weight

## [1] 5.073

sd.weight  # see how the mean and sd have now been stored?

## [1] 0.7011918
```

Now let's create the custom histogram. Note that because the x-axis range does not extend to zero some people would argue that the **R** defaults are preferable to the format shown here. It is an open question.

```
# plot our histogram
with(PlantGrowth, hist(weight,xlim = c(3,8),ylim = c(0,10), xaxs="i", yaxs="i"))
# add vertical lines for the mean and 1 sd above and below the mean
abline(v= mean.weight, lty=1, lwd=3,col="blue")
```

```r
abline(v= mean.weight - (sd.weight), lty=2, lwd=1, col="red")
abline(v= mean.weight + (sd.weight), lty=2, lwd=1, col="red")
# add our legend
legend("topright",              # specify location
       legend=c("Mean", "Mean +/- SD"), # specify contents of legend
       col=c("blue","red"),     # specify colours in order of legend contents
       lwd=c(3,1),              # specify line widths in order
       lty=c(1,2))              # specify line types in order
```

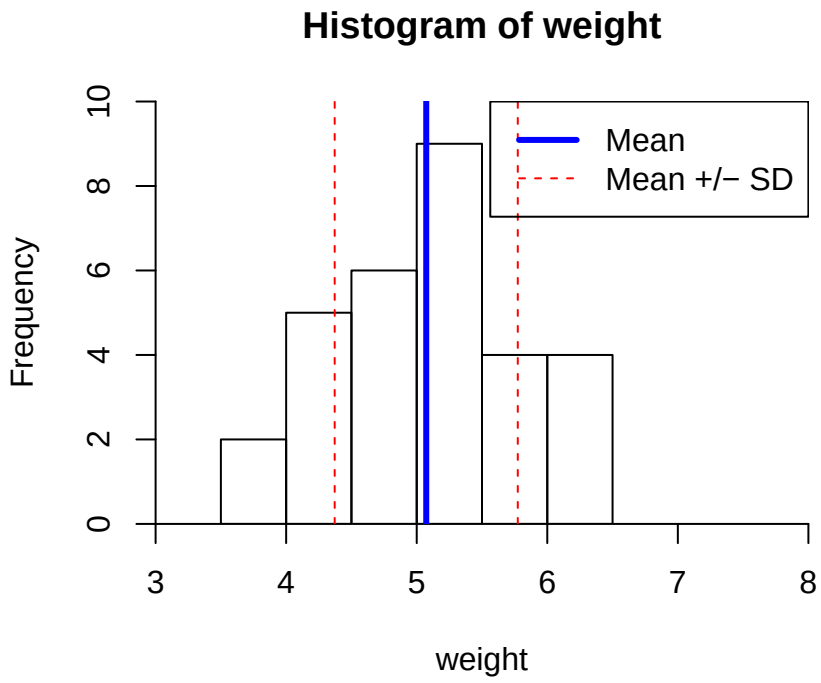


Figure 3.4: Histogram displaying measures of central tendency.

4 Boxplots

Histograms are most effective when you want to describe a single distribution. When you want to compare groups a boxplot is more effective. A boxplot is a visual representation of the five number summary of a data set. The five number summary of a distribution is an alternative to using the mean and standard deviation to describe the distribution. Recall that the five number summary consists of: the maximum and minimum values, the interquartile range (IQR), and the median.

In a boxplot the box shows the IQR, and the whiskers show the maximum and minimum values. A horizontal line is used to indicate the median. Sometimes drawing the whiskers all the way to the maximum or minimum values distorts the visual representation of the data. As such, a limiting rule for how far the whiskers should extend past the box is also applied. The default decision rule in most software programs is to extend the whiskers no more than 1.5 times the IQR. When a value extends past this point, rather than extending the whisker to that value it is typically identified with a dot or other symbol. Sometimes these observations are referred to as 'outliers'.

4.1 Creating a Basic Boxplot

This example uses the **PlantGrowth** data set in base **R**. The **PlantGrowth** data set has two variables and the data set is organised in what is called long-format or stacked-format. Long-format data means one column contains numerical values (`weight`) and one column lists the context of the value (`group`). We say, `weight` is a continuous/numerical variable and `group` is a factor/categorical variable. The alternative to long-format data is wide-format data. Wide-format data has the measurement values listed in separate columns for each group of the factor variable. In Science it is common to see data organised in long-format. In the Social Sciences it is common to see data organised in wide-format.

Let's create a boxplot to show the `weight` distributions of the three plant groups: ctrl, trt1, and trt2. To create a boxplot we use a formula that takes the general form: `boxplot(y ~ x)`. In the formula y is the numerical (weight) variable and x the factor (group) variable. We apply the command to the data **PlantGrowth** using the function *with()*. For our initial plot we use the default boxplot parameters in **R**. In subsequent sections we explore how to customize boxplots.

Remember, before creating any plots you must first look at the data to make sure it has been read in correctly and there are no obvious errors. Here we use *str()* to look at the structure of the data, and then *summary()* to obtain 'summary' information on the data values. The *str()* function tells us we have two columns of data: the numerical values and a column with the names that identify the grouping. The *summary()* function provides a five number summary of the weights data (ignoring the grouping structure), plus the mean; and also lists the different 'levels' of our factor variable (ctrl, trt1, trt2). The output also shows how many observations we have for each level; which in this case is ten observations for each level.

long format

weight	group
4.17	ctrl
5.58	ctrl
5.18	ctrl
6.11	ctrl
4.17	trt1
4.41	trt1
3.59	trt1
5.87	trt1
6.31	trt2
5.12	trt2
5.54	trt2
5.50	trt2

wide format

ctrl	trt1	trt2
4.17	4.17	6.31
5.58	4.41	5.12
5.18	3.59	5.54
6.11	5.87	5.50

Figure 4.1: Long and wide format data

```
str(PlantGrowth)  # we have 30 observations, 2 variables

## 'data.frame': 30 obs. of  2 variables:
##  $ weight: num  4.17 5.58 5.18 6.11 4.5 4.61 5.17 4.53 5.33 5.14 ...
##  $ group : Factor w/ 3 levels "ctrl","trt1",..: 1 1 1 1 1 1 1 1 1 1 ...

summary(PlantGrowth)  # a summary for each variable in PlantGrowth

##      weight       group
##  Min.   :3.590   ctrl:10
##  1st Qu.:4.550   trt1:10
##  Median :5.155   trt2:10
##  Mean   :5.073
##  3rd Qu.:5.530
##  Max.   :6.310
```

Once we have a clear understanding of the data structure we can then create our boxplot.

```
# now create your box plot
with(PlantGrowth, boxplot(weight ~ group))  # long data, use ~ not ,
```

Now look at the boxplot. With a simple visualisation we can already see several important things. First, the dispersion of measurements for treatment 2 looks to be smaller than for treatment 1 and the control. Second, the observations for treatment 2 are, on average, a little higher than for the control and treatment 1. We have not conducted any formal

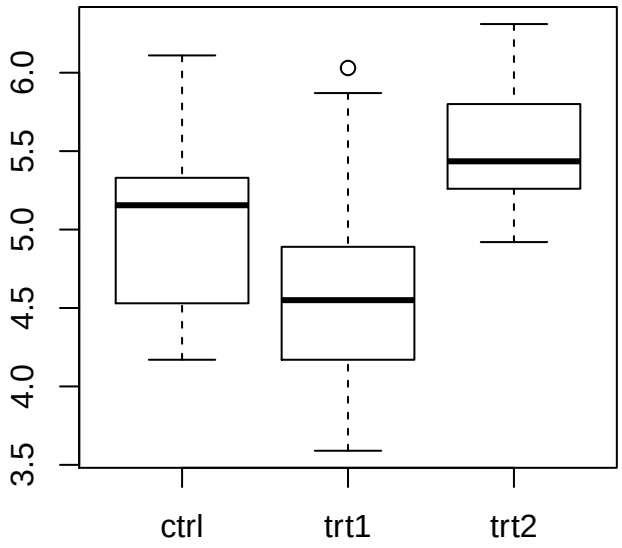


Figure 4.2: Illustration of a basic box plot using the default parameter options in R.

statistical tests, and these differences could be due to sampling variation only, but the visual plot provides us with important information about the data we are working with.

4.2 Optional Parameters

Now let's modify our boxplot to get something that is publication standard, something that is almost impossible with MS Excel. All but one of the parameters we use here – `boxwex`, which is used for changing the width of your boxes – were covered in the histogram example. If you are feeling lost with the material you can refer back to the histogram example.

```
with(PlantGrowth,boxplot(weight ~ group,
    col= "gray",                    # box colour (US and UK spelling both work)
     border= "black",                    # box border
     main= "Box Plot of Plant Weights",  # figure title
     xlab= "Treatment",                  # x-axis label
     ylab= "Dried weight (g)",           # y-axis label
     ylim= c(3,7),                       # y-axis range
     boxwex= 0.6))                       # set box widths (60% of default)
```

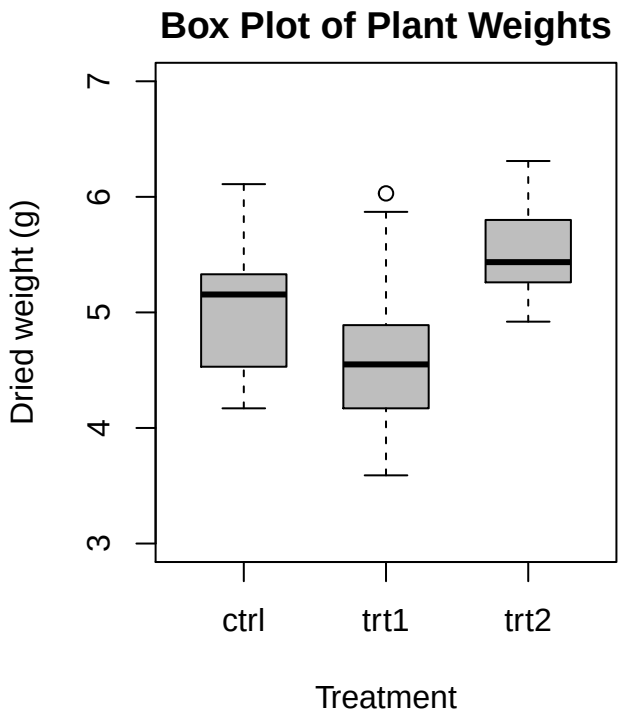


Figure 4.3: Illustration of a box plot with customised parameters in R.

The boxes are now narrower; we have clean axis labels that are informative; and the shading for the boxes looks like something we would see in a science journal.

4.3 Horizontal Box Plots

There are several reasons you may want a horizontal boxplot. One reason is to draw more attention to the continuous variable's values by placing them on the horizontal axis. Also, when you only have one group, a horizontal boxplot looks better than a vertical boxplot. In **R** it is simple to create a horizontal boxplot; you use the same command as before, but set the parameter `horizontal` as TRUE.

```
with(PlantGrowth,boxplot(weight ~ group,
     col= "gray",
     main= "Horizontal Box Plot",
     xlab= "Dried weight (g)",
     ylab= "Treatment",
     horizontal = TRUE))  # set horizontal to TRUE
```

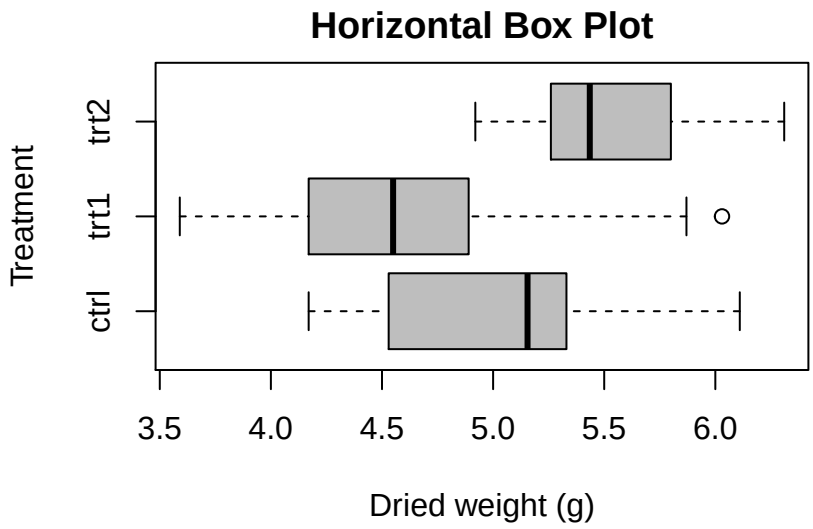


Figure 4.4: Illustration of a horizontal box plot in R.

4.4 Advanced Boxplot Features

Now let's explore how we can change: the fill colour and outlier format; group names and the order of our groups (factor levels); and set the orientation of the axis labels so they read horizontally. This material is advanced and not relevant for students in an introductory course such as SCIE1104.

The default group order in **R** is for the groups to be ordered alphabetically. However, we might prefer to have the treatments listed first and the control group last. As 'c' in 'ctrl' comes before 't' in 'trt' we will need to tell **R** the order we want using the function *factor()*.

We might also want to use full names and spaces in the labels to more clearly illustrate their meanings. Let's rename the levels of our factor variable as 'Treatment 1', 'Treatment 2', and 'Control'. Since we still want the abbreviations to be used in our actual data we will change the names only in the plot, using the *boxplot()* parameter `names`, not the raw data set. We will also provide a list of colours to better contrast the control group with the treatment groups and change the orientation of the y-axis values (weights) so they are horizontal. As a final step we will also change the symbol and colour used to identify the outlier observation in treatment 1 group.

The following steps will be followed to produce our new boxplot:

1. Change the order of factor levels with *factor()*, by listing them in parameter `levels`.
2. Create your boxplot as before, but setting the following parameters:
   (a) use `names` to list the labels for each box using the new order of groups
   (b) use `col` to list the colours to fill the boxes with

(c) set `las` to 1 to change your y-axis values to horizontal (0=parallel to axis (default), 1=horizontal, 2=perpendicular to axis, 3=vertical)

(d) use `outpch` to change the change the outlier symbol (20= solid dot), and `outcol` to tell R what colour to use for the symbol

Note: to see a list of the colours available in **R**, type the following command into your console: `colours(distinct = FALSE)`, and see also Chapter 6, Graphical Parameters.

```
#1: Reorder x-axis factor levels
# Note: data$variable is used to specify your data here, instead of with()
PlantGrowth$group<-factor(PlantGrowth$group, levels= c("trt1", "trt2", "ctrl"))
# check the order has changed
summary(PlantGrowth$group)

## trt1 trt2 ctrl
##   10   10   10

#2. create a plot
with(PlantGrowth,boxplot(weight ~ group,
     col= c("grey70","grey70","grey40"),      # list colours to fill boxes
     names= c("Treatment 1", "Treatment 2","Control"), # list box names
     main= "Publication quality example",
     ylab= "Dried weight (g)",
     outpch=20,                           # change outlier symbol
     outcol="grey70",                     # change outlier colour
     ylim= c(3,7),                        # y-axis range
     boxwex= 0.6,                         # make the box width thinner
     las= 1))                             # set axis values to horizontal
```

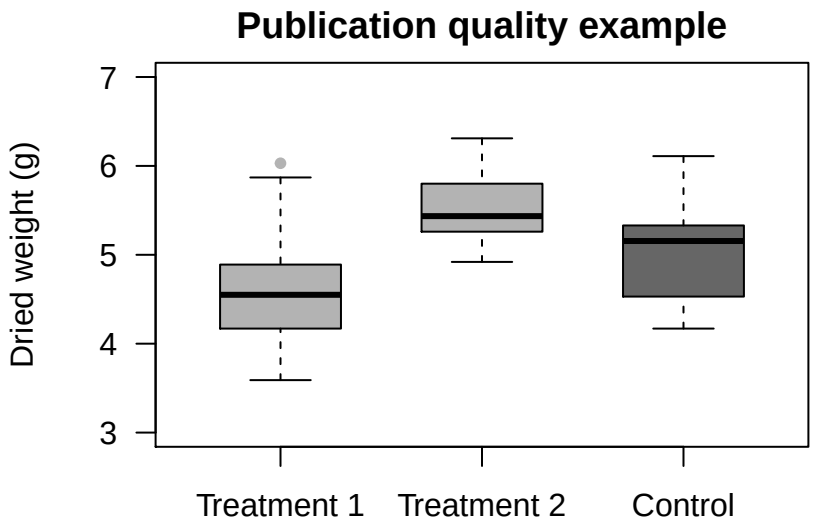


Figure 4.5: Illustration of a box plot with customised axes in R.

5 Scatterplots

When we have measurements on two dimensions for each observation – for example the height and dry weight of a plant – rather than plot a separate histogram for height and a separate histogram for weight it is generally more useful to plot both the height and weight information in a single figure. As the two measurements on each item generally have different units of measurement – e.g. weight in (g) or (kg) and height in (cm) or (in) – it is not possible to use a boxplot to represent the information. For such data the most effective visual representation is likely to be a scatter plot.

5.1 Creating a Basic Scatter Plot

In this example we will use the **iris** data set in base **R**. This is a famous data set that is associated with the work of R. Fisher (although he did not actually collect the data). The data set has four numerical variables related to Irises: petal length, petal width, sepal length, and sepal width; and one factor/grouping variable: species. To look at the data structure you can use use *str()* and *summary()* as shown previously. An additional **R** command that allows you to look at the data structure is: *head()*, and this is the command used here. This command shows the first six rows of the data, including variable names, and is also a good way to make sure you understand the way your data has been read into **R**.

```
head(iris)  # prints the first few rows of the iris data set
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

We will use the generic *plot()* function to create a scatter plot of petal lengths by petal widths, and specify our formula the same way we did for boxplots: `plot(y ~ x)`. **R** will automatically create a scatter plot because the command refers to two numerical variables. As in previous guides, let's use the default **R** parameters first. For our initial plot we will ignore the grouping variable.

```
with(iris, plot(Petal.Length ~ Petal.Width))
```

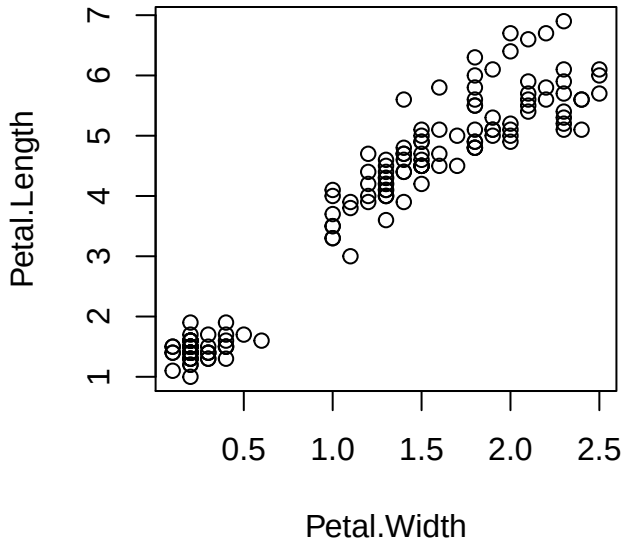


Figure 5.1: Illustration of a scatter box plot using the default parameter options in R.

5.2 Optional Parameters

From looking at the default plot, can you think of parameters you would want to change or add to improve your scatter plot? Let's use the same parameters covered in previous guides, adding just one new thing in this section. We saw in the histograms guide how to change line type using `lty`; here we will use `pch` (Point CHange) to set the symbol to be used as our points. The default is 1, a circle, and options range from 0 to 25. To see all symbols available see Chapter 6, Graphical Parameters.

```
with(iris,plot(Petal.Length ~ Petal.Width,
     col= "purple",                    # points colour
     pch= 5,                           # symbol 5 for rhombus/diamond
     main= "Petal Lengths and Widths", # figure title
     ylab= "Petal length (cm)",        # y-axis label
     xlab= "Petal width (cm)",         # x-axis label
     ylim= c(0,8),                     # y-axis range
     xlim= c(0,3),                     # x-axis range
     las= 1))                          # horizontal axis labels
```

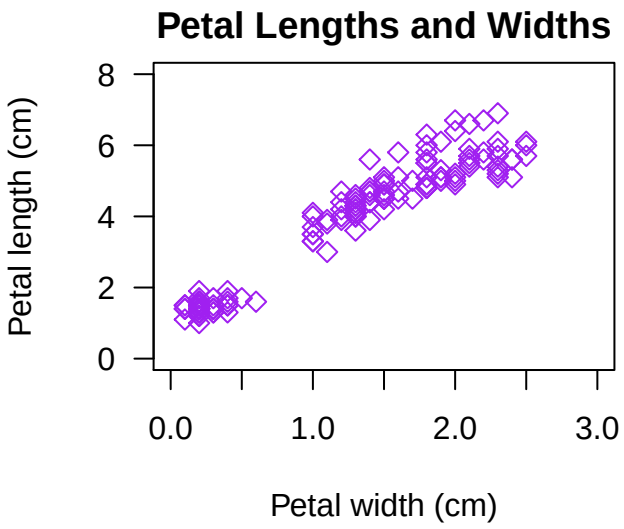


Figure 5.2: Illustration of a scatter plot with customised parameters in R.

Note: In many (most) applications it would not be necessary to have both a figure title and a caption label. In these guides both are used as it is necessary to describe both the figure and the **R** features that are covered in each example.

5.3 Scatter Plots with a Grouping Variable

From our scatter plot above we can see that petal length increases with petal width, but why is there a gap in the apparent linear relationship we see? Could this gap be explained by another variable in our data such as species? By differentiating the points by a third variable we can more clearly see what is going on in our data. Here we will differentiate points for each species by colour and in the next section we will make them different symbols and colours.

We don't use any new parameters or functions to create our plot. We just define our `col` as the variable we would like points to be differentiated by. The colours that are used can be specified, but here we will let **R** automatically choose the first 1 to 3 colours since we have three species.

The last thing we will do is add a legend. We did this in the last section of the histograms guide and here the legend is specified in almost the same format. We add only a single new parameter, `bty`. With `bty` we can specify the Box TYpe around the legend, or remove it altogether as we will do here. We first confirm the order of our species and then list them in this order for the parameter `legend`. In the legend `col` is set as 1:3 because in the plot we let **R** select colours for use, and **R** selected for first three colours. The colon here indicates the start and end points for our data range, where for `ylim` and `xlim` we are providing a list of the min and max.

```r
with(iris,plot(Petal.Length ~ Petal.Width,
     col= Species,    # colour points by variable 'Species'
     pch= 16,         # 16 = filled circle (easier to see differences)
     main= "Petal Lengths and Widths by Species",
     ylab= "Petal length (cm)",
     xlab= "Petal width (cm)",
     ylim= c(0,8),
     xlim= c(0,3),
     las= 1))

# check order of species to make sure we get the legend details correct
summary(iris$Species) # gives summary for only Species in iris data

##     setosa versicolor  virginica
##         50         50         50

# create legend
legend("bottomright",                          # put in bottom right corner
       title= "Species",                       # add a title
       legend= c("setosa","versicolor","virginica"), # specify species order
       col= 1:3,                               # add range of colours
       pch= 16,                                # use same symbol as in plot
       bty= "n")                               # remove legend box outline
```

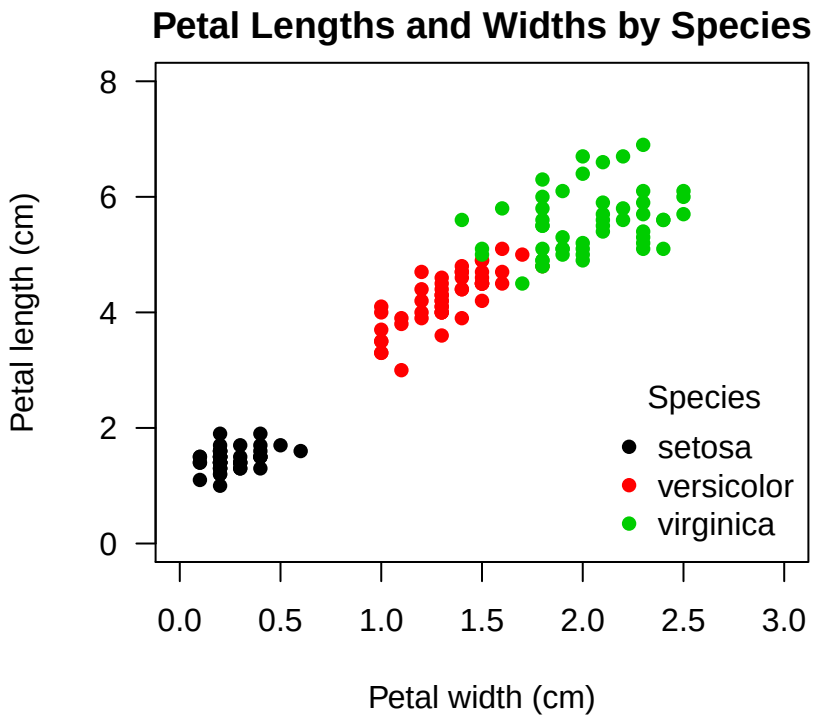


Figure 5.3: Illustration of a scatter plot with coloured grouping variable in R.

5.4 Advanced Scatter Plot Features

Now, what if you were submitting an academic paper that needed to be in black and white? Or, what if you were presenting your results and using colour could really help convey your message? For example, what if each of these species was a consistent different shade of purple and this trait explained why some iris species' petals grow larger? Sometimes it is necessary to specify the colours used for groups, or to use specific symbols to show differentiation instead of colour.

When specifying `col` and `pch` in this way there are two things you must include: (i) the list of colours or symbols to use; and (ii) the variable in your data to apply this 'list' to, which must be specified as a numerical variable (this doesn't mean it must be numbers). The format for `col` and `pch` will then look something like this:

```
c('col1','col2','col3')[as.numeric(group_variable)]

c(pt1,pt2,pt3)[as.numeric(group_variable)]
```

The only difference is that colour names must be enclosed in either single 'quote marks' or double "quote marks." Notice how we have (parentheses) around the list of colours/symbols and [brackets] around the function defining which variable to apply the list to. Let's try

creating two plots using this new format. First we will create a plot where we specify the colours as different shades of purple:

```
#1. scatter plot with specified colours for grouping variable
with(iris,plot(Petal.Length ~ Petal.Width,
    # specify colours and tell R to apply them to the variable 'Species'
     col= c("darkorchid1","darkorchid","darkorchid4")[as.numeric(Species)],
     pch= 19))                                              # define symbol
# add legend
legend("topleft",
       title= "Species",
       legend= c("setosa","versicolor","virginica"),      # list species' order
       col= c("darkorchid1","darkorchid","darkorchid4"), # list colours
       pch= 19,                                           # define symbol
       bty="n")                                           # remove outline
```

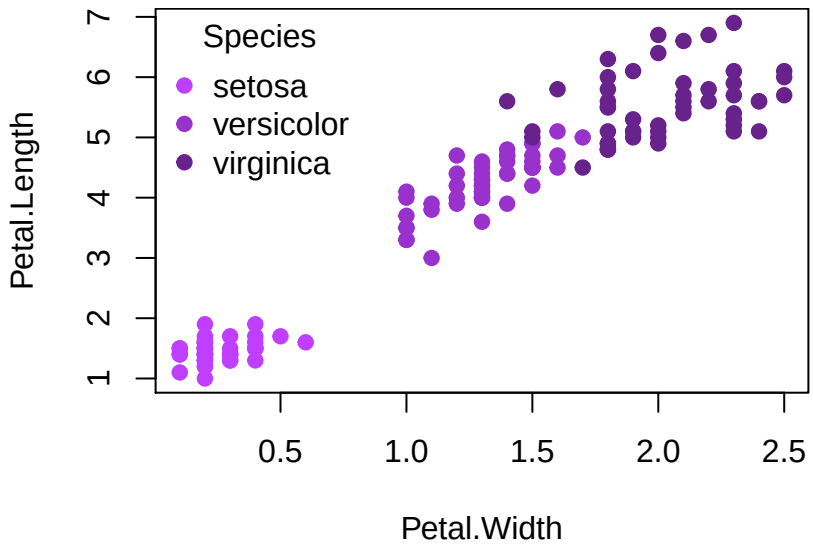


Figure 5.4: Illustration of a scatter plot with specified colours for grouping variable in R.

Now, let's create a plot that defines shades of grey (good for black and while publications) as well as different symbols to represent species. For this plot, let's also use other optional parameters we have learnt about:

```
#2. scatter plot with specified colours and symbols for grouping variable
with(iris,plot(Petal.Length ~ Petal.Width,
     # specify colours and tell R to apply them to the variable 'Species'
     col= c("grey10","grey50","grey80")[as.numeric(Species)],
     # specify symbols and tell R to apply them to the variable 'Species'
     pch= c(6,19,21)[as.numeric(Species)],
```

```
     main= "Petal Lengths and Widths",
     ylab= "Petal length (cm)",
     xlab= "Petal width (cm)",
     ylim= c(0,8),
     xlim= c(0,3),
     las= 1))
# add legend
legend("bottomright",
       title= "Species",
      legend= c("setosa","versicolor","virginica"),    # specify species' order
       col= c("grey10","grey50","grey80"),               # specify colours
       pch= c(6,19,21),                                  # specify symbols
       bty="n")                                          # remove outline
```

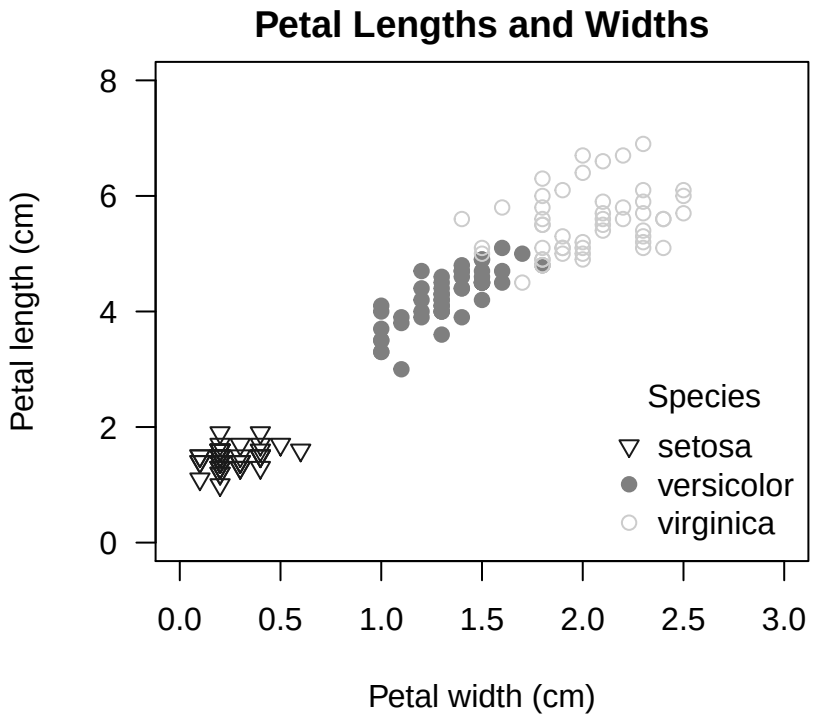


Figure 5.5: Illustration of a scatter plot with specified symbols and colours for grouping variable in R.

It is hard to think of a creating a plot with greater fine level control of the output!

6 Graphical Parameters

One of the beneficial things about using **R** is that it allows you flexibility when creating figures. With R you can easily customise plots using a seemingly endless array of functions and parameters. This guide will cover how to specify **colour**, **symbols** and **line types**, and how to save plots. Many other graphical parameters and functions are covered within other guides for this unit.

6.1 Defining Colours

Colours can be defined for a number of different attributes in R (e.g. points, lines and text). The standard parameter for most functions with colour capabilities is `col`, and this can be defined by a number or name, or a vector of either of these using the function concatenate *c()*. When defining a vector, colours are separated by a comma and enclosed in (parentheses). If colours are defined as names, they must be each enclosed in "double quotes". The default colour is almost always black, and the standard colour numbers run from 1 to 8 (see Figure 6.1). Outside of these colours there are hundreds more available - more easily by name - which can be found by executing this command in your R console: `colours()`. You can also find colour charts online by searching 'R colours'. The following are just a few examples of how `col` can be defined in R:

```
# specify colour by number
  col= 1,           # black
  col= c(1,2,3),    # vector of colours 1,2,3
  col= c(1:3),      # range of colours 1 to 3 - same as above
# specify colour by name
  col= "grey",
  col= c("grey10","grey30","grey80"),
```

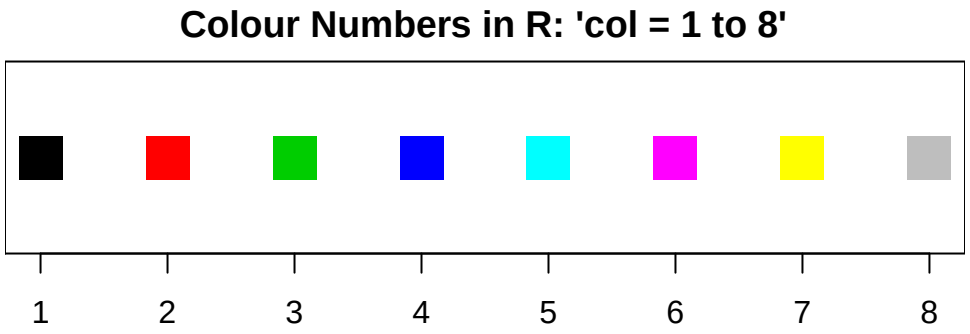


Figure 6.1: Plot showing standard colour numbers (1 to 8) in R.

6.2 Defining Symbols (Points)

Symbols can be defined to customize plot points, or differentiate points for different groups of a categorical variable (e.g. species or treatments). The parameter for point type is `pch` (Point CHange) and this can be defined as a number relating to one of the default symbols (1 to 25 - see Figure 6.2) or as a keyboard character or character string (i.e. group name). If points are defined as characters they must be enclosed in "double quotes". Points can also be given as a vector using the function `c()`. The size of your points can be changed with parameter `cex` by providing the proportion to increase or decrease from the default size. Lastly, while the colour can be changed for all points, default points 21 to 25 can have both the fill and outline defined using parameters `bg` and `col`. Here are a few examples:

```
# default point types by number
  pch= 1,              # circle outline
  pch= c(1,2,3),       # vector of points 1,2,3
  pch= c(1:3),         # range of points 1 to 3 - same as above
# setting points as characters
  pch= "A",
  pch= c("N","S","E","W"),
# point size (cex)
  pch= 1, cex= 0.5,          # circle outline, 50% the default size
  pch= c(1,2,3), cex= 1.2,   # points 1,2,3, all 20% larger than default
# add colour (col,bg)
  pch= 1, col= 4,            # blue circle outline
  pch= 21, col= 3, bg= 1,    # filled green circle with black outline
# combine parameters
# points 1,2,5 all 20% larger, and using colours 2 to 4
  pch= c(1,2,5), cex= 1.2, col= c(2:4),

# the most common situation for changing the points is for a scatter plot
# for example:
with(my.data,plot(Y.values ~ X.values, cex= 2, col=1, bg=8))
```

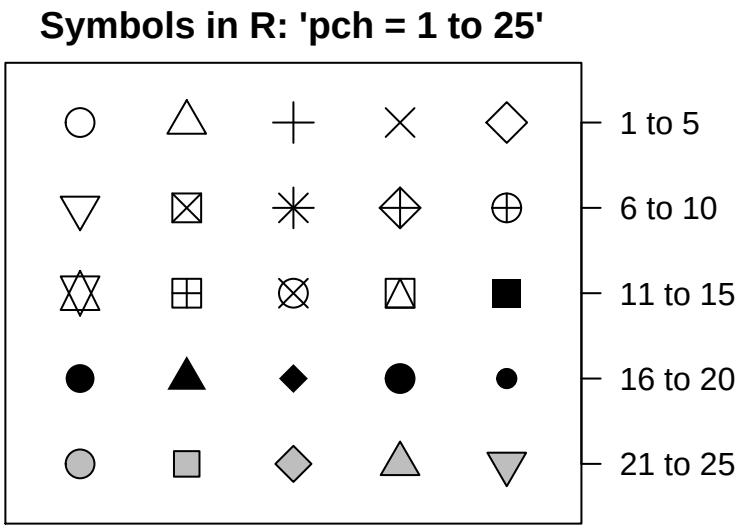


Figure 6.2: Plot showing available symbols by number in R. Note symbols 21 to 25 have background/fill defined as grey (bg = 8) and outline defined as black (col = 1)

6.3 Defining Line Types

Line types are defined in R by the parameter `lty` (Line TYpe) and there are 6 options to choose from (see Figure 6.3). You can also set the width of your line with parameter `lwd` (Line WiDth), with larger numbers giving thicker lines.

```
# line type
  lty= 1,              # solid line
  lty= c(1,5,3),       # list of lines 1,5,3
  lty= c(1:6),         # range of points 1 to 3 - same as above
# line width
  lty= 1, lwd= 0.5,          # solid lines, 50% default size
  lty= c(1,2,3), lwd= 1.2,   # lines 1,2,3, all 20% larger
# add colour
  lty= 1, col= 4,            # blue solid line
  lty= c(1,5), col= c(3,4), # green solid and blue dashed lines
# combine all parameters
# lines 1,2,3, all red, and using line widths 2 to 4
  lty= c(1,2,3), col= 2, lwd= c(2:4),
#  the most common situation for controlling the line type is adding
#  the trend line to a scatter plot.  For example:
abline(lm.1, col = 2, lty= 2, lwd= 2)
```

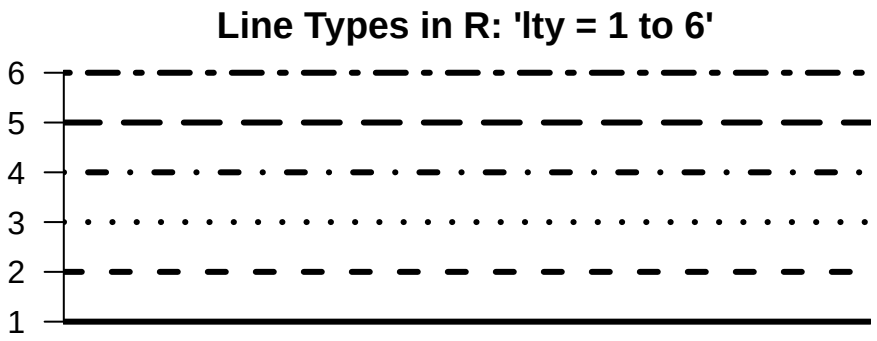


Figure 6.3: Plot showing available line types in R (1 to 6).

6.4 Saving Plots

The best and most efficient way to get high quality graphics from **R** is to save them from the command line. There are three steps to save a plot:

1. specify the file details - this starts a device which creates an empty file and waits for you to provide the contents (i.e. a plot). Here you can specify the following parameters:
    - file type: this determines the command you use, e.g. `png()`, `jpeg()`, `tiff()`, `bmp()`
    - file name: choose something intuitive; file name and extension must be included and enclosed in "double quotes" (e.g. "`My_file_name.png`")
    - directory: by default the plot will be saved to your working directory; if you prefer another, specify this with the file name using forward slashes between directories (Windows operating systems use backward slashes to indicate paths and R will not recognize this!)
    - figure size: specify this with optional parameters `width`, `height` and `units` ('mm', 'cm', 'in', default= pixels); since figures can easily be enlarged and compressed, e.g. when pasted into a MS Word document, this determines the plot size relative to its contents, axes and title. E.g. if your boxes seem squished in your box plot at a height of 70 mm, increase it to 100 mm. This will NOT change the actual axis value.
    - resolution: the default for parameter `res` is 72 pixels per inch (ppi), a good quality photograph generally requires 300. Set this if you require high quality figures (i.e. for a report/assignment), recognizing resolution is relevant for the size the plot is saved at.
2. create your plot as you would normally
3. close the device - this is the step that actually SAVES your plot

Here is an example of what this looks like in R:

```r
#1. Create an empty file in your working directory with your
#   specified parameters
png("Figure1.png", width = 100, height = 100, units = 'mm', res = 300)

#2. Create your plot
with(PlantGrowth,hist(weight,
      xlim= c(3,7),                        # set the x-axis range
      ylim= c(0,10),                       # set the y-axis range
      col= "lightgray",                    # fill the columns
      border= "black",                     # column border
      main= "Histogram of Plant Weights", # figure title
      xlab= "Dried weight (g)",            # x-axis label
      ylab= "Frequency"))                  # y-axis label

#3. Save your plot by turning device off
dev.off()
```

For information on which file type to choose, see the R help file: BMP, JPEG, PNG and TIFF graphics devices. In general, tiff files are larger but better quality figures. Also note that some 'viewer' applications do not support all file types (tiff are widely accepted). If when trying to view your saved plot you receive an error, try to open it with another application.

7 Probability

Probability forms a foundation for statistics. You might already be familiar with many aspects of probability, however, formalization of the concepts is new for most. This chapter aims to introduce probability on familiar terms using processes most people have seen before.

7.1 Defining Probability

Example 7.1 A "die", the singular of dice, is a cube with six faces numbered 1, 2, 3, 4, 5, and 6. What is the chance of getting 1 when rolling a die?

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

Example 7.2 What is the chance of getting a 1 or 2 in the next roll?

---

1 and 2 constitute two of the six equally likely possible outcomes, so the chance of getting one of these two outcomes must be $2/6 = 1/3$.

Example 7.3 What is the chance of getting either 1, 2, 3, 4, 5, or 6 on the next roll?

100%. The outcome must be one of these numbers.

Example 7.4 What is the chance of not rolling a 2?

Since the chance of rolling a 2 is 1/6 or $16.\bar{6}\%$, the chance of not rolling a 2 must be $100\% - 16.\bar{6}\% = 83.\bar{3}\%$ or 5/6.

Alternatively, we could have noticed that 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 7.5 Consider rolling two dice. If $1/6^{th}$ of the time the first die is a 1 and $1/6^{th}$ of those times the second die is a 1, what is the chance of getting two 1s?

If $16.\bar{6}\%$ 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 $(1/6) \times (1/6)$ or 1/36.

Probability

We use probability to build tools to describe and understand apparent randomness. We often frame probability in terms of a **random process** giving rise to an **outcome**.

Roll a die	$\rightarrow$	1, 2, 3, 4, 5, or 6
Flip a coin	$\rightarrow$	H or T

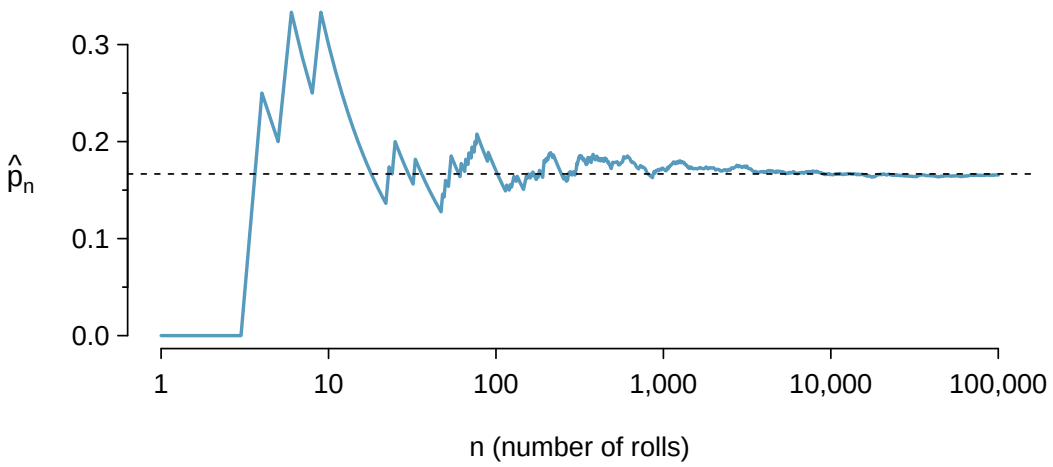


Figure 7.1: The fraction of die rolls that are 1 at each stage in a simulation. The proportion tends to get closer to the probability $1/6 \approx 0.167$ as the number of rolls increases.

Rolling a die or flipping a coin is a seemingly random process and each gives rise to an outcome.

Probability

The **probability** of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.

Probability is defined as a proportion, and it always takes values between 0 and 1 (inclusively). It may also be displayed as a percentage between 0% and 100%.

Probability can be illustrated by rolling a die 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 7.1 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**.

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.

Occasionally the proportion will veer off from the probability and appear to defy the Law of Large Numbers, as $\hat{p}_n$ does many times in Figure 7.1. However, these deviations become smaller as the number of rolls increases.

Above we write p as the probability of rolling a 1. We can also write this probability as

$$P(\text{rolling a 1})$$

As we become more comfortable with this notation, we will abbreviate it further. For

$P(A)$
Probability of outcome A

instance, if it is clear that the process is "rolling a die", we could abbreviate $P(\text{rolling a } 1)$ as $P(1)$.

⊙ **Guided Practice 7.6** Random processes include rolling a die and flipping a coin. (a) Think of another random process. (b) Describe all the possible outcomes of that process. For instance, rolling a die is a random process with possible outcomes 1, 2, ..., 6.[19]

Disjoint or mutually exclusive outcomes

Two outcomes are called **disjoint** or **mutually exclusive** if they cannot both happen. For instance, if we roll a die, the outcomes 1 and 2 are disjoint since they cannot both occur. On the other hand, the outcomes 1 and "rolling an odd number" are not disjoint since both occur if the outcome of the roll is a 1. The terms *disjoint* and *mutually exclusive* are equivalent and interchangeable.

Calculating the probability of disjoint outcomes is easy. When rolling a die, the outcomes 1 and 2 are disjoint, and we compute the probability that one of these outcomes will occur 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 we add the probabilities:

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

The **Addition Rule** guarantees the accuracy of this approach when the outcomes are disjoint.

[19] Here are four examples. (i) Whether someone gets sick in the next month or not is an apparently random process with outcomes `sick` and `not`. (ii) We can *generate* a random process by randomly picking a person and measuring that person's height. The outcome of this process will be a positive number. (iii) Whether the stock market goes up or down next week is a seemingly random process with possible outcomes `up`, `down`, and `no_change`. Alternatively, we could have used the percent change in the stock market as a numerical outcome. (iv) Whether your roommate cleans her dishes tonight probably seems like a random process with possible outcomes `cleans_dishes` and `leaves_dishes`.

Addition Rule of disjoint outcomes

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

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

If there are many disjoint outcomes A_1, ..., A_k, then the probability that one of these outcomes will occur is

$$P(A_1) + P(A_2) + \cdots + P(A_k) \tag{7}$$

⊙ **Guided Practice 7.8** We are interested in the probability of rolling a `1`, `4`, or `5`. (a) Explain why the outcomes `1`, `4`, and `5` are disjoint. (b) Apply the Addition Rule for disjoint outcomes to determine $P($`1` or `4` or `5`$)$.[20]

⊙ **Guided Practice 7.9** Load and inspect the `email` data set as follows.

```
install.packages("openintro")  #install the package containing the data
library(openintro)  #load the package that contains the data
help("email")  # inspect the help file describing the data
head(email)  # inspect the first few rows of the data
```

The `number` variable described whether no number (labeled `none`), only one or more small numbers (`small`), or whether at least one big number appeared in an email (`big`).

```
length(email$number)

## [1] 3921

table(email$number)

##
##  none small   big
##   549  2827   545
```

Of the 3,921 emails, 549 had no numbers, 2,827 had only one or more small numbers, and 545 had at least one big number. (a) Are the outcomes `none`, `small`, and `big` disjoint? (b) Determine the proportion of emails with value `small` and `big` separately.

[20] (a) The random process is a die roll, and at most one of these outcomes can come up. This means they are disjoint outcomes. (b) $P($`1` or `4` or `5`$) = P(1) + P(4) + P(5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$

(c) Use the Addition Rule for disjoint outcomes to compute the probability a randomly selected email from the data set has a number in it, small or big.[21]

Statisticians rarely work with individual outcomes and instead consider *sets* or *collections* of outcomes. Let A represent the event where 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 7.2.

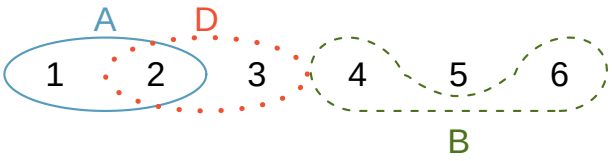


Figure 7.2: Three events, A, B, and D, consist of outcomes from rolling a die. A and B are disjoint since they do not have any outcomes in common.

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 7.10** (a) Verify the probability of event A, $P(A)$, is 1/3 using the Addition Rule. (b) Do the same for event B.[22]

⊙ **Guided Practice 7.11** (a) Using Figure 7.2 as a reference, what outcomes are represented by event D? (b) Are events B and D disjoint? (c) Are events A and D disjoint?[23]

⊙ **Guided Practice 7.12** In Guided Practice 11, you confirmed B and D from Figure 7.2 are disjoint. Compute the probability that event B or event D occurs.[24]

Probabilities when events are not disjoint

Let's consider calculations for two events that are not disjoint in the context of a regular deck of 52 cards, represented in Table 7.1. If you are unfamiliar with the cards in a regular deck, please see the footnote.[25]

[21](a) Yes. Each email is categorized in only one level of `number`. (b) Small: $\frac{2827}{3921} = 0.721$. Big: $\frac{545}{3921} = 0.139$. (c) $P(\texttt{small or big}) = P(\texttt{small}) + P(\texttt{big}) = 0.721 + 0.139 = 0.860$.

[22](a) $P(A) = P(\texttt{1 or 2}) = P(\texttt{1}) + P(\texttt{2}) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$. (b) Similarly, $P(B) = 1/3$.

[23](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.

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

[25]The 52 cards are split into four **suits**: ♣ (club), ♢ (diamond), ♡ (heart), ♠ (spade). Each suit has its 13 cards labeled: `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♣. The 12 cards represented by the jacks, queens, and kings are called **face cards**. The cards that are ♢ or ♡ are typically colored red while the other two suits are typically colored black.

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♠

Table 7.1: Representations of the 52 unique cards in a deck.

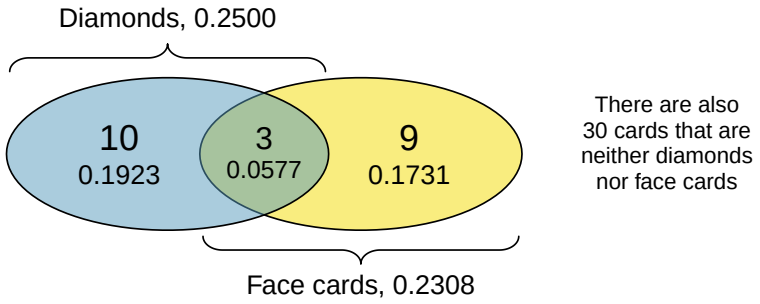


Figure 7.3: A Venn diagram for diamonds and face cards.

⊙ **Guided Practice 7.13** (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?[26]

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 7.3 uses a circle to represent diamonds and another to represent face cards. If a card is both a diamond and a face card, it falls into the intersection of the circles. If it is a diamond but not a face card, it will be in part of the left circle that is not in the right circle (and so on). The total number of cards that are diamonds is given by the total number of cards in the diamonds circle: $10 + 3 = 13$. The probabilities are also shown (e.g. $10/52 = 0.1923$).

Let A represent the event that a randomly selected card is a diamond and B represent the event that it is a face card. How do we compute $P(A \text{ or } B)$? Events A and B are not disjoint – the cards $J\diamondsuit$, $Q\diamondsuit$, and $K\diamondsuit$ fall into both categories – so we cannot use the Addition Rule for disjoint events. Instead we use the Venn diagram. We start by adding the probabilities of the two events:

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

However, the three cards that are in both events were counted twice, once in each probability.

[26] (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 $P(\diamondsuit) = \frac{13}{52} = 0.250$. (b) Likewise, there are 12 face cards, so $P(\text{face card}) = \frac{12}{52} = \frac{3}{13} = 0.231$.

We must correct this double counting:

$$\begin{aligned} P(A \text{ or } B) &= P(\diamondsuit \text{ or face card}) \\ &= P(\diamondsuit) + P(\text{face card}) - P(\diamondsuit \text{ and face card}) \\ &= 13/52 + 12/52 - 3/52 \\ &= 22/52 = 11/26 \end{aligned} \tag{14}$$

Equation (14) 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

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \tag{15}$$

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

TIP: "or" is inclusive

When we write "or" in statistics, we mean "and/or" unless we explicitly state otherwise. Thus, A or B occurs means A, B, or both A and B occur.

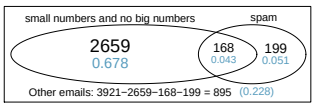
Guided Practice 7.16 (a) If A and B are disjoint, describe why this implies $P(A \text{ and } B) = 0$. (b) Using part (a), verify that the General Addition Rule simplifies to the simpler Addition Rule for disjoint events if A and B are disjoint.[27]

Guided Practice 7.17 In the `email` data set with 3,921 emails, 367 were spam, 2,827 contained some small numbers but no big numbers, and 168 had both characteristics. Create a Venn diagram for this setup.[28]

Guided Practice 7.18 (a) Use your Venn diagram from Guided Practice 17 to determine the probability a randomly drawn email from the `email` data set is spam and had small numbers (but not big numbers). (b) What is the probability that the email had either of these attributes?[29]

[27] (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 (15) is 0 (see part (a)) and we are left with the Addition Rule for disjoint events.

[28] Both the counts and corresponding probabilities (e.g. $2659/3921 = 0.678$) are shown. Notice that the number of emails represented in the left circle corresponds to $2659 + 168 = 2827$, and the number represented in the right circle is $168 + 199 = 367$.



[29] (a) The solution is represented by the intersection of the two circles: 0.043. (b) This is the sum of the three disjoint probabilities shown in the circles: $0.678 + 0.043 + 0.051 = 0.772$.

Probability distributions

A **probability distribution** is a table of all disjoint outcomes and their associated probabilities. Table 7.2 shows the probability distribution for the sum of two dice.

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

Table 7.2: Probability distribution for the sum of two dice.

Rules for probability distributions

A probability distribution is a list of the possible outcomes with corresponding 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 1.

⊙ **Guided Practice 7.19** Table 7.3 suggests three distributions for household income in Australia. Only one is correct. Which one must it be? What is wrong with the other two?[30]

Weekly income range ($1000s)	0-1	1-2	2-3	3-4	4-5	5+
(a)	0.33	0.28	0.19	0.10	0.04	0.05
(b)	0.33	-0.28	0.19	0.10	0.04	0.05
(c)	0.23	0.18	0.19	0.10	0.04	0.05

Table 7.3: Proposed distributions of Australian household incomes (Guided Practice 19).

Probability distributions can be summarized in a bar plot. For instance, the distribution of Australian household incomes [31]is shown in Figure 7.4 as a bar plot. The probability distribution for the sum of two dice is shown in Table 7.2 and plotted in Figure 7.5.

In these bar plots, the bar heights represent the probabilities of outcomes. If the outcomes are numerical and discrete, it is usually (visually) convenient to make a bar plot that resembles a histogram, as in the case of the sum of two dice.

[30]The probabilities of (c) do not sum to 1. The second probability in (b) is negative. This leaves (a), which sure enough satisfies the requirements of a distribution. One of the three was said to be the actual distribution of Australian household incomes, so it must be (a).

[31]http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/6523.02013-14?OpenDocument

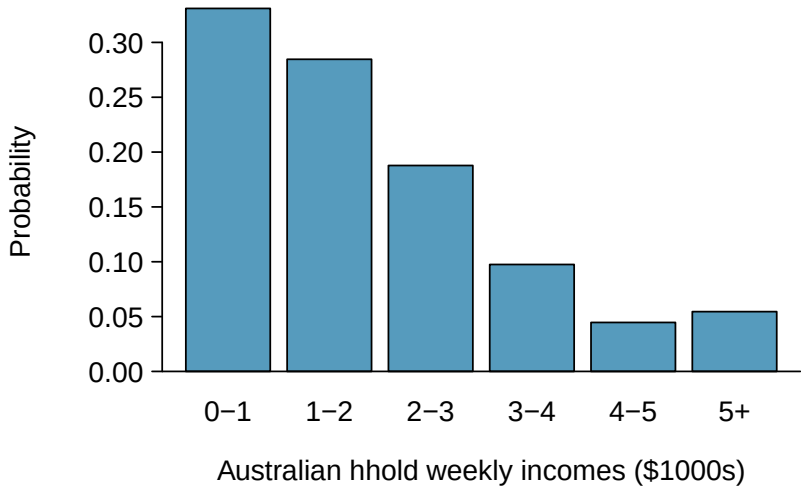


Figure 7.4: The probability distribution of Australian household income.

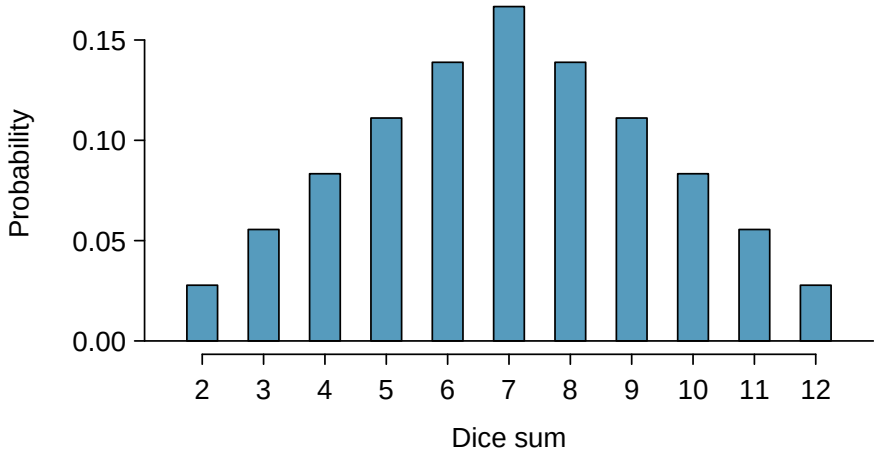


Figure 7.5: The probability distribution of the sum of two dice.

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. We often use the sample space to examine the scenario where an event does not occur.

S
Sample space

Let $D = \{2, 3\}$ represent the event that the outcome of a die roll is 2 or 3. Then the **complement** of D represents all outcomes in our 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 7.6 shows the relationship between D, D^c, and the sample space S.

A^c
Complement of outcome A

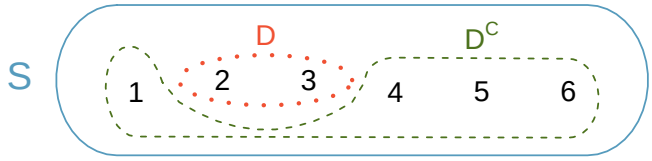


Figure 7.6: Event $D = \{2, 3\}$ and its complement, $D^c = \{1, 4, 5, 6\}$. S represents the sample space, which is the set of all possible events.

Guided Practice 7.20 (a) Compute $P(D^c) = P(\text{rolling a } 1, 4, 5, \text{ or } 6)$. (b) What is $P(D) + P(D^c)$?[32]

Guided Practice 7.21 Events $A = \{1, 2\}$ and $B = \{4, 6\}$ are shown in Figure 7.2 on page 49. (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 $P(B) + P(B^c)$.[33]

A complement of an event A is constructed to have two very important properties: (i) every possible outcome not in A is in A^c, and (ii) A and A^c are disjoint. Property (i) implies

$$P(A \text{ or } A^c) = 1 \tag{22}$$

That is, if the outcome is not in A, it must be represented in A^c. We use the Addition Rule for disjoint events to apply Property (ii):

$$P(A \text{ or } A^c) = P(A) + P(A^c) \tag{23}$$

Combining Equations (22) and (23) yields a very useful relationship between the probability of an event and its complement.

[32] (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, $P(D) + P(D^c) = 1$.

[33] Brief 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:

$$P(A) + P(A^c) = 1, \quad \text{i.e.} \quad P(A) = 1 - P(A^c) \tag{24}$$

In simple examples, computing A or A^c is feasible in a few steps. However, using the complement can save a lot of time as problems grow in complexity.

⊙ **Guided Practice 7.25** Let A represent the event where we roll two dice and their total is less than 12. (a) What does the event A^c represent? (b) Determine $P(A^c)$ from Table 7.2 on page 52. (c) Determine $P(A)$.[34]

⊙ **Guided Practice 7.26** Consider again the probabilities from Table 7.2 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 $B = \{4, 5, ..., 12\}$. (c) The sum is no more than 10. That is, determine the probability of the event $D = \{2, 3, ..., 10\}$.[35]

Independence

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

Example 5 provides a basic example of two independent processes: rolling two dice. We want to determine the probability that both will be 1. Suppose one of the dice is red and the other white. If the outcome of the red die is a 1, it provides no information about the outcome of the white die. We first encountered this same question in Example 5 (page 45), where we calculated the probability using the following reasoning: $1/6^{th}$ of the time the red die is a 1, and $1/6^{th}$ of *those* times the white die will also be 1. This is illustrated in Figure 7.7. Because the rolls are independent, the probabilities of the corresponding outcomes can be multiplied to get the final answer: $(1/6) \times (1/6) = 1/36$. This can be generalized to many independent processes.

[34](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 (24): $P(\text{less than } 12) = 1 - P(12) = 1 - 1/36 = 35/36$.

[35](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 clever way to determine $P(D)$. First find $P(D^c) = P(11 \text{ or } 12) = 2/36 + 1/36 = 1/12$. Then calculate $P(D) = 1 - P(D^c) = 11/12$.

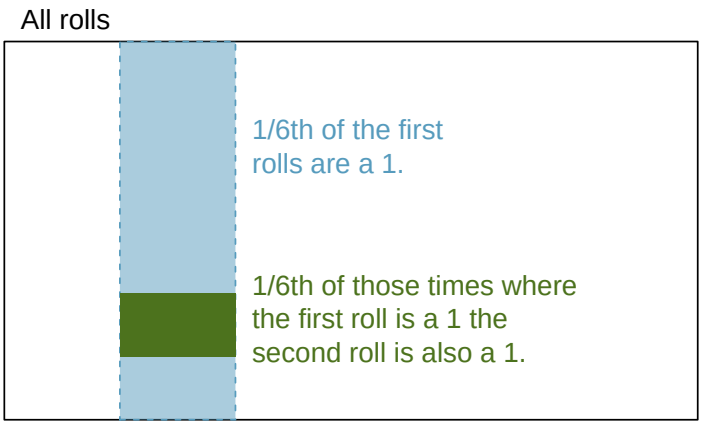


Figure 7.7: $1/6^{th}$ of the time, the first roll is a 1. Then $1/6^{th}$ of *those* times, the second roll will also be a 1.

Example 7.27 What if there was also a blue die independent of the other two? What is the probability of rolling the three dice and getting all 1s?

The same logic applies from Example 5. If $1/36^{th}$ of the time the white and red dice are both 1, then $1/6^{th}$ of *those* times the blue die will also be 1, so multiply:

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

Example 27 illustrates what is called the Multiplication Rule for independent processes.

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 can be calculated as the product of their separate probabilities:

$$P(A \text{ and } B) = P(A) \times P(B) \tag{28}$$

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

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

Sometimes we wonder if one outcome provides useful information about another outcome. The question we are asking is, are the occurrences of the two events independent? We say that two events A and B are independent if they satisfy Equation (28).

Example 7.29 If we shuffle up a deck of cards and draw one, is the event that the card is a heart independent of the event that the card is an ace?

The probability the card is a heart is 1/4 and the probability that it is an ace is 1/13. The probability the card is the ace of hearts is 1/52. We check whether Equation 28 is satisfied:

$$P(\heartsuit) \times P(\text{ace}) = \frac{1}{4} \times \frac{1}{13} = \frac{1}{52} = P(\heartsuit \text{ and ace})$$

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

7.2 Conditional Probability

The `family_college` data set contains a sample of 792 cases with two variables, `teen` and `parents`, and is summarized in Table 7.4.[36] The `teen` variable is either `college` or `not`, where the `college` label means the teen went to college immediately after high school. The `parents` variable takes the value `degree` if at least one parent of the teenager completed a college degree.

		parents		
		degree	not	Total
teen	college	231	214	445
	not	49	298	347
	Total	280	512	792

Table 7.4: Contingency table summarizing the `family_college` data set.

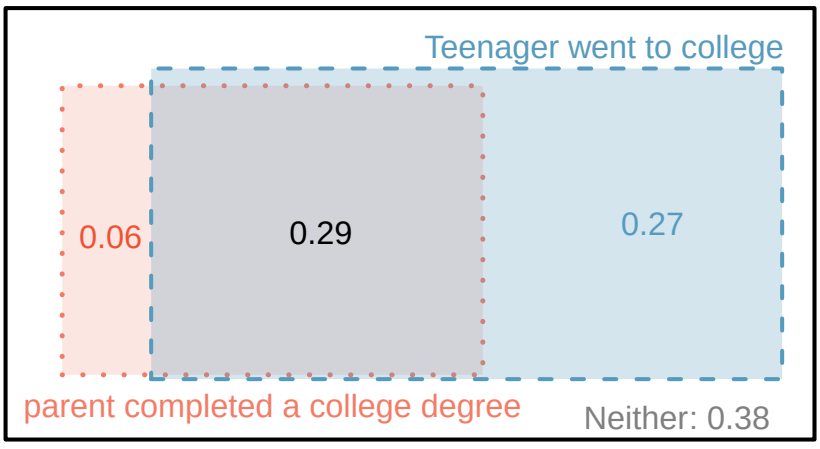


Figure 7.8: A Venn diagram using boxes for the `family_college` data set.

Example 7.30 If at least one parent of a teenager completed a college degree, what is the chance the teenager attended college right after high school?

[36] A simulated data set based on real population summaries at nces.ed.gov/pubs2001/2001126.pdf.

We can estimate this probability using the data. Of the 280 cases in this data set where `parents` takes value `degree`, 231 represent cases where the `teen` variable takes value `college`:

$$P(\texttt{teen college given parents degree}) = \frac{231}{280} = 0.825$$

Example 7.31 A teenager is randomly selected from the sample and she did not attend college right after high school. What is the probability that at least one of her parents has a college degree?

If the teenager did not attend, then she is one of the 347 teens in the second row. Of these 347 teens, 49 had at least one parent who got a college degree:

$$P(\texttt{parents degree given teen not}) = \frac{49}{347} = 0.141$$

Marginal and joint probabilities

Table 7.4 includes row and column totals for each variable separately in the `family_college` data set. These totals represent **marginal probabilities** for the sample, which are the probabilities based on a single variable without regard to any other variables. For instance, a probability based solely on the `teen` variable is a marginal probability:

$$P(\texttt{teen college}) = \frac{445}{792} = 0.56$$

A probability of outcomes for two or more variables or processes is called a **joint probability**:

$$P(\texttt{teen college and parents not}) = \frac{214}{792} = 0.27$$

It is common to substitute a comma for "and" in a joint probability, although either is acceptable. That is,

$$P(\texttt{teen college, parents not})$$

means the same thing as

$$P(\texttt{teen college and parents not})$$

Marginal and joint probabilities

If a probability is based on a single variable, it is a *marginal probability*. The probability of outcomes for two or more variables or processes is called a *joint probability*.

We use **table proportions** to summarize joint probabilities for the `family_college` sample. These proportions are computed by dividing each count in Table 7.4 by the table's

total, 792, to obtain the proportions in Table 7.5. The joint probability distribution of the `parents` and `teen` variables is shown in Table 7.6.

	`parents: degree`	`parents: not`	Total
`teen: college`	0.29	0.27	0.56
`teen: not`	0.06	0.38	0.44
Total	0.35	0.65	1.00

Table 7.5: Probability table summarizing whether at least one parent had a college degree and the teenager attended college.

Joint outcome	Probability
`parents degree` and `teen college`	0.29
`parents degree` and `teen not`	0.06
`parents not` and `teen college`	0.27
`parents not` and `teen not`	0.38
Total	1.00

Table 7.6: Joint probability distribution for the `family_college` data set.

⊙ **Guided Practice 7.32** Verify Table 7.6 represents a probability distribution: events are disjoint, all probabilities are non-negative, and the probabilities sum to 1.[37]

We can compute marginal probabilities using joint probabilities in simple cases. For example, the probability a random teenager from the study went to college is found by summing the outcomes where `teen` takes value `college`:

$$\begin{aligned} P(\underline{\texttt{teen college}}) &= P(\texttt{parents degree and } \underline{\texttt{teen college}}) \\ &\quad + P(\texttt{parents not and } \underline{\texttt{teen college}}) \\ &= 0.29 + 0.27 \\ &= 0.56 \end{aligned}$$

Defining conditional probability

There is some connection between education level of parents and of the teenager: a college degree by a parent is associated with college attendance of the teenager. In this section, we discuss how to use information about associations between two variables to improve probability estimation.

The probability that a random teenager from the study attended college is 0.56. Could we update this probability if we knew that one of the teen's parents has a college degree?

[37] Each of the four outcome combination are disjoint, all probabilities are indeed non-negative, and the sum of the probabilities is $0.29 + 0.06 + 0.27 + 0.38 = 1.00$.

Absolutely. To do so, we limit our view to only those 280 cases where a parent has a college degree and look at the fraction where the teenager attended college:

$$P(\texttt{teen college} \text{ given } \texttt{parents degree}) = \frac{231}{280} = 0.825$$

We call this a **conditional probability** because we computed the probability under a condition: a parent has a college degree. There are two parts to a conditional probability, the **outcome of interest** and the **condition**. It is useful to think of the condition as information we know to be true, and this information usually can be described as a known outcome or event.

We separate the text inside our probability notation into the outcome of interest and the condition:

$$\begin{aligned} &P(\texttt{teen college} \text{ given } \texttt{parents degree}) \\ &= P(\texttt{teen college} \mid \texttt{parents degree}) = \frac{231}{280} = 0.825 \end{aligned} \tag{33}$$

$P(A|B)$
Probability of outcome A given B

The vertical bar "|" is read as *given*.

In Equation (33), we computed the probability a teen attended college based on the condition that at least one parent has a college degree as a fraction:

$$\begin{aligned} &P(\texttt{teen college} \mid \texttt{parents degree}) \\ &\quad = \frac{\#\text{ cases where } \texttt{teen college} \text{ and } \texttt{parents degree}}{\#\text{ cases where } \texttt{parents degree}} \\ &\quad = \frac{231}{280} = 0.825 \end{aligned} \tag{34}$$

We considered only those cases that met the condition, `parents degree`, and then we computed the ratio of those cases that satisfied our outcome of interest, the teenager attended college.

Frequently, marginal and joint probabilities are provided instead of count data. For example, disease rates are commonly listed in percentages rather than in a count format. We would like to be able to compute conditional probabilities even when no counts are available, and we use Equation (34) as a template to understand this technique.

We considered only those cases that satisfied the condition, `parents degree`. Of these cases, the conditional probability was the fraction who represented the outcome of interest, `teen college`. Suppose we were provided only the information in Table 7.5, i.e. only probability data. Then if we took a sample of 1000 people, we would anticipate about 35% or $0.35 \times 1000 = 350$ would meet the information criterion (`parents degree`). Similarly, we would expect about 29% or $0.29 \times 1000 = 290$ to meet both the information criteria and

represent our outcome of interest. Then the conditional probability can be computed as

$$\begin{aligned} &P(\texttt{teen college} \mid \texttt{parents degree}) \\ &= \frac{\#\ (\texttt{teen college and parents degree})}{\#\ (\texttt{parents degree})} \\ &= \frac{290}{350} = \frac{0.29}{0.35} = 0.829 \quad \text{(different from 0.825 due to rounding error)} \end{aligned} \tag{35}$$

In Equation (35), we examine exactly the fraction of two probabilities, 0.29 and 0.35, which we can write as

$$P(\texttt{teen college and parents degree}) \quad \text{and} \quad P(\texttt{parents degree}).$$

The fraction of these probabilities is an example of the general formula for conditional probability.

Conditional probability

The conditional probability of the outcome of interest A given condition B is computed as the following:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \tag{36}$$

⊙ **Guided Practice 7.37** (a) Write out the following statement in conditional probability notation: "*The probability a random case where neither parent has a college degree if it is known that the teenager didn't attend college right after high school*". Notice that the condition is now based on the teenager, not the parent.

(b) Determine the probability from part (a). Table 7.5 on page 59 may be helpful.[38]

[38] (a) $P(\texttt{parents not} \mid \texttt{teen not})$. (b) Equation (36) for conditional probability indicates we should first find $P(\texttt{parents not and teen not}) = 0.38$ and $P(\texttt{teen not}) = 0.44$. Then the ratio represents the conditional probability: $0.38/0.44 = 0.864$.

Guided Practice 7.38 (a) Determine the probability that one of the parents has a college degree if it is known the teenager did not attend college.

(b) Using the answers from part (a) and Guided Practice 37(b), compute

$$P(\texttt{parents degree} \mid \texttt{teen not}) \ + \ P(\texttt{parents not} \mid \texttt{teen not})$$

(c) Provide an intuitive argument to explain why the sum in (b) is 1.[39]

Guided Practice 7.39 The data indicate there is an association between parents having a college degree and their teenager attending college. Does this mean the parents' college degree(s) *caused* the teenager to go to college?[40]

Smallpox in Boston, 1721

The `smallpox` data set provides a sample of 6,224 individuals from the year 1721 who were exposed to smallpox in Boston.[41] Doctors at the time believed that inoculation, which involves exposing a person to the disease in a controlled form, could reduce the likelihood of death.

Each case represents one person with two variables: `inoculated` and `result`. The variable `inoculated` takes two levels: `yes` or `no`, indicating whether the person was inoculated or not. The variable `result` has outcomes `lived` or `died`. These data are summarized in Tables 7.7 and 7.8.

		inoculated		
		yes	no	Total
result	lived	238	5136	5374
	died	6	844	850
	Total	244	5980	6224

Table 7.7: Contingency table for the `smallpox` data set.

		inoculated		
		yes	no	Total
result	lived	0.0382	0.8252	0.8634
	died	0.0010	0.1356	0.1366
	Total	0.0392	0.9608	1.0000

Table 7.8: Table proportions for the `smallpox` data, computed by dividing each count by the table total, 6224.

[39](a) This probability is $\frac{P(\texttt{parents degree, teen not})}{P(\texttt{teen not})} = \frac{0.06}{0.44} = 0.136$. (b) The total equals 1. (c) Under the condition the teenager didn't attend college, the parents must either have a college degree or not. The complement still works for conditional probabilities, provided the probabilities are conditioned on the same information.

[40]No. While there is an association, the data are observational. Two potential confounding variables include `income` and `region`. Can you think of others?

[41]Fenner F. 1988. *Smallpox and Its Eradication (History of International Public Health, No. 6)*. Geneva: World Health Organization. ISBN 92-4-156110-6.

Guided Practice 7.40 Write out, in formal notation, the probability a randomly selected person who was not inoculated died from smallpox, and find this probability.[42]

Guided Practice 7.41 Determine the probability that an inoculated person died from smallpox. How does this result compare with the result of Guided Practice 40?[43]

Guided Practice 7.42 The people of Boston self-selected whether or not to be inoculated. (a) Is this study observational or was this an experiment? (b) Can we infer any causal connection using these data? (c) What are some potential confounding variables that might influence whether someone `lived` or `died` and also affect whether that person was inoculated?[44]

General multiplication rule

Section 7.1 introduced the Multiplication Rule for independent processes. Here we provide the **General Multiplication Rule** for events that might not be independent.

General Multiplication Rule

If A and B represent two outcomes or events, then

$$P(A \text{ and } B) = P(A|B) \times P(B)$$

It is useful to think of A as the outcome of interest and B as the condition.

This General Multiplication Rule is simply a rearrangement of the definition for conditional probability in Equation (36) on page 61.

Example 7.43 Consider the `smallpox` data set. Suppose we are given only two pieces of information: 96.08% of residents were not inoculated, and 85.88% of the residents who were not inoculated ended up surviving. How could we compute the probability that a resident was not inoculated and lived?

We will compute our answer using the General Multiplication Rule and then verify it using Table 7.8. We want to determine

$$P(\texttt{result} = \texttt{lived} \text{ and } \texttt{inoculated} = \texttt{no})$$

---

[42] $P(\texttt{result} = \texttt{died} \mid \texttt{inoculated} = \texttt{no}) = \frac{P(\texttt{result} = \texttt{died and inoculated} = \texttt{no})}{P(\texttt{inoculated} = \texttt{no})} = \frac{0.1356}{0.9608} = 0.1411.$

[43] $P(\texttt{result} = \texttt{died} \mid \texttt{inoculated} = \texttt{yes}) = \frac{P(\texttt{result} = \texttt{died and inoculated} = \texttt{yes})}{P(\texttt{inoculated} = \texttt{yes})} = \frac{0.0010}{0.0392} = 0.0255$. The death rate for individuals who were inoculated is only about 1 in 40 while the death rate is about 1 in 7 for those who were not inoculated.

[44] Brief answers: (a) Observational. (b) No, we cannot infer causation from this observational study. (c) Accessibility to the latest and best medical care. There are other valid answers for part (c).

and we are given that

$$P(\texttt{result} = \texttt{lived} \mid \texttt{inoculated} = \texttt{no}) = 0.8588$$
$$P(\texttt{inoculated} = \texttt{no}) = 0.9608$$

Among the 96.08% of people who were not inoculated, 85.88% survived:

$$P(\texttt{result} = \texttt{lived} \text{ and } \texttt{inoculated} = \texttt{no}) = 0.8588 \times 0.9608 = 0.8251$$

This is equivalent to the General Multiplication Rule. We can confirm this probability in Table 7.8 at the intersection of `no` and `lived` (with a small rounding error).

⊙ **Guided Practice 7.44** Use $P(\texttt{inoculated} = \texttt{yes}) = 0.0392$ and $P(\texttt{result} = \texttt{lived} \mid \texttt{inoculated} = \texttt{yes}) = 0.9754$ to determine the probability that a person was both inoculated and lived.[45]

⊙ **Guided Practice 7.45** If 97.54% of the people who were inoculated lived, what proportion of inoculated people must have died?[46]

Sum of conditional probabilities

Let A_1, ..., A_k represent all the disjoint outcomes for a variable or process. Then if B is an event, possibly for another variable or process, we have:

$$P(A_1|B) + \cdots + P(A_k|B) = 1$$

The rule for complements also holds when an event and its complement are conditioned on the same information:

$$P(A|B) = 1 - P(A^c|B)$$

⊙ **Guided Practice 7.46** Based on the probabilities computed above, does it appear that inoculation is effective at reducing the risk of death from smallpox?[47]

Independence considerations in conditional probability

If two events are independent, then knowing the outcome of one should provide no information about the other. We can show this is mathematically true using conditional probabilities.

[45] The answer is 0.0382, which can be verified using Table 7.8.

[46] There were only two possible outcomes: `lived` or `died`. This means that 100% - 97.45% = 2.55% of the people who were inoculated died.

[47] The samples are large relative to the difference in death rates for the "inoculated" and "not inoculated" groups, so it seems there is an association between `inoculated` and `outcome`. However, as noted in the solution to Guided Practice 42, this is an observational study and we cannot be sure if there is a causal connection. (Further research has shown that inoculation is effective at reducing death rates.)

⊙ **Guided Practice 7.47** Let X and Y represent the outcomes of rolling two dice.[48]

(a) What is the probability that the first die, X, is 1?

(b) What is the probability that both X and Y are 1?

(c) Use the formula for conditional probability to compute $P(Y = 1 \mid X = 1)$.

(d) What is $P(Y = 1)$? Is this different from the answer from part (c)? Explain.

[48] Brief solutions: (a) 1/6. (b) 1/36. (c) $\frac{P(Y=1 \text{ and } X=1)}{P(X=1)} = \frac{1/36}{1/6} = 1/6$. (d) The probability is the same as in part (c): $P(Y = 1) = 1/6$. The probability that $Y = 1$ was unchanged by knowledge about X, which makes sense as X and Y are independent.

We can show in Guided Practice 47(c) that the conditioning information has no influence by using the Multiplication Rule for independence processes:

$$\begin{aligned} P(Y = 1 \mid X = 1) &= \frac{P(Y = 1 \text{ and } X = 1)}{P(X = 1)} \\ &= \frac{P(Y = 1) \times P(X = 1)}{P(X = 1)} \\ &= P(Y = 1) \end{aligned}$$

⊙ **Guided Practice 7.48** Ron is watching a roulette table in a casino and notices that the last five outcomes were `black`. He figures that the chances of getting `black` six times in a row is very small (about 1/64) and puts his paycheck on red. What is wrong with his reasoning?[49]

Tree diagrams

Tree diagrams are a tool to organize outcomes and probabilities around the structure of the data. They are most useful when two or more processes occur in a sequence and each process is conditioned on its predecessors.

The `smallpox` data fit this description. We see the population as split by `inoculation`: `yes` and `no`. Following this split, survival rates were observed for each group. This structure is reflected in the **tree diagram** shown in Figure 7.9. The first branch for `inoculation` is said to be the **primary** branch while the other branches are **secondary**.

Tree diagrams are annotated with marginal and conditional probabilities, as shown in Figure 7.9. This tree diagram splits the smallpox data by `inoculation` into the `yes` and `no` groups with respective marginal probabilities 0.0392 and 0.9608. The secondary branches are conditioned on the first, so we assign conditional probabilities to these branches. For example, the top branch in Figure 7.9 is the probability that `result` = `lived` conditioned on the information that `inoculated` = `yes`. We may (and usually do) construct joint probabilities at the end of each branch in our tree by multiplying the numbers we come across as we move from left to right. These joint probabilities are computed using the General Multiplication Rule:

$$\begin{aligned} &P(\texttt{inoculated} = \texttt{yes} \text{ and } \texttt{result} = \texttt{lived}) \\ &\quad = P(\texttt{inoculated} = \texttt{yes}) \times P(\texttt{result} = \texttt{lived} | \texttt{inoculated} = \texttt{yes}) \\ &\quad = 0.0392 \times 0.9754 = 0.0382 \end{aligned}$$

● **Example 7.49** Consider the midterm and final for a statistics class. Suppose 13% of students earned an `A` on the midterm. Of those students who earned an `A` on the midterm, 47% received an `A` on the final, and 11% of the students who earned lower than an `A` on the midterm received an `A` on the final. You randomly pick up a final

[49] He has forgotten that the next roulette spin is independent of the previous spins. Casinos do employ this practice; they post the last several outcomes of many betting games to trick unsuspecting gamblers into believing the odds are in their favor. This is called the **gambler's fallacy**.

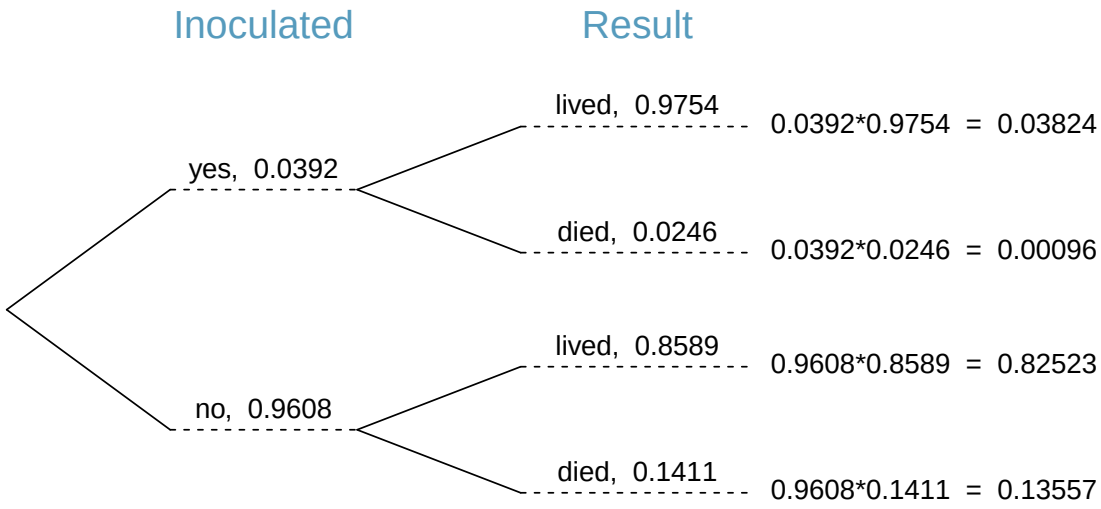


Figure 7.9: A tree diagram of the `smallpox` data set.

exam and notice the student received an `A`. What is the probability that this student earned an `A` on the midterm?

The end-goal is to find $P(\texttt{midterm} = \texttt{A} | \texttt{final} = \texttt{A})$. To calculate this conditional probability, we need the following probabilities:

$$P(\texttt{midterm} = \texttt{A} \text{ and } \texttt{final} = \texttt{A}) \qquad \text{and} \qquad P(\texttt{final} = \texttt{A})$$

However, this information is not provided, and it is not obvious how to calculate these probabilities. Since we aren't sure how to proceed, it is useful to organize the information into a tree diagram, as shown in Figure 7.10. When constructing a tree diagram, variables provided with marginal probabilities are often used to create the tree's primary branches; in this case, the marginal probabilities are provided for midterm grades. The final grades, which correspond to the conditional probabilities provided, will be shown on the secondary branches.

With the tree diagram constructed, we may compute the required probabilities:

$$\begin{aligned} &P(\texttt{midterm} = \texttt{A} \text{ and } \texttt{final} = \texttt{A}) = 0.0611 \\ &P(\underline{\texttt{final} = \texttt{A}}) \\ &\quad = P(\texttt{midterm} = \texttt{other} \text{ and } \underline{\texttt{final} = \texttt{A}}) + P(\texttt{midterm} = \texttt{A} \text{ and } \underline{\texttt{final} = \texttt{A}}) \\ &\quad = 0.0957 + 0.0611 = 0.1568 \end{aligned}$$

The marginal probability, $P(\texttt{final} = \texttt{A})$, was calculated by adding up all the joint probabilities on the right side of the tree that correspond to $\texttt{final} = \texttt{A}$. We may now

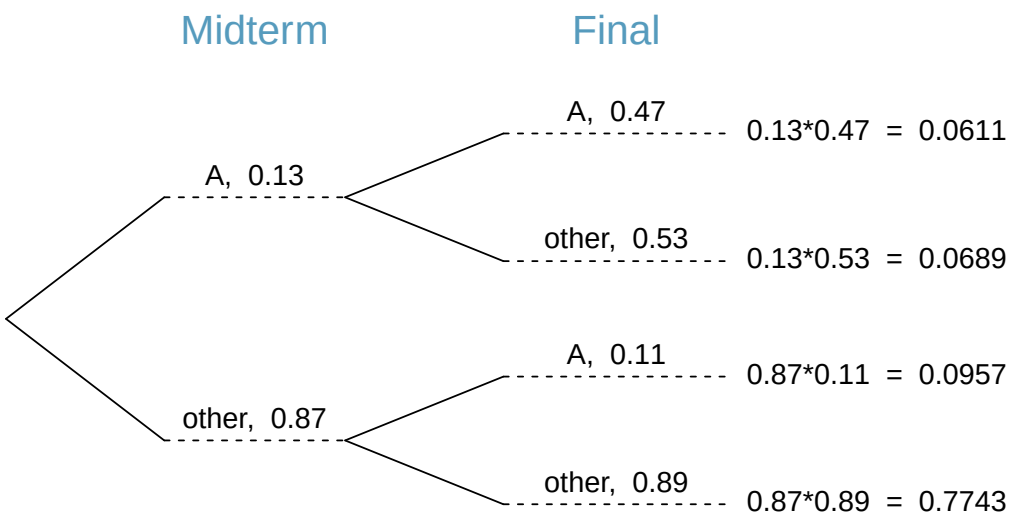


Figure 7.10: A tree diagram describing the `midterm` and `final` variables.

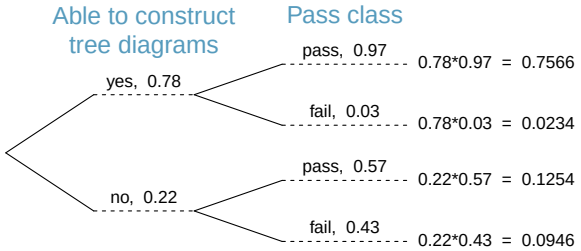
finally take the ratio of the two probabilities:

$$P(\texttt{midterm} = \texttt{A} | \texttt{final} = \texttt{A}) = \frac{P(\texttt{midterm} = \texttt{A} \text{ and } \texttt{final} = \texttt{A})}{P(\texttt{final} = \texttt{A})}$$
$$= \frac{0.0611}{0.1568} = 0.3897$$

The probability the student also earned an A on the midterm is about 0.39.

⊙ **Guided Practice 7.50** After an introductory statistics course, 78% of students can successfully construct tree diagrams. Of those who can construct tree diagrams, 97% passed, while only 57% of those students who could not construct tree diagrams passed. (a) Organize this information into a tree diagram. (b) What is the probability that a randomly selected student passed? (c) Compute the probability a student is able to construct a tree diagram if it is known that she passed.[50]

[50] (a) The tree diagram is shown to the right. (b) Identify which two joint probabilities represent students who passed, and add them: $P(\text{passed}) = 0.7566 + 0.1254 = 0.8820$. (c) $P(\text{construct tree diagram} \mid \text{passed}) = \frac{0.7566}{0.8820} = 0.8578$.



Bayes' Theorem

In many instances, we are given a conditional probability of the form

$$P(\text{statement about variable 1} \mid \text{statement about variable 2})$$

but we would really like to know the inverted conditional probability:

$$P(\text{statement about variable 2} \mid \text{statement about variable 1})$$

Tree diagrams can be used to find the second conditional probability when given the first. However, sometimes it is not possible to draw the scenario in a tree diagram. In these cases, we can apply a very useful and general formula: Bayes' Theorem.

We first take a critical look at an example of inverting conditional probabilities where we still apply a tree diagram.

Example 7.51 In Canada, about 0.35% of women over 40 will develop breast cancer in any given year. A common screening test for cancer is the mammogram, but this test is not perfect. In about 11% of patients with breast cancer, the test gives a **false negative**: it indicates a woman does not have breast cancer when she does have breast cancer. Similarly, the test gives a **false positive** in 7% of patients who do not have breast cancer: it indicates these patients have breast cancer when they actually do not.[51] If we tested a random woman over 40 for breast cancer using a mammogram and the test came back positive – that is, the test suggested the patient has cancer – what is the probability that the patient actually has breast cancer?

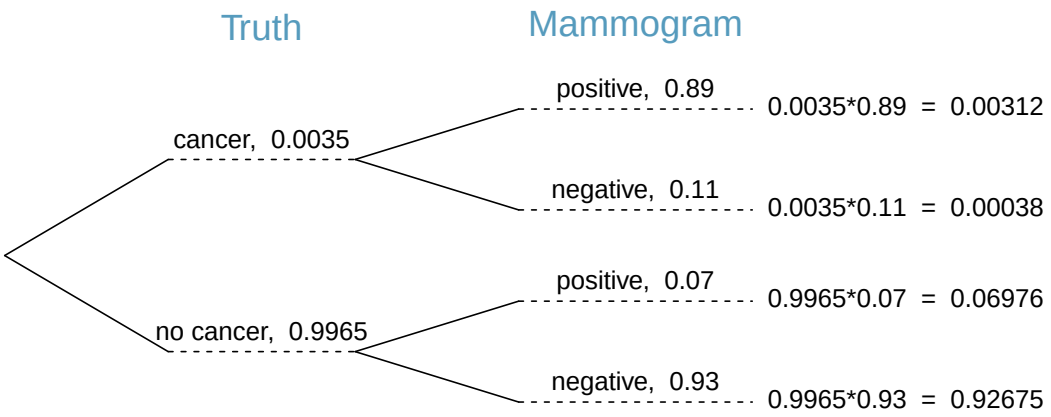


Figure 7.11: Tree diagram for Example 51, computing the probability a random patient who tests positive on a mammogram actually has breast cancer.

Notice that we are given sufficient information to quickly compute the probability of testing positive if a woman has breast cancer ($1.00 - 0.11 = 0.89$). However, we seek

[51] The probabilities reported here were obtained using studies reported at www.breastcancer.org and www.ncbi.nlm.nih.gov/pmc/articles/PMC1173421.

the inverted probability of cancer given a positive test result. (Watch out for the non-intuitive medical language: a *positive* test result suggests the possible presence of cancer in a mammogram screening.) This inverted probability may be broken into two pieces:

$$P(\text{has BC} \mid \text{mammogram}^+) = \frac{P(\text{has BC and mammogram}^+)}{P(\text{mammogram}^+)}$$

where "has BC" is an abbreviation for the patient actually having breast cancer and "mammogram$^+$" means the mammogram screening was positive. A tree diagram is useful for identifying each probability and is shown in Figure 7.11. The probability the patient has breast cancer and the mammogram is positive is

$$\begin{aligned} P(\text{has BC and mammogram}^+) &= P(\text{mammogram}^+ \mid \text{has BC})P(\text{has BC}) \\ &= 0.89 \times 0.0035 = 0.00312 \end{aligned}$$

The probability of a positive test result is the sum of the two corresponding scenarios:

$$\begin{aligned} P(\underline{\text{mammogram}^+}) &= P(\underline{\text{mammogram}^+} \text{ and has BC}) + P(\underline{\text{mammogram}^+} \text{ and no BC}) \\ &= P(\text{has BC})P(\text{mammogram}^+ \mid \text{has BC}) \\ &\qquad + P(\text{no BC})P(\text{mammogram}^+ \mid \text{no BC}) \\ &= 0.0035 \times 0.89 + 0.9965 \times 0.07 = 0.07288 \end{aligned}$$

Then if the mammogram screening is positive for a patient, the probability the patient has breast cancer is

$$\begin{aligned} P(\text{has BC} \mid \text{mammogram}^+) &= \frac{P(\text{has BC and mammogram}^+)}{P(\text{mammogram}^+)} \\ &= \frac{0.00312}{0.07288} \approx 0.0428 \end{aligned}$$

That is, even if a patient has a positive mammogram screening, there is still only a 4% chance that she has breast cancer.

Example 51 highlights why doctors often run more tests regardless of a first positive test result. When a medical condition is rare, a single positive test isn't generally definitive.

Consider again the last equation of Example 51. Using the tree diagram, we can see that the numerator (the top of the fraction) is equal to the following product:

$$P(\text{has BC and mammogram}^+) = P(\text{mammogram}^+ \mid \text{has BC})P(\text{has BC})$$

The denominator – the probability the screening was positive – is equal to the sum of probabilities for each positive screening scenario:

$$P(\underline{\text{mammogram}^+}) = P(\underline{\text{mammogram}^+} \text{ and no BC}) + P(\underline{\text{mammogram}^+} \text{ and has BC})$$

In the example, each of the probabilities on the right side was broken down into a product

of a conditional probability and marginal probability using the tree diagram.

$$\begin{aligned} P(\text{mammogram}^+) &= P(\text{mammogram}^+ \text{ and no BC}) + P(\text{mammogram}^+ \text{ and has BC}) \\ &= P(\text{mammogram}^+ \mid \text{no BC})P(\text{no BC}) \\ &\quad + P(\text{mammogram}^+ \mid \text{has BC})P(\text{has BC}) \end{aligned}$$

We can see an application of Bayes' Theorem by substituting the resulting probability expressions into the numerator and denominator of the original conditional probability.

$$P(\text{has BC} \mid \text{mammogram}^+) = \frac{P(\text{mammogram}^+ \mid \text{has BC})P(\text{has BC})}{P(\text{mammogram}^+ \mid \text{no BC})P(\text{no BC}) + P(\text{mammogram}^+ \mid \text{has BC})P(\text{has BC})}$$

Bayes' Theorem: inverting probabilities

Consider the following conditional probability for variable 1 and variable 2:

$$P(\text{outcome } A_1 \text{ of variable 1} \mid \text{outcome } B \text{ of variable 2})$$

Bayes' Theorem states that this conditional probability can be identified as the following fraction:

$$\frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)} \tag{52}$$

where A_2, A_3, ..., and A_k represent all other possible outcomes of the first variable.

Bayes' Theorem is just a generalization of what we have done using tree diagrams. The numerator identifies the probability of getting both A_1 and B. The denominator is the marginal probability of getting B. This bottom component of the fraction appears long and complicated since we have to add up probabilities from all of the different ways to get B. We always completed this step when using tree diagrams. However, we usually did it in a separate step so it didn't seem as complex.

To apply Bayes' Theorem correctly, there are two preparatory steps:

(1) First identify the marginal probabilities of each possible outcome of the first variable: $P(A_1)$, $P(A_2)$, ..., $P(A_k)$.

(2) Then identify the probability of the outcome B, conditioned on each possible scenario for the first variable: $P(B|A_1)$, $P(B|A_2)$, ..., $P(B|A_k)$.

Once each of these probabilities are identified, they can be applied directly within the formula.

TIP: Only use Bayes' Theorem when tree diagrams are difficult

Drawing a tree diagram makes it easier to understand how two variables are connected. Use Bayes' Theorem only when there are so many scenarios that drawing a tree diagram would be complex.

Guided Practice 7.53 Jose visits campus every Thursday evening. However, some days the parking garage is full, often due to college events. There are academic events on 35% of evenings, sporting events on 20% of evenings, and no events on 45% of evenings. When there is an academic event, the garage fills up about 25% of the time, and it fills up 70% of evenings with sporting events. On evenings when there are no events, it only fills up about 5% of the time. If Jose comes to campus and finds the garage full, what is the probability that there is a sporting event? Use a tree diagram to solve this problem.[52]

Example 7.54 Here we solve the same problem presented in Guided Practice 53, except this time we use Bayes' Theorem.

The outcome of interest is whether there is a sporting event (call this A_1), and the condition is that the lot is full (B). Let A_2 represent an academic event and A_3 represent there being no event on campus. Then the given probabilities can be written as

$$P(A_1) = 0.2 \qquad P(A_2) = 0.35 \qquad P(A_3) = 0.45$$
$$P(B|A_1) = 0.7 \qquad P(B|A_2) = 0.25 \qquad P(B|A_3) = 0.05$$

Bayes' Theorem can be used to compute the probability of a sporting event (A_1) under the condition that the parking lot is full (B):

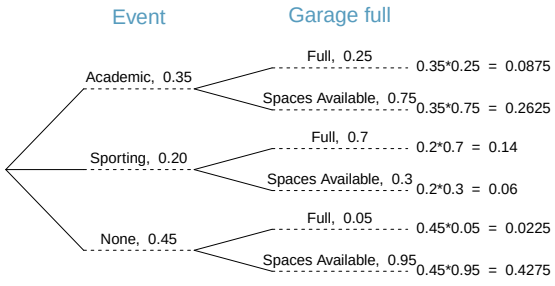
$$\begin{aligned} P(A_1|B) &= \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)} \\ &= \frac{(0.7)(0.2)}{(0.7)(0.2) + (0.25)(0.35) + (0.05)(0.45)} \\ &= 0.56 \end{aligned}$$

Based on the information that the garage is full, there is a 56% probability that a sporting event is being held on campus that evening.

Guided Practice 7.55 Use the information in the previous exercise and example to verify the probability that there is an academic event conditioned on the parking lot

---

[52] The tree diagram, with three primary branches, is shown to the right. Next, we identify two probabilities from the tree diagram. (1) The probability that there is a sporting event and the garage is full: 0.14. (2) The probability the garage is full: 0.0875 + 0.14 + 0.0225 = 0.25. Then the solution is the ratio of these probabilities: $\frac{0.14}{0.25} = 0.56$. If the garage is full, there is a 56% probability that there is a sporting event.



being full is 0.35.[53]

⊙ **Guided Practice 7.56** In Guided Practice 53 and 55, you found that if the parking lot is full, the probability there is a sporting event is 0.56 and the probability there is an academic event is 0.35. Using this information, compute $P(\text{no event} \mid \text{the lot is full})$.[54]

The last several exercises offered a way to update our belief about whether there is a sporting event, academic event, or no event going on at the school based on the information that the parking lot was full. This strategy of *updating beliefs* using Bayes' Theorem is actually the foundation of an entire section of statistics called **Bayesian statistics**. While Bayesian statistics is very important and useful, we will not have time to cover much more of it in this book.

[53]Short answer:

$$\begin{aligned} P(A_2|B) &= \frac{P(B|A_2)P(A_2)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)} \\ &= \frac{(0.25)(0.35)}{(0.7)(0.2) + (0.25)(0.35) + (0.05)(0.45)} \\ &= 0.35 \end{aligned}$$

[54]Each probability is conditioned on the same information that the garage is full, so the complement may be used: $1.00 - 0.56 - 0.35 = 0.09$.

8 Distributions

8.1 Normal Distribution

Among all the distributions we see in practice, one is overwhelmingly the most common. The symmetric, unimodal, bell curve is ubiquitous throughout statistics. Indeed it is so common, that people often know it as the **normal curve** or **normal distribution**,[55] shown in Figure 8.1. Many variables observed in nature closely follow the normal distribution.

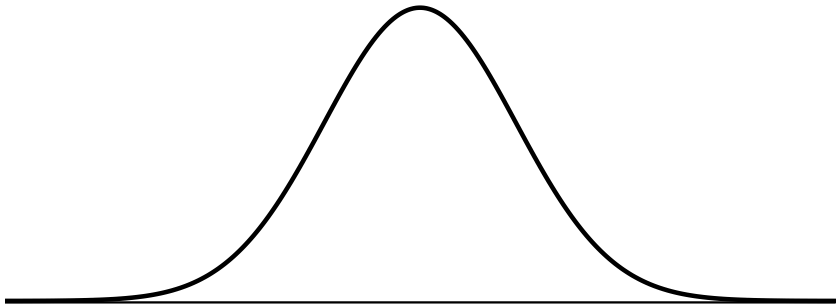


Figure 8.1: A normal curve.

Normal distribution facts

Many variables are nearly normal, but none are exactly normal. Thus the normal distribution, while not perfect for any single problem, is very useful for a variety of problems. We will use it in data exploration and to solve important problems in statistics.

Normal distribution model

The normal distribution model always describes a symmetric, unimodal, bell-shaped curve. However, these curves can look different depending on the details of the model. Specifically, the normal distribution model can be adjusted using two parameters: mean and standard deviation. As you can probably guess, changing the mean shifts the bell curve to the left or right, while changing the standard deviation stretches or constricts the curve. Figure 8.2 shows the normal distribution with mean 0 and standard deviation 1 in the left panel and the normal distributions with mean 19 and standard deviation 4 in the right panel. Figure 8.3 shows these distributions on the same axis.

If a normal distribution has mean μ and standard deviation σ, we may write the distribution as $N(\mu, \sigma)$. The two distributions in Figure 8.3 can be written as

$$N(\mu = 0, \sigma = 1) \quad \text{and} \quad N(\mu = 19, \sigma = 4)$$

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

[55] It is also introduced as the Gaussian distribution after Carl Friedrich Gauss, the first person to formalize its mathematical expression.

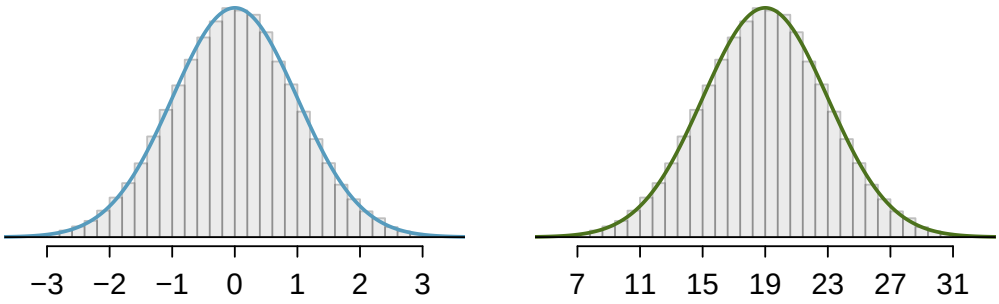


Figure 8.2: Both curves represent the normal distribution, however, they differ in their centre and spread. The normal distribution with mean 0 and standard deviation 1 is called the **standard normal distribution**.

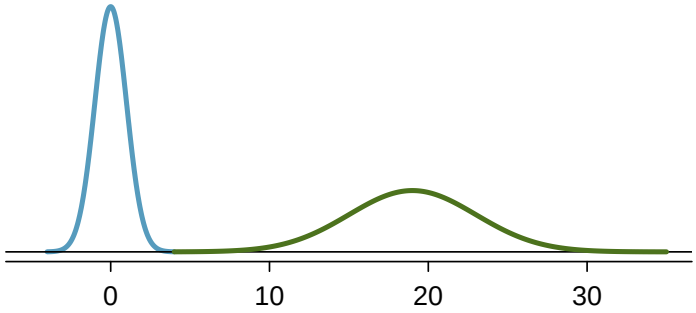


Figure 8.3: The normal models shown in Figure 8.2 but plotted together and on the same scale.

Because the mean and standard deviation describe a normal distribution exactly, they are called the distribution's **parameters**.

⊙ **Guided Practice 8.1** Write down the short-hand for a normal distribution with[56]

(a) mean 5 and standard deviation 3,
(b) mean -100 and standard deviation 10, and
(c) mean 2 and standard deviation 9.

Standardizing with Z-scores

● **Example 8.2** Table 8.1 on the next page shows the mean and standard deviation for total scores on two aptitude tests for entrance to US universities, the SAT and ACT. The distribution of SAT and ACT scores are both nearly normal. Suppose Ann scored 1800 on her SAT and Tom scored 24 on his ACT. Who performed better?

We use the standard deviation as a guide. Ann is 1 standard deviation above average on the SAT: $1500 + 300 = 1800$. Tom is 0.6 standard deviations above the mean on

[56] (a) $N(\mu = 5, \sigma = 3)$. (b) $N(\mu = -100, \sigma = 10)$. (c) $N(\mu = 2, \sigma = 9)$.

	SAT	ACT
Mean	1500	21
SD	300	5

Table 8.1: Mean and standard deviation for the SAT and ACT.

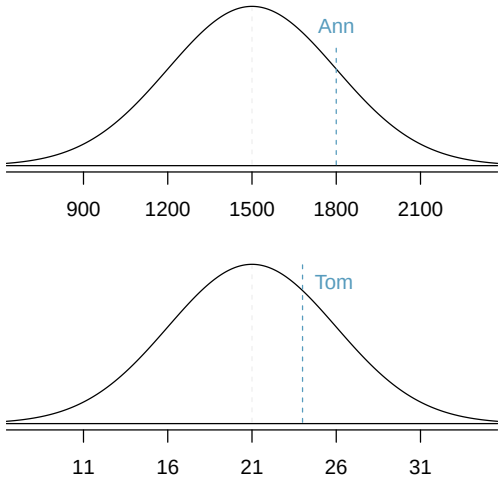


Figure 8.4: Ann's and Tom's scores shown with the distributions of SAT and ACT scores.

the ACT: $21 + 0.6 \times 5 = 24$. In Figure 8.4, we can see that Ann tends to do better with respect to everyone else than Tom did, so her score was better.

Example 2 used a standardization technique called a Z-score, a method most commonly employed for nearly normal observations but that may be used with any distribution. The **Z-score** of an observation is defined as the number of standard deviations it falls above or below the mean. If the observation is one standard deviation above the mean, its Z-score is 1. If it is 1.5 standard deviations *below* the mean, then its Z-score is -1.5. If x is an observation from a distribution $N(\mu, \sigma)$, we define the Z-score mathematically as

Z

Z-score, the standardized observation

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

Using $\mu_{SAT} = 1500$, $\sigma_{SAT} = 300$, and $x_{Ann} = 1800$, we find Ann's Z-score:

$$Z_{Ann} = \frac{x_{Ann} - \mu_{SAT}}{\sigma_{SAT}} = \frac{1800 - 1500}{300} = 1$$

The Z-score

The Z-score of an observation is the number of standard deviations it falls above or below the mean. We compute the Z-score for an observation x that follows a distribution with mean μ and standard deviation σ using

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

⊙ **Guided Practice 8.3** Use Tom's ACT score, 24, along with the ACT mean and standard deviation to compute his Z-score.[57]

Observations above the mean always have positive Z-scores while those below the mean have negative Z-scores. If an observation is equal to the mean (e.g. SAT score of 1500), then the Z-score is 0.

⊙ **Guided Practice 8.4** Let X represent a random variable from $N(\mu = 3, \sigma = 2)$, and suppose we observe $x = 5.19$. (a) Find the Z-score of x. (b) Use the Z-score to determine how many standard deviations above or below the mean x falls.[58]

⊙ **Guided Practice 8.5** Head lengths of brushtail possums follow a nearly normal distribution with mean 92.6 mm and standard deviation 3.6 mm. Compute the Z-scores for possums with head lengths of 95.4 mm and 85.8 mm.[59]

We can use Z-scores to roughly identify which observations are more unusual than others. 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: $|Z_1| > |Z_2|$. This technique is especially insightful when a distribution is symmetric.

⊙ **Guided Practice 8.6** Which of the observations in Guided Practice 5 is more unusual?[60]

We can use the normal model to find percentiles. A **normal probability table**, which lists Z-scores and corresponding percentiles used to be used to identify a percentile based on the Z-score (and vice versa). Looking up statistical tables is no longer needed as finding these values in **R** and other software is very simple.

---

[57] $Z_{Tom} = \frac{x_{Tom} - \mu_{ACT}}{\sigma_{ACT}} = \frac{24-21}{5} = 0.6$

[58] (a) Its Z-score is given by $Z = \frac{x-\mu}{\sigma} = \frac{5.19-3}{2} = 2.19/2 = 1.095$. (b) The observation x is 1.095 standard deviations *above* the mean. We know it must be above the mean since Z is positive.

[59] For $x_1 = 95.4$ mm: $Z_1 = \frac{x_1-\mu}{\sigma} = \frac{95.4-92.6}{3.6} = 0.78$. For $x_2 = 85.8$ mm: $Z_2 = \frac{85.8-92.6}{3.6} = -1.89$.

[60] Because the *absolute value* of Z-score for the second observation is larger than that of the first, the second observation has a more unusual head length.

```r
# calculate the probability P(Z<0.43) - i.e. what is the probability of
# standard normal variable taking a value less than 0.43?
pnorm(q = 0.43, mean = 0, sd = 1)

## [1] 0.6664022

# calculate the $Z$-score satisfying P(Z<z)=0.80 - i.e. what value is at the
# 80th percentile of a standard normal distribution?
qnorm(p = 0.8, mean = 0, sd = 1)

## [1] 0.8416212

# to see more options, type '?rnorm'
```

⊙ **Guided Practice 8.7** Determine the proportion of SAT test takers who scored better than Ann on the SAT.[61]

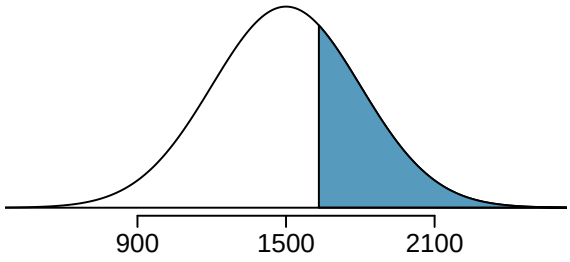
[61] If 84% had lower scores than Ann, the proportion of people who had better scores must be 16%. (Generally ties are ignored when the normal model, or any other continuous distribution, is used.)

Normal probability examples

Cumulative SAT scores are approximated well by a normal model, $N(\mu = 1500, \sigma = 300)$.

Example 8.8 Shannon is a randomly selected SAT taker, and nothing is known about Shannon's SAT aptitude. What is the probability Shannon scores at least 1630 on her SATs?

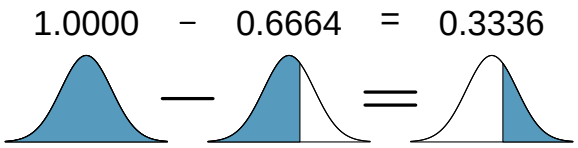
First, always draw and label a picture of the normal distribution. (Drawings need not be exact to be useful.) We are interested in the chance she scores above 1630, so we shade this upper tail:



The picture shows the mean and the values at 2 standard deviations above and below the mean. The simplest way to find the shaded area under the curve makes use of the Z-score of the cutoff value. With $\mu = 1500$, $\sigma = 300$, and the cutoff value $x = 1630$, the Z-score is computed as

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

We look up the percentile of $Z = 0.43$ in the normal probability table or using statistical software, which yields 0.6664. However, the percentile describes those who had a Z-score *lower* than 0.43. To find the area *above* $Z = 0.43$, we compute one minus the area of the lower tail:



The probability Shannon scores at least 1630 on the SAT is 0.3336.

TIP: always draw a picture first, and find the Z-score second

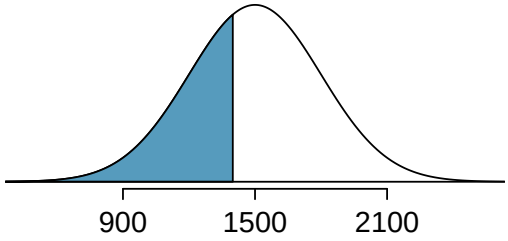
For any normal probability situation, *always always always* draw and label the normal curve and shade the area of interest first. The picture will provide an estimate of the probability.

After drawing a figure to represent the situation, identify the Z-score for the observation of interest.

Guided Practice 8.9 If the probability of Shannon scoring at least 1630 is 0.3336, then what is the probability she scores less than 1630? Draw the normal curve representing this exercise, shading the lower region instead of the upper one.[62]

Example 8.10 Edward earned a 1400 on his SAT. What is his percentile?

First, a picture is needed. Edward's percentile is the proportion of people who do not get as high as a 1400. These are the scores to the left of 1400.



Identifying the mean $\mu = 1500$, the standard deviation $\sigma = 300$, and the cutoff for the tail area $x = 1400$ makes it easy to compute the Z-score:

$$Z = \frac{x - \mu}{\sigma} = \frac{1400 - 1500}{300} = -0.33$$

Using the normal probability table, identify the row of -0.3 and column of 0.03, which corresponds to the probability 0.3707. Edward is at the 37^{th} percentile.

Guided Practice 8.11 Use the results of Example 10 to compute the proportion of SAT takers who did better than Edward. Also draw a new picture.[63]

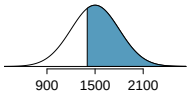
TIP: areas to the right

The normal probability table in most books gives the area to the left. If you would like the area to the right, first find the area to the left and then subtract this amount from one.

Guided Practice 8.12 Stuart earned an SAT score of 2100. Draw a picture for each part. (a) What is his percentile? (b) What percent of SAT takers did better than Stuart?[64]

[62] We found the probability in Example 8: 0.6664. A picture for this exercise is represented by the shaded area below "0.6664" in Example 8.

[63] If Edward did better than 37% of SAT takers, then about 63% must have done better than him.



[64] Numerical answers: (a) 0.9772. (b) 0.0228.

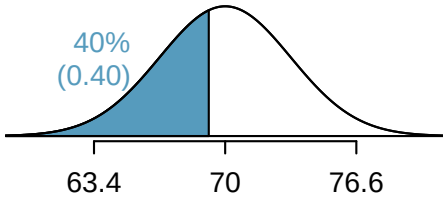
Based on a sample of 100 men,[65] the heights of male adults between the ages 20 and 62 in the US is nearly normal with mean 70.0" and standard deviation 3.3".

Guided Practice 8.13 Mike is 5'7" and Jim is 6'4". (a) What is Mike's height percentile? (b) What is Jim's height percentile? Also draw one picture for each part.[66]

The last several problems have focused on finding the probability or percentile for a particular observation. What if you would like to know the observation corresponding to a particular percentile?

Example 8.14 Erik's height is at the 40^{th} percentile. How tall is he?

As always, first draw the picture.



In this case, the lower tail probability is known (0.40), which can be shaded on the diagram. We want to find the observation that corresponds to this value. As a first step in this direction, we determine the Z-score associated with the 40^{th} percentile.

Because the percentile is below 50%, we know Z will be negative. Looking in the negative part of the normal probability table, we search for the probability *inside* the table closest to 0.4000. We find that 0.4000 falls in row -0.2 and between columns 0.05 and 0.06. Since it falls closer to 0.05, we take this one: $Z = -0.25$.

Knowing $Z_{Erik} = -0.25$ and the population parameters $\mu = 70$ and $\sigma = 3.3$ inches, the Z-score formula can be set up to determine Erik's unknown height, labelled x_{Erik}:

$$-0.25 = Z_{Erik} = \frac{x_{Erik} - \mu}{\sigma} = \frac{x_{Erik} - 70}{3.3}$$

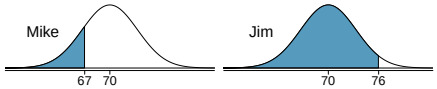
Solving for x_{Erik} yields the height 69.18 inches. That is, Erik is about 5'9" (this is notation for 5-feet, 9-inches).

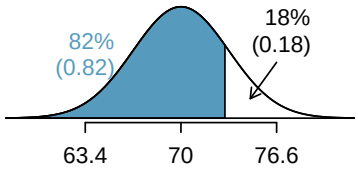
Example 8.15 What is the adult male height at the 82^{nd} percentile?

Again, we draw the figure first.

[65] This sample was taken from the USDA Food Commodity Intake Database.

[66] First put the heights into inches: 67 and 76 inches. Figures are shown below. (a) $Z_{Mike} = \frac{67-70}{3.3} = -0.91 \rightarrow 0.1814$. (b) $Z_{Jim} = \frac{76-70}{3.3} = 1.82 \rightarrow 0.9656$.





Next, we want to find the Z-score at the 82^{nd} percentile, which will be a positive value. Looking in the Z-table, we find Z falls in row 0.9 and the nearest column is 0.02, i.e. $Z = 0.92$. Finally, the height x is found using the Z-score formula with the known mean μ, standard deviation σ, and Z-score $Z = 0.92$:

$$0.92 = Z = \frac{x - \mu}{\sigma} = \frac{x - 70}{3.3}$$

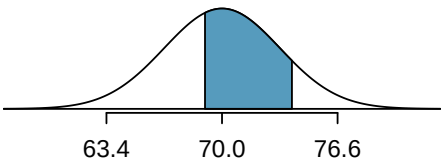
This yields 73.04 inches or about 6'1" as the height at the 82^{nd} percentile.

Guided Practice 8.16 (a) What is the 95^{th} percentile for SAT scores? (b) What is the 97.5^{th} percentile of the male heights? As always with normal probability problems, first draw a picture.[67]

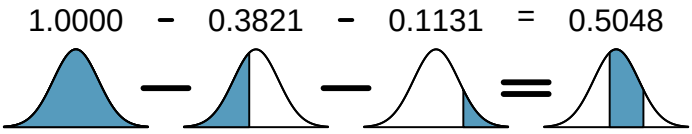
Guided Practice 8.17 (a) What is the probability that a randomly selected male adult is at least 6'2" (74 inches)? (b) What is the probability that a male adult is shorter than 5'9" (69 inches)?[68]

Example 8.18 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 no longer an upper or lower tail.



The total area under the curve is 1. If we find the area of the two tails that are not shaded (from Guided Practice 17, these areas are 0.3821 and 0.1131), then we can find the middle area:



[67] Remember: draw a picture first, then find the Z-score. (We leave the pictures to you.) The Z-score can be found by using the percentiles and the normal probability table. (a) We look for 0.95 in the probability portion (middle part) of the normal probability table, which leads us to row 1.6 and (about) column 0.05, i.e. $Z_{95} = 1.65$. Knowing $Z_{95} = 1.65$, $\mu = 1500$, and $\sigma = 300$, we setup the Z-score formula: $1.65 = \frac{x_{95} - 1500}{300}$. We solve for x_{95}: $x_{95} = 1995$. (b) Similarly, we find $Z_{97.5} = 1.96$, again setup the Z-score formula for the heights, and calculate $x_{97.5} = 76.5$.

[68] Numerical answers: (a) 0.1131. (b) 0.3821.

That is, the probability of being between 5'9" and 6'2" is 0.5048.

⊙ **Guided Practice 8.19** What percent of SAT takers get between 1500 and 2000?[69]

⊙ **Guided Practice 8.20** What percent of adult males are between 5'5" and 5'7"?[70]

68-95-99.7 rule

Here, we present a useful rule of thumb for the probability of falling within 1, 2, and 3 standard deviations of the mean in the normal distribution. This will be useful in a wide range of practical settings, especially when trying to make a quick estimate without a calculator or Z-table.

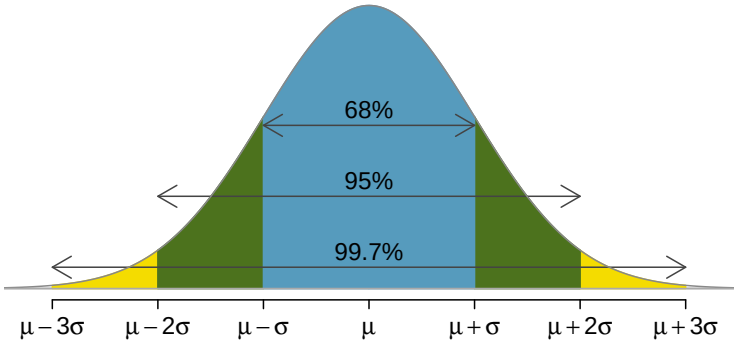


Figure 8.5: Probabilities for falling within 1, 2 and 3 standard deviations of the mean in a normal distribution.

⊙ **Guided Practice 8.21** Use the Z-table to confirm that about 68%, 95%, and 99.7% of observations fall within 1, 2, and 3, standard deviations of the mean in the normal distribution, respectively. For instance, first find the area that falls between $Z = -1$ and $Z = 1$, which should have an area of about 0.68. Similarly there should be an area of about 0.95 between $Z = -2$ and $Z = 2$.[71]

It is possible for a normal random variable to fall 4, 5, or even more standard deviations from the mean. However, these occurrences are very rare if the data are nearly normal. The probability of being further than 4 standard deviations from the mean is about 1-in-15,000. For 5 and 6 standard deviations, it is about 1-in-2 million and 1-in-500 million, respectively.

[69] This is an abbreviated solution. (Be sure to draw a figure!) First find the percent who get below 1500 and the percent that get above 2000: $Z_{1500} = 0.00 \rightarrow 0.5000$ (area below), $Z_{2000} = 1.67 \rightarrow 0.0475$ (area above). Final answer: $1.0000 - 0.5000 - 0.0475 = 0.4525$.

[70] 5'5" is 65 inches. 5'7" is 67 inches. Numerical solution: $1.000 - 0.0649 - 0.8183 = 0.1168$, i.e. 11.68%.

[71] First draw the pictures. To find the area between $Z = -1$ and $Z = 1$, use the normal probability table to determine the areas below $Z = -1$ and above $Z = 1$. Next verify the area between $Z = -1$ and $Z = 1$ is about 0.68. Repeat this for $Z = -2$ to $Z = 2$ and also for $Z = -3$ to $Z = 3$.

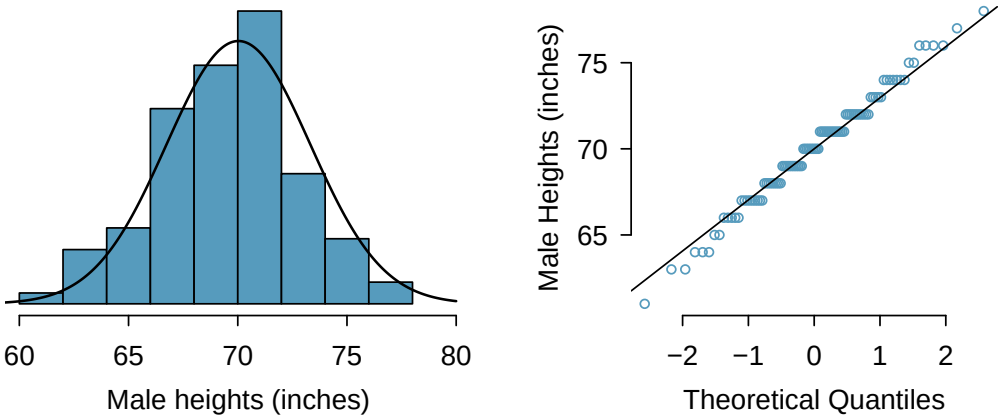


Figure 8.6: A sample of 100 male heights. The observations are rounded to the nearest whole inch, explaining why the points appear to jump in increments in the normal probability plot.

⊙ **Guided Practice 8.22** SAT scores closely follow the normal model with mean $\mu = 1500$ and standard deviation $\sigma = 300$. (a) About what percent of test takers score 900 to 2100? (b) What percent score between 1500 and 2100?[72]

8.2 Evaluating the Normal Approximation

Many processes can be well approximated by the normal distribution. We have already seen two good examples: SAT scores and the heights of US adult males. While using a normal model can be extremely convenient and helpful, it is important to remember normality is always an approximation. Testing the appropriateness of the normal assumption is a key step in many data analyses.

Example 14 suggests the distribution of heights of US males is well approximated by the normal model. We are interested in proceeding under the assumption that the data are normally distributed, but first we must check to see if this is reasonable.

There are two visual methods for checking the assumption of normality, which can be implemented and interpreted quickly. The first is a simple histogram with the best fitting normal curve overlaid on the plot, as shown in the left panel of Figure 8.6. The sample mean $\bar{x}$ and standard deviation s are used as the parameters of the best fitting normal curve. The closer this curve fits the histogram, the more reasonable the normal model assumption. Another more common method is examining a **normal probability plot**,[73] shown in the right panel of Figure 8.6. The closer the points are to a perfect straight line, the more confident we can be that the data follow the normal model.

[72] (a) 900 and 2100 represent two standard deviations above and below the mean, which means about 95% of test takers will score between 900 and 2100. (b) Since the normal model is symmetric, then half of the test takers from part (a) ($\frac{95\%}{2} = 47.5\%$ of all test takers) will score 900 to 1500 while 47.5% score between 1500 and 2100.

[73] Also commonly called a **quantile-quantile plot**.

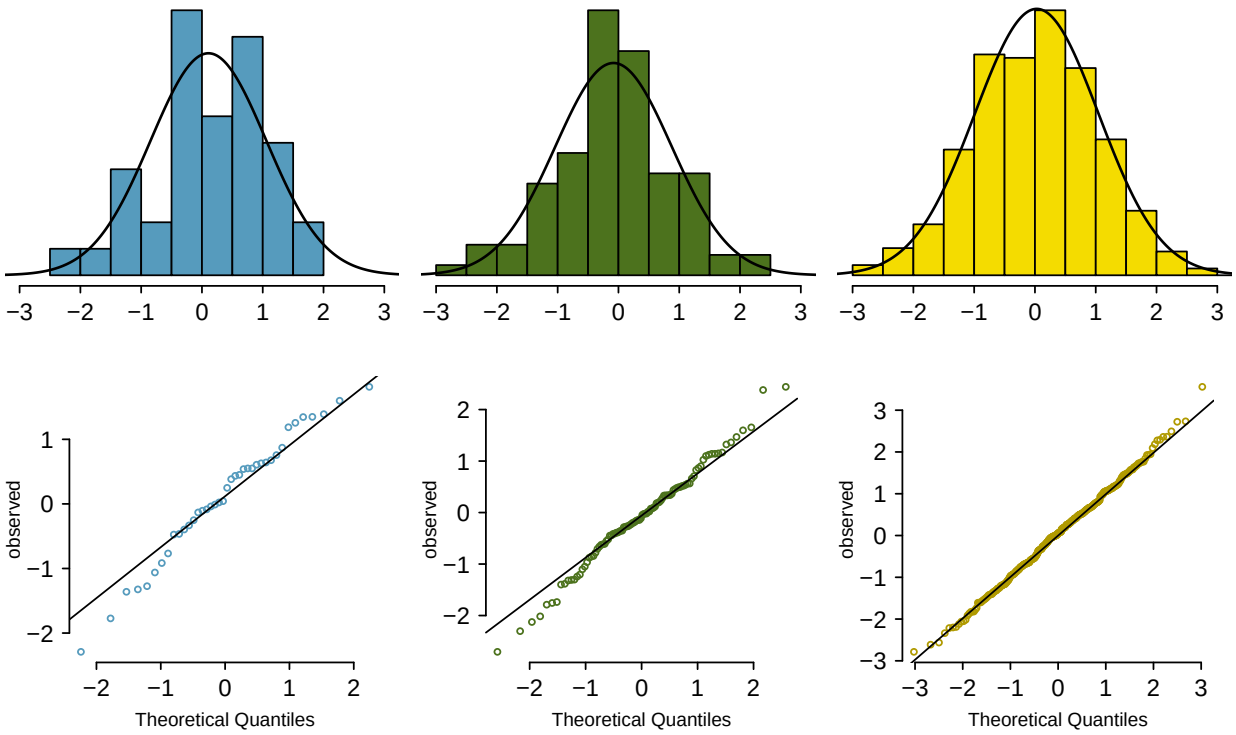


Figure 8.7: Histograms and normal probability plots for three simulated normal data sets; $n = 40$ (left), $n = 100$ (middle), $n = 400$ (right).

Example 8.23 Three data sets of 40, 100, and 400 samples were simulated from a normal distribution, and the histograms and normal probability plots of the data sets are shown in Figure 8.7. These will provide a benchmark for what to look for in plots of real data.

The left panels show the histogram (top) and normal probability plot (bottom) for the simulated data set with 40 observations. The data set is too small to really see clear structure in the histogram. The normal probability plot also reflects this, where there are some deviations from the line. We should expect deviations of this amount for such a small data set.

The middle panels show diagnostic plots for the data set with 100 simulated observations. The histogram shows more normality and the normal probability plot shows a better fit. While there are a few observations that deviate noticeably from the line, they are not particularly extreme.

The data set with 400 observations has a histogram that greatly resembles the normal distribution, while the normal probability plot is nearly a perfect straight line. Again in the normal probability plot there is one observation (the largest) that deviates slightly from the line. If that observation had deviated 3 times further from the line, it would be of greater importance in a real data set. Apparent outliers can occur in normally distributed data but they are rare.

Notice the histograms look more normal as the sample size increases, and the normal probability plot becomes straighter and more stable.

Example 8.24 Are NBA player heights normally distributed? Consider all 435 NBA players from the 2008-9 season presented in Figure 8.8.[74]

We first create a histogram and normal probability plot of the NBA player heights. The histogram in the left panel is slightly left skewed, which contrasts with the symmetric normal distribution. The points in the normal probability plot do not appear to closely follow a straight line but show what appears to be a "wave". We can compare these characteristics to the sample of 400 normally distributed observations in Example 23 and see that they represent much stronger deviations from the normal model. NBA player heights do not appear to come from a normal distribution.

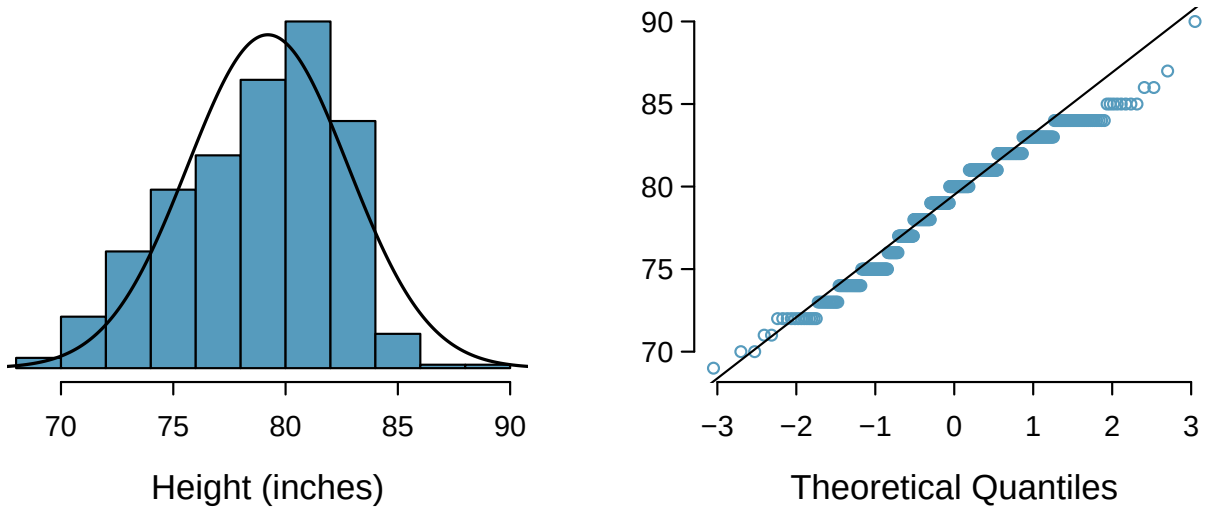


Figure 8.8: Histogram and normal probability plot for the NBA heights from the 2008-9 season.

Example 8.25 Can we approximate poker winnings by a normal distribution? We consider the poker winnings of an individual over 50 days. A histogram and normal probability plot of these data are shown in Figure 8.9.

The data are very strongly right skewed in the histogram, which corresponds to the very strong deviations on the upper right component of the normal probability plot. If we compare these results to the sample of 40 normal observations in Example 23, it is apparent that these data show very strong deviations from the normal model.

[74] These data were collected from www.nba.com.

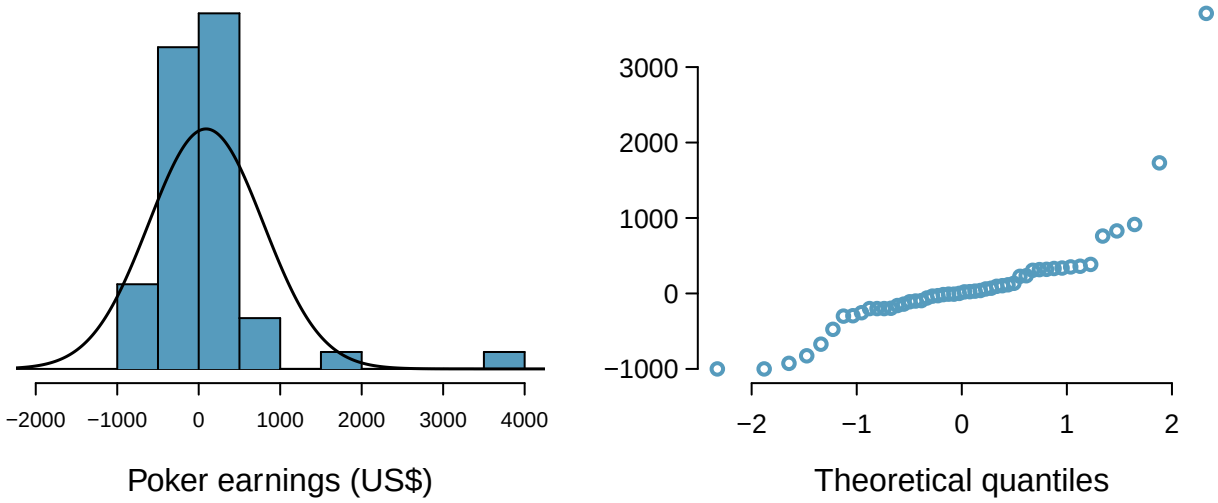


Figure 8.9: A histogram of poker data with the best fitting normal plot and a normal probability plot.

Guided Practice 8.26 Determine which data sets represented in Figure 8.10 plausibly come from a nearly normal distribution. Are you confident in all of your conclusions? There are 100 (top left), 50 (top right), 500 (bottom left), and 15 points (bottom right) in the four plots.[75]

Guided Practice 8.27 Figure 8.11 shows normal probability plots for two distributions that are skewed. One distribution is skewed to the low end (left skewed) and the other to the high end (right skewed). Which is which?[76]

8.3 Geometric Distribution

How long should we expect to flip a coin until it turns up `heads`? Or how many times should we expect to roll a die until we get a 1? These questions can be answered using the geometric distribution. We first formalize each trial – such as a single coin flip or die toss – using the Bernoulli distribution, and then we combine these with our tools from probability (Chapter 7) to construct the geometric distribution.

[75] Answers may vary a little. The top-left plot shows some deviations in the smallest values in the data set; specifically, the left tail of the data set has some outliers we should be wary of. The top-right and bottom-left plots do not show any obvious or extreme deviations from the lines for their respective sample sizes, so a normal model would be reasonable for these data sets. The bottom-right plot has a consistent curvature that suggests it is not from the normal distribution. If we examine just the vertical coordinates of these observations, we see that there is a lot of data between -20 and 0, and then about five observations scattered between 0 and 70. This describes a distribution that has a strong right skew.

[76] Examine where the points fall along the vertical axis. In the first plot, most points are near the low end with fewer observations scattered along the high end; this describes a distribution that is skewed to the high end. The second plot shows the opposite features, and this distribution is skewed to the low end.

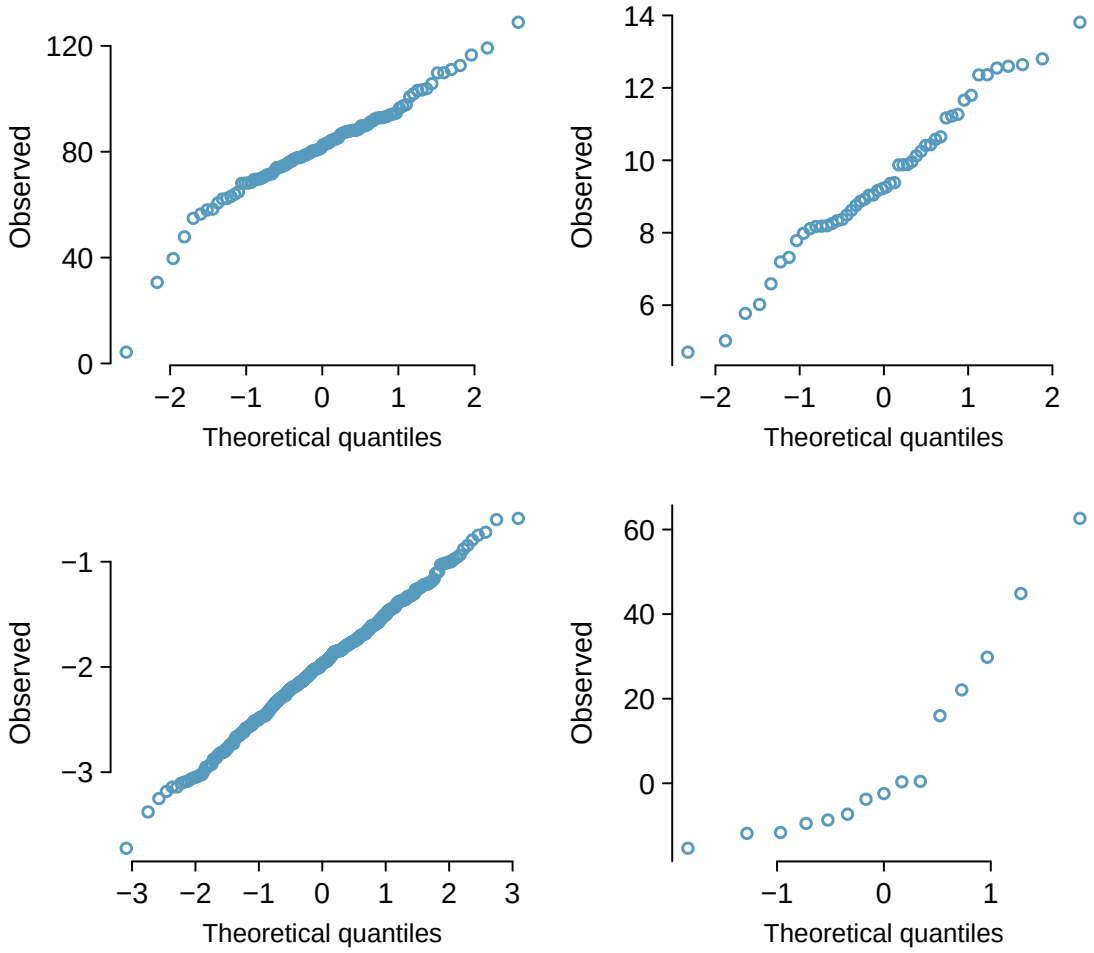


Figure 8.10: Four normal probability plots for Guided Practice 26.

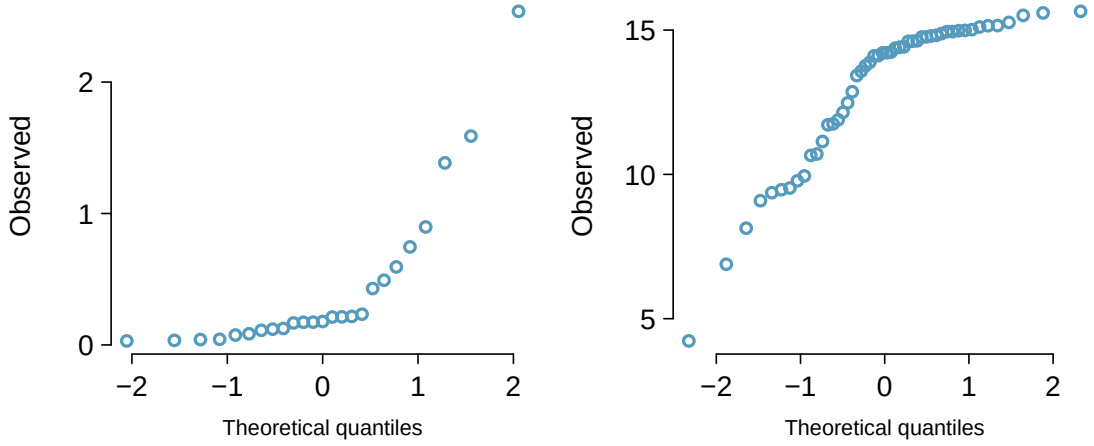


Figure 8.11: Normal probability plots for Guided Practice 27.

Bernoulli distribution

Stanley Milgram began a series of experiments in 1963 to estimate what proportion of people would willingly obey an authority and give severe shocks to a stranger. Milgram found that about 65% of people would obey the authority and give such shocks. Over the years, additional research suggested this number is approximately consistent across communities and time.[77]

Each person in Milgram's experiment can be thought of as a **trial**. We label a person a **success** if she refuses to administer the worst shock. A person is labelled a **failure** if she administers the worst shock. Because only 35% of individuals refused to administer the most severe shock, we denote the **probability of a success** with $p = 0.35$. The probability of a failure is sometimes denoted with $q = 1 - p$.

Thus, `success` or `failure` is recorded for each person in the study. When an individual trial only has two possible outcomes, it is called a **Bernoulli random variable**.

Bernoulli random variable, descriptive

A Bernoulli random variable has exactly two possible outcomes. We typically label one of these outcomes a "success" and the other outcome a "failure". We may also denote a success by `1` and a failure by `0`.

TIP: "success" need not be something positive

We chose to label a person who refuses to administer the worst shock a "success" and all others as "failures". However, we could just as easily have reversed these labels. The mathematical framework we will build does not depend on which outcome is labelled a success and which a failure, as long as we are consistent.

Bernoulli random variables are often denoted as `1` for a success and `0` for a failure. In addition to being convenient in entering data, it is also mathematically handy. Suppose we observe ten trials:

`0 1 1 1 1 0 1 1 0 0`

Then the **sample proportion**, $\hat{p}$, is the sample mean of these observations:

$$\hat{p} = \frac{\text{\# of successes}}{\text{\# of trials}} = \frac{0+1+1+1+1+0+1+1+0+0}{10} = 0.6$$

This mathematical inquiry of Bernoulli random variables can be extended even further. Because `0` and `1` are numerical outcomes, we can define the mean and standard deviation of

[77] Find further information on Milgram's experiment at www.cnr.berkeley.edu/ucce50/ag-labor/7article/article35.htm.

a Bernoulli random variable.[78]

Bernoulli random variable, mathematical

If X is a random variable that takes value 1 with probability of success p and 0 with probability $1 - p$, then X is a Bernoulli random variable with mean and standard deviation

$$\mu = p \qquad\qquad \sigma = \sqrt{p(1-p)}$$

In general, it is useful to think about a Bernoulli random variable as a random process with only two outcomes: a success or failure. Then we build our mathematical framework using the numerical labels `1` and `0` for successes and failures, respectively.

Geometric distribution

Example 8.28 Dr. Smith wants to repeat Milgram's experiments but she only wants to sample people until she finds someone who will not inflict the worst shock.[79] If the probability a person will *not* give the most severe shock is still 0.35 and the subjects are independent, what are the chances that she will stop the study after the first person? The second person? The third? What about if it takes her $n - 1$ individuals who will administer the worst shock before finding her first success, i.e. the first success is on the n^{th} person? (If the first success is the fifth person, then we say $n = 5$.)

The probability of stopping after the first person is just the chance the first person will not administer the worst shock: $1 - 0.65 = 0.35$. The probability it will be the second person is

$$\begin{aligned} &P(\text{second person is the first to not administer the worst shock}) \\ &\quad = P(\text{the first will, the second won't}) = (0.65)(0.35) = 0.228 \end{aligned}$$

Likewise, the probability it will be the third person is $(0.65)(0.65)(0.35) = 0.148$.

---

[78] If p is the true probability of a success, then the mean of a Bernoulli random variable X is given by

$$\begin{aligned} \mu = E[X] &= P(X = 0) \times 0 + P(X = 1) \times 1 \\ &= (1 - p) \times 0 + p \times 1 = 0 + p = p \end{aligned}$$

Similarly, the variance of X can be computed:

$$\begin{aligned} \sigma^2 &= P(X = 0)(0 - p)^2 + P(X = 1)(1 - p)^2 \\ &= (1 - p)p^2 + p(1 - p)^2 = p(1 - p) \end{aligned}$$

The standard deviation is $\sigma = \sqrt{p(1-p)}$.

[79] This is hypothetical since, in reality, this sort of study probably would not be permitted any longer under current ethical standards.

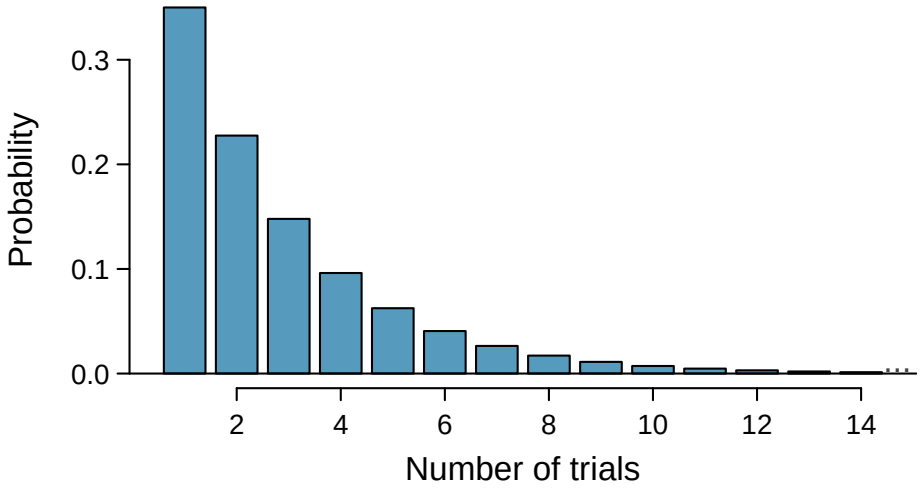


Figure 8.12: The geometric distribution when the probability of success is $p = 0.35$.

If the first success is on the n^{th} person, then there are $n-1$ failures and finally 1 success, which corresponds to the probability $(0.65)^{n-1}(0.35)$. This is the same as $(1-0.35)^{n-1}(0.35)$.

Example 28 illustrates what is called the geometric distribution, which describes the waiting time until a success for **independent and identically distributed (iid)** Bernoulli random variables. In this case, the *independence* aspect just means the individuals in the example don't affect each other, and *identical* means they each have the same probability of success.

The geometric distribution from Example 28 is shown in Figure 8.12. In general, the probabilities for a geometric distribution decrease **exponentially** fast.

While this text will not derive the formulas for the mean (expected) number of trials needed to find the first success or the standard deviation or variance of this distribution, we present general formulas for each.

Geometric Distribution

If the probability of a success in one trial is p and the probability of a failure is $1-p$, then the probability of finding the first success in the n^{th} trial is given by

$$(1-p)^{n-1}p \tag{29}$$

The mean (i.e. expected value), variance, and standard deviation of this wait time are given by

$$\mu = \frac{1}{p} \qquad \sigma^2 = \frac{1-p}{p^2} \qquad \sigma = \sqrt{\frac{1-p}{p^2}} \tag{30}$$

It is no accident that we use the symbol μ for both the mean and expected value. The mean and the expected value are one and the same.

The left side of Equation (30) says that, on average, it takes $1/p$ trials to get a success. This mathematical result is consistent with what we would expect intuitively. If the probability of a success is high (e.g. 0.8), then we don't usually wait very long for a success: $1/0.8 = 1.25$ trials on average. If the probability of a success is low (e.g. 0.1), then we would expect to view many trials before we see a success: $1/0.1 = 10$ trials.

⊙ **Guided Practice 8.31** The probability that an individual would refuse to administer the worst shock is said to be about 0.35. If we were to examine individuals until we found one that did not administer the shock, how many people should we expect to check? The first expression in Equation (30) may be useful.[80]

● **Example 8.32** What is the chance that Dr. Smith will find the first success within the first 4 people?

This is the chance it is the first ($n = 1$), second ($n = 2$), third ($n = 3$), or fourth ($n = 4$) person as the first success, which are four disjoint outcomes. Because the individuals in the sample are randomly sampled from a large population, they are independent. We compute the probability of each case and add the separate results:

$$\begin{aligned} &P(n = 1, 2, 3, \text{ or } 4) \\ &\quad = P(n = 1) + P(n = 2) + P(n = 3) + P(n = 4) \\ &\quad = (0.65)^{1-1}(0.35) + (0.65)^{2-1}(0.35) + (0.65)^{3-1}(0.35) + (0.65)^{4-1}(0.35) \\ &\quad = 0.82 \end{aligned}$$

There is an 82% chance that she will end the study within 4 people.

⊙ **Guided Practice 8.33** Determine a more clever way to solve Example 32. Show that you get the same result.[81]

● **Example 8.34** Suppose in one region it was found that the proportion of people who would administer the worst shock was "only" 55%. If people were randomly selected from this region, what is the expected number of people who must be checked before one was found that would be deemed a success? What is the standard deviation of this waiting time?

A success is when someone will **not** inflict the worst shock, which has probability $p = 1 - 0.55 = 0.45$ for this region. The expected number of people to be checked is $1/p = 1/0.45 = 2.22$ and the standard deviation is $\sqrt{(1-p)/p^2} = 1.65$.

[80] We would expect to see about $1/0.35 = 2.86$ individuals to find the first success.

[81] First find the probability of the complement: $P(\text{no success in first 4 trials}) = 0.65^4 = 0.18$. Next, compute one minus this probability: $1 - P(\text{no success in 4 trials}) = 1 - 0.18 = 0.82$.

Guided Practice 8.35 Using the results from Example 34, $\mu = 2.22$ and $\sigma = 1.65$, would it be appropriate to use the normal model to find what proportion of experiments would end in 3 or fewer trials?[82]

The independence assumption is crucial to the geometric distribution's accurate description of a scenario. Mathematically, we can see that to construct the probability of the success on the n^{th} trial, we had to use the Multiplication Rule for Independent Processes. It is no simple task to generalize the geometric model for dependent trials.

8.4 Binomial Distribution

Example 8.36 Suppose we randomly selected four individuals to participate in the "shock" study. What is the chance exactly one of them will be a success? Let's call the four people Allen (A), Brittany (B), Caroline (C), and Damian (D) for convenience. Also, suppose 35% of people are successes as in the previous version of this example.

Let's consider a scenario where one person refuses:

$$\begin{aligned} &P(A = \texttt{refuse},\ B = \texttt{shock},\ C = \texttt{shock},\ D = \texttt{shock}) \\ &\quad = P(A = \texttt{refuse})\ P(B = \texttt{shock})\ P(C = \texttt{shock})\ P(D = \texttt{shock}) \\ &\quad = (0.35)(0.65)(0.65)(0.65) = (0.35)^1(0.65)^3 = 0.096 \end{aligned}$$

But there are three other scenarios: Brittany, Caroline, or Damian could have been the one to refuse. In each of these cases, the probability is again $(0.35)^1(0.65)^3$. These four scenarios exhaust all the possible ways that exactly one of these four people could refuse to administer the most severe shock, so the total probability is $4 \times (0.35)^1(0.65)^3 = 0.38$.

Guided Practice 8.37 Verify that the scenario where Brittany is the only one to refuse to give the most severe shock has probability $(0.35)^1(0.65)^3$. [83]

[82] No. The geometric distribution is always right skewed and can never be well-approximated by the normal model.

[83] $P(A = \texttt{shock},\ B = \texttt{refuse},\ C = \texttt{shock},\ D = \texttt{shock}) = (0.65)(0.35)(0.65)(0.65) = (0.35)^1(0.65)^3$.

The binomial distribution

The scenario outlined in Example 36 is a special case of what is called the binomial distribution. The **binomial distribution** describes the probability of having exactly k successes in n independent Bernoulli trials with probability of a success p (in Example 36, $n = 4$, $k = 1$, $p = 0.35$). We would like to determine the probabilities associated with the binomial distribution more generally, i.e. we want a formula where we can use n, k, and p to obtain the probability. To do this, we reexamine each part of the example.

There were four individuals who could have been the one to refuse, and each of these four scenarios had the same probability. Thus, we could identify the final probability as

$$[\#\text{ of scenarios}] \times P(\text{single scenario}) \tag{38}$$

The first component of this equation is the number of ways to arrange the $k = 1$ successes among the $n = 4$ trials. The second component is the probability of any of the four (equally probable) scenarios.

Consider $P(\text{single scenario})$ under the general case of k successes and $n - k$ failures in the n trials. In any such scenario, we apply the Multiplication Rule for independent events:

$$p^k(1-p)^{n-k}$$

This is our general formula for $P(\text{single scenario})$.

Secondly, we introduce a general formula for the number of ways to choose k successes in n trials, i.e. arrange k successes and $n - k$ failures:

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

The quantity $\binom{n}{k}$ is read **n choose k**.[84] The exclamation point notation (e.g. $k!$) denotes a **factorial** expression.

$$\begin{aligned}
0! &= 1 \\
1! &= 1 \\
2! &= 2 \times 1 = 2 \\
3! &= 3 \times 2 \times 1 = 6 \\
4! &= 4 \times 3 \times 2 \times 1 = 24 \\
&\vdots \\
n! &= n \times (n-1) \times ... \times 3 \times 2 \times 1
\end{aligned}$$

Using the formula, we can compute the number of ways to choose $k = 1$ successes in $n = 4$ trials:

$$\binom{4}{1} = \frac{4!}{1!(4-1)!} = \frac{4!}{1!3!} = \frac{4 \times 3 \times 2 \times 1}{(1)(3 \times 2 \times 1)} = 4$$

[84] Other notation for n choose k includes ${}_nC_k$, C_n^k, and $C(n, k)$.

This result is exactly what we found by carefully thinking of each possible scenario in Example 36.

Substituting n choose k for the number of scenarios and $p^k(1-p)^{n-k}$ for the single scenario probability in Equation (38) yields the general binomial formula.

Binomial distribution

Suppose the probability of a single trial being a success is p. Then the probability of observing exactly k successes in n independent trials is given by

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \tag{39}$$

Additionally, the mean, variance, and standard deviation of the number of observed successes are

$$\mu = np \qquad \sigma^2 = np(1-p) \qquad \sigma = \sqrt{np(1-p)} \tag{40}$$

TIP: Is it binomial? Four conditions to check.

(1) The trials are independent.
(2) The number of trials, n, is fixed.
(3) Each trial outcome can be classified as a *success* or *failure*.
(4) The probability of a success, p, is the same for each trial.

Example 8.41 What is the probability that 3 of 8 randomly selected students will refuse to administer the worst shock, i.e. 5 of 8 will?

We would like to apply the binomial model, so we check our conditions. The number of trials is fixed ($n = 8$) (condition 2) and each trial outcome can be classified as a success or failure (condition 3). Because the sample is random, the trials are independent (condition 1) and the probability of a success is the same for each trial (condition 4).

In the outcome of interest, there are $k = 3$ successes in $n = 8$ trials, and the probability of a success is $p = 0.35$. So the probability that 3 of 8 will refuse is given by

$$\begin{aligned} \binom{8}{3}(0.35)^3(1-0.35)^{8-3} &= \frac{8!}{3!(8-3)!}(0.35)^3(1-0.35)^{8-3} \\ &= \frac{8!}{3!5!}(0.35)^3(0.65)^5 \end{aligned}$$

Dealing with the factorial part:

$$\frac{8!}{3!5!} = \frac{8\times7\times6\times5\times4\times3\times2\times1}{(3\times2\times1)(5\times4\times3\times2\times1)} = \frac{8\times7\times6}{3\times2\times1} = 56$$

Using $(0.35)^3(0.65)^5 \approx 0.005$, the final probability is about $56 * 0.005 = 0.28$.

TIP: computing binomial probabilities
The first step in using the binomial model is to check that the model is appropriate. The second step is to identify n, p, and k. The final step is to apply the formulas and interpret the results.

TIP: computing n choose k
In general, it is useful to do some cancelation in the factorials immediately. Alternatively, many computer programs and calculators have built in functions to compute n choose k, factorials, and even entire binomial probabilities.

⊙ **Guided Practice 8.42** If you ran a study and randomly sampled 40 students, how many would you expect to refuse to administer the worst shock? What is the standard deviation of the number of people who would refuse? Equation (40) may be useful.[85]

⊙ **Guided Practice 8.43** The probability that a random smoker will develop a severe lung condition in his or her lifetime is about 0.3. If you have 4 friends who smoke, are the conditions for the binomial model satisfied?[86]

⊙ **Guided Practice 8.44** Suppose these four friends do not know each other and we can treat them as if they were a random sample from the population. Is the binomial model appropriate? What is the probability that (a) none of them will develop a severe lung condition? (b) One will develop a severe lung condition? (c) That no more than one will develop a severe lung condition?[87]

⊙ **Guided Practice 8.45** What is the probability that at least 2 of your 4 smoking

[85] We are asked to determine the expected number (the mean) and the standard deviation, both of which can be directly computed from the formulas in Equation (40): $\mu = np = 40 \times 0.35 = 14$ and $\sigma = \sqrt{np(1-p)} = \sqrt{40 \times 0.35 \times 0.65} = 3.02$. Because very roughly 95% of observations fall within 2 standard deviations of the mean (see Section 2.5), we would probably observe at least 8 but less than 20 individuals in our sample who would refuse to administer the shock.

[86] One possible answer: if the friends know each other, then the independence assumption is probably not satisfied. For example, acquaintances may have similar smoking habits.

[87] To check if the binomial model is appropriate, we must verify the conditions. (i) Since we are supposing we can treat the friends as a random sample, they are independent. (ii) We have a fixed number of trials ($n = 4$). (iii) Each outcome is a success or failure. (iv) The probability of a success is the same for each trials since the individuals are like a random sample ($p = 0.3$ if we say a "success" is someone getting a lung condition, a morbid choice). Compute parts (a) and (b) from the binomial formula in Equation (39): $P(0) = \binom{4}{0}(0.3)^0(0.7)^4 = 1 \times 1 \times 0.7^4 = 0.2401$, $P(1) = \binom{4}{1}(0.3)^1(0.7)^3 = 0.4116$. Note: $0! = 1$, as shown on page 95. Part (c) can be computed as the sum of parts (a) and (b): $P(0) + P(1) = 0.2401 + 0.4116 = 0.6517$. That is, there is about a 65% chance that no more than one of your four smoking friends will develop a severe lung condition.

friends will develop a severe lung condition in their lifetimes?[88]

Guided Practice 8.46 Suppose you have 7 friends who are smokers and they can be treated as a random sample of smokers. (a) How many would you expect to develop a severe lung condition, i.e. what is the mean? (b) What is the probability that at most 2 of your 7 friends will develop a severe lung condition.[89]

Next we consider the first term in the binomial probability, n choose k under some special scenarios.

Guided Practice 8.47 Why is it true that $\binom{n}{0} = 1$ and $\binom{n}{n} = 1$ for any number n?[90]

Guided Practice 8.48 How many ways can you arrange one success and $n - 1$ failures in n trials? How many ways can you arrange $n - 1$ successes and one failure in n trials?[91]

Normal approximation to the binomial distribution

The binomial formula is cumbersome when the sample size (n) is large, particularly when we consider a range of observations. In some cases we may use the normal distribution as an easier and faster way to estimate binomial probabilities.

Example 8.49 Approximately 14.5% of the Australian adult population smokes cigarettes [92].

A local government believed their community had a lower smoker rate and commissioned a survey of 400 randomly selected individuals. The survey found that only 39 of the 400 participants smoke cigarettes. If the true proportion of smokers in the community was really 14.5%, what is the probability of observing 39 or fewer smokers in a sample of 400 people?

We leave the usual verification that the four conditions for the binomial model are valid as an exercise.

[88] The complement (no more than one will develop a severe lung condition) as computed in Guided Practice 44 as 0.6517, so we compute one minus this value: 0.3483.

[89] (a) $\mu = 0.3 \times 7 = 2.1$. (b) $P(0, 1, \text{or } 2 \text{ develop severe lung condition}) = P(k = 0) + P(k = 1) + P(k = 2) = 0.6471$.

[90] Frame these expressions into words. How many different ways are there to arrange 0 successes and n failures in n trials? (1 way.) How many different ways are there to arrange n successes and 0 failures in n trials? (1 way.)

[91] One success and $n - 1$ failures: there are exactly n unique places we can put the success, so there are n ways to arrange one success and $n - 1$ failures. A similar argument is used for the second question. Mathematically, we show these results by verifying the following two equations:

$$\binom{n}{1} = n, \qquad \binom{n}{n-1} = n$$

[92] http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/4364.0.55.0012014-15?OpenDocument

The question posed is equivalent to asking, what is the probability of observing $k = 0$, 1, ..., 38, or 39 smokers in a sample of $n = 400$ when $p = 0.145$? We can compute these 40 different probabilities and add them together to find the answer:

$$\begin{aligned} &P(k = 0 \text{ or } k = 1 \text{ or } \cdots \text{ or } k = 39) \\ &\quad = P(k = 0) + P(k = 1) + \cdots + P(k = 39) \\ &\quad = 0.0030 \end{aligned}$$

If the true proportion of smokers in the community is $p = 0.145$, then the probability of observing 39 or fewer smokers in a sample of $n = 400$ is 0.0030. In R this can be calculated as follows:

```
# calculate the probability P(X<=39) when X has Bin(n=400, p=0.145)
# distribution
pbinom(q = 39, size = 400, prob = 0.145, lower.tail = TRUE)

## [1] 0.003025954
```

Although the once tedious computation in Example 49 is now easily performed using the software, there are still occasions when calculating the exact binomial probabilities is best avoided. We might wonder, is it reasonable to use the normal model in place of the binomial distribution? Surprisingly, yes, if certain conditions are met.

⊙ **Guided Practice 8.50** Here we consider the binomial model when the probability of a success is $p = 0.10$. Figure 8.13 shows four hollow histograms for simulated samples from the binomial distribution using four different sample sizes: $n = 10, 30, 100, 300$. What happens to the shape of the distributions as the sample size increases? What distribution does the last hollow histogram resemble?[93]

Normal approximation of the binomial distribution

The binomial distribution with probability of success p is nearly normal when the sample size n is sufficiently large that np and $n(1 - p)$ are both at least 10. The approximate normal distribution has parameters corresponding to the mean and standard deviation of the binomial distribution:

$$\mu = np \qquad\qquad \sigma = \sqrt{np(1 - p)}$$

The normal approximation may be used when computing the range of many possible successes. For instance, we may apply the normal distribution to the setting of Example 49.

[93] The distribution is transformed from a blocky and skewed distribution into one that rather resembles the normal distribution in last hollow histogram

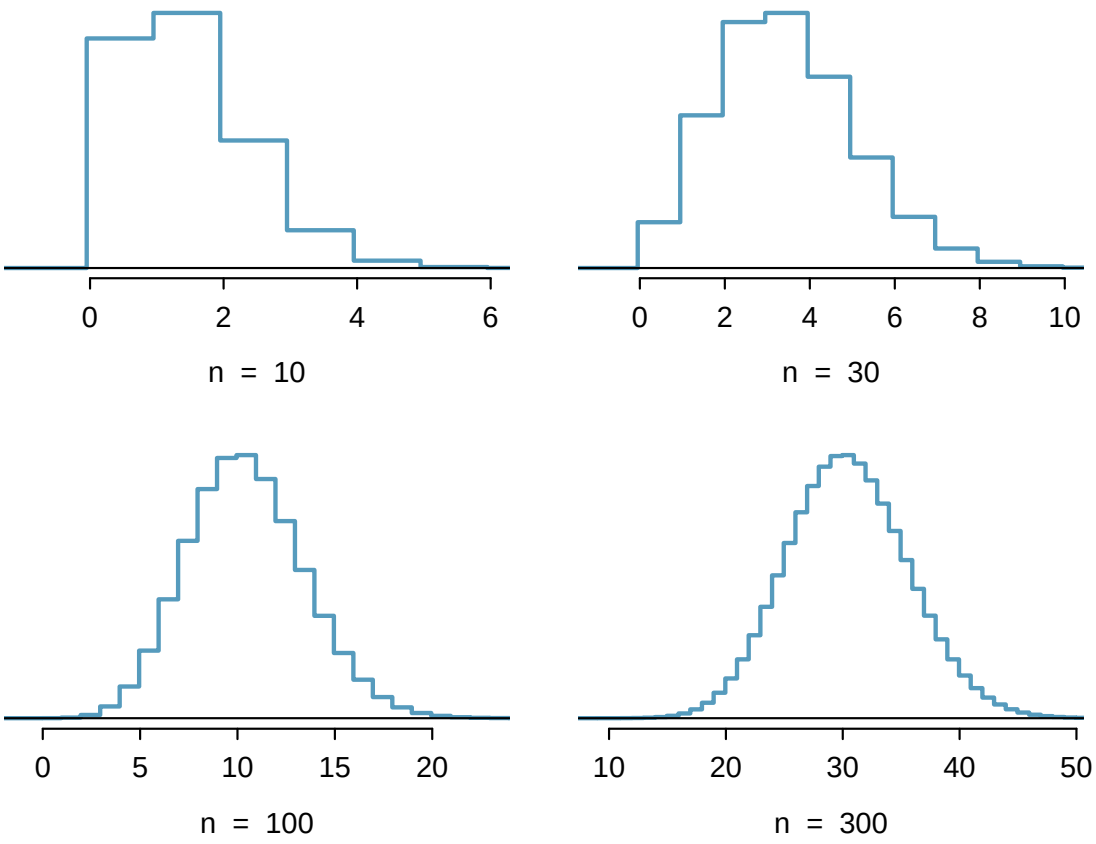


Figure 8.13: Hollow histograms of samples from the binomial model when $p = 0.10$. The sample sizes for the four plots are $n = 10$, 30, 100, and 300, respectively.

Example 8.51 How can we use the normal approximation to estimate the probability of observing 39 or fewer smokers in a sample of 400, if the true proportion of smokers is $p = 0.145$?

Showing that the binomial model is reasonable was a suggested exercise in Example 49. We also verify that both np and $n(1-p)$ are at least 10:

$$np = 400 \times 0.145 = 58 \qquad n(1-p) = 400 \times 0.855 = 342$$

With these conditions checked, we may use the normal approximation in place of the binomial distribution using the mean and standard deviation from the binomial model:

$$\mu = np = 58 \qquad \sigma = \sqrt{np(1-p)} = 7.04$$

We want to find the probability of observing fewer than 39 smokers using this model.

Guided Practice 8.52 Use the normal model $N(\mu = 58, \sigma = 7.04)$ to estimate the probability of observing fewer than 39 smokers. Your answer should be approximately

equal to the solution of Example 49: 0.0030.[94]

[94]Compute the Z-score first: $Z = \frac{39-58}{7.04} = -2.70$. The corresponding left tail area is 0.0035.

9 Inference [optional technical background]

Statistical inference is concerned primarily with understanding the quality of parameter estimates. For example, a classic inferential question is, "How sure are we that the estimated mean, $\bar{x}$, is near the true population mean, μ?" While the equations and details change depending on the setting, the foundations for inference are the same throughout all of statistics. We introduce these common themes in Sections 9.1-9.4 by discussing inference about the population mean, μ.

Throughout the next few sections we consider a data set called `yrbss`, which represents all 13,583 high school students in the Youth Risk Behavior Surveillance System (YRBSS) from 2013.[95] Part of this data set is shown in Table 9.1, and the variables are described in Table 9.2.

Table 9.1: Five cases from the `yrbss` data set. Some observations are blank since there are missing data. For example, the height and weight of students 1 and 2 are missing.

ID	age	gender	grade	height	weight	helmet	active	lifting
1	14	female	9			never	4	0
2	14	female	9			never	2	0
3	15	female	9	1.73	84.37	never	7	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
13582	17	female	12	1.60	77.11	sometimes	5	
13583	17	female	12	1.57	52.16	did not ride	5	

Table 9.2: Variables and their descriptions for the `yrbss` data set.

age	**Age of the student.**
`gender`	Sex of the student.
`grade`	Grade in high school
`height`	Height, in meters. There are 3.28 feet in a meter.
`weight`	Weight, in kilograms (2.2 pounds per kilogram).
`helmet`	Frequency that the student wore a helmet while biking in the last 12 months.
`active`	Number of days physically active for 60+ minutes in the last 7 days.
`lifting`	Number of days of strength training (e.g. lifting weights) in the last 7 days.

We're going to consider the population of high school students who participated in the 2013 YRBSS. We took a simple random sample of this population, which is represented in Table 9.3. We will use this sample, which we refer to as the `yrbss_samp` data set, to draw conclusions about the population of YRBSS participants. This is the practice of statistical

[95] www.cdc.gov/healthyyouth/data/yrbs/data.htm

Table 9.3: Four observations for the `yrbss_samp` data set, which represents a simple random sample of 100 high schoolers from the 2013 YRBSS.

ID	age	gender	grade	height	weight	helmet	active	lifting
5653	16	female	11	1.50	52.62	never	0	0
9437	17	male	11	1.78	74.84	rarely	7	5
2021	17	male	11	1.75	106.60	never	7	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2325	14	male	9	1.70	55.79	never	1	0

inference in the broadest sense. Two histograms summarizing the `height`, `weight`, `active`, and `lifting` variables from `yrbss_samp` data set are shown in Figure 9.1.

9.1 Variability in Estimates

We would like to estimate four features of the high schoolers in YRBSS using the sample.

(1) What is the average height of the YRBSS high schoolers?

(2) What is the average weight of the YRBSS high schoolers?

(3) On average, how many days per week are YRBSS high schoolers physically active?

(4) On average, how many days per week do YRBSS high schoolers do weight training?

While we focus on the mean in this chapter, questions regarding variation are often just as important in practice. For instance, if students are either very active or almost entirely inactive (the distribution is bimodal), we might try different strategies to promote a healthy lifestyle among students than if all high schoolers were already somewhat active.

Point estimates

We want to estimate the **population mean** based on the sample. The most intuitive way to go about doing this is to simply take the **sample mean**. That is, to estimate the average height of all YRBSS students, take the average height for the sample:

$$\bar{x}_{height} = \frac{1.50 + 1.78 + \cdots + 1.70}{100} = 1.697$$

The sample mean $\bar{x} = 1.697$ meters (5 feet, 6.8 inches) is called a **point estimate** of the population mean: if we can only choose one value to estimate the population mean, this is our best guess. Suppose we take a new sample of 100 people and recompute the mean; we will probably not get the exact same answer that we got using the `yrbss_samp` data set. Estimates generally vary from one sample to another, and this **sampling variation** suggests our estimate may be close, but it will not be exactly equal to the parameter.

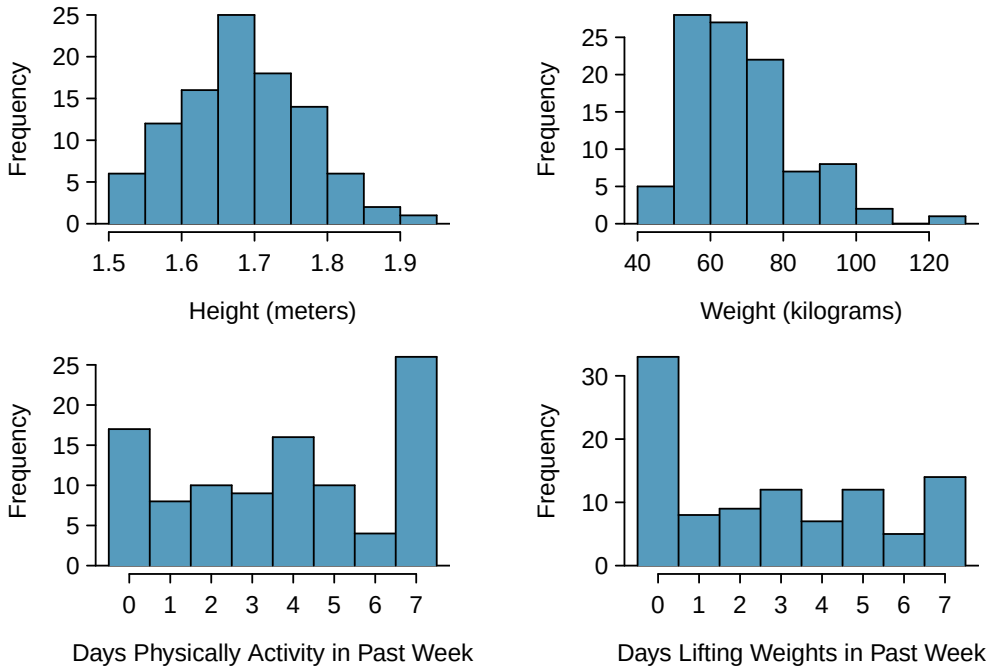


Figure 9.1: Histograms of `height`, `weight`, `activity`, and `lifting` for the sample YRBSS data. The `height` distribution is approximately symmetric, `weight` is moderately skewed to the right, `activity` is bimodal or multimodal (with unclear skew), and `lifting` is strongly right skewed.

We can also estimate the average weight of YRBSS respondents by examining the sample mean of `weight` (in kg), and average number of days physically active in a week:

$$\bar{x}_{weight} = \frac{52.6 + 74.8 + \cdots + 55.8}{100} = 68.89 \qquad \bar{x}_{active} = \frac{0 + 7 + \cdots + 1}{100} = 3.75$$

The average weight is 68.89 kilograms, which is about 151.6 pounds.

What about generating point estimates of other **population parameters**, such as the population median or population standard deviation? Once again we might estimate parameters based on sample statistics, as shown in Table 9.4. For example, the population standard deviation of `active` using the sample standard deviation, 2.56 days.

Table 9.4: Point estimates and parameter values for the `active` variable. The parameters were obtained by computing the mean, median, and SD for all YRBSS respondents.

`active`	estimate	parameter
mean	3.75	3.90
median	4.00	4.00
st. dev.	2.556	2.564

Point estimates are not exact

Estimates are usually not exactly equal to the truth, but they get better as more data become available. We can see this by plotting a running mean from `yrbss_samp`. A **running mean** is a sequence of means, where each mean uses one more observation in its calculation than the mean directly before it in the sequence. For example, the second mean in the sequence is the average of the first two observations and the third in the sequence is the average of the first three. The running mean for the `active` variable in the `yrbss_samp` is shown in Figure 9.2, and it approaches the true population average, 3.90 days, as more data become available.

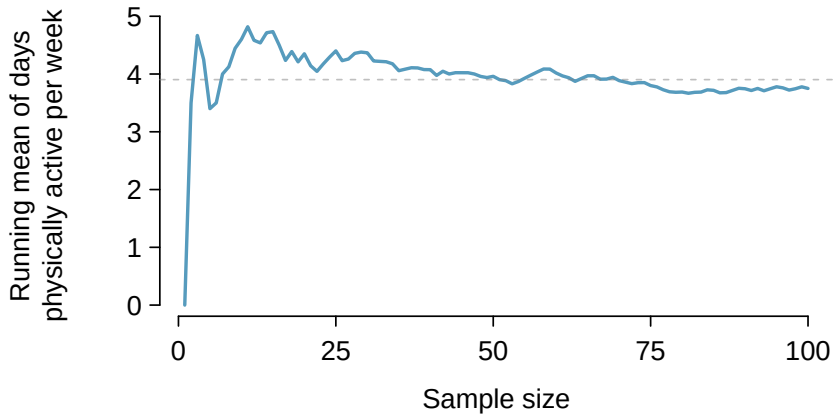


Figure 9.2: The mean computed after adding each individual to the sample. The mean tends to approach the true population average as more data become available.

Sample point estimates only approximate the population parameter, and they vary from one sample to another. If we took another simple random sample of the YRBSS students, we would find that the sample mean for the number of days active would be a little different. It will be useful to quantify how variable an estimate is from one sample to another. If this variability is small (i.e. the sample mean doesn't change much from one sample to another) then that estimate is probably very accurate. If it varies widely from one sample to another, then we should not expect our estimate to be very good.

Standard error of the mean

From the random sample represented in `yrbss_samp`, we guessed the average number of days a YRBSS student is physically active is 3.75 days. Suppose we take another random sample of 100 individuals and take its mean: 3.22 days. Suppose we took another (3.67 days) and another (4.10 days), and so on. If we do this many many times – which we can do only because we have all YRBSS students – we can build up a **sampling distribution** for the sample mean when the sample size is 100, shown in Figure 9.3.

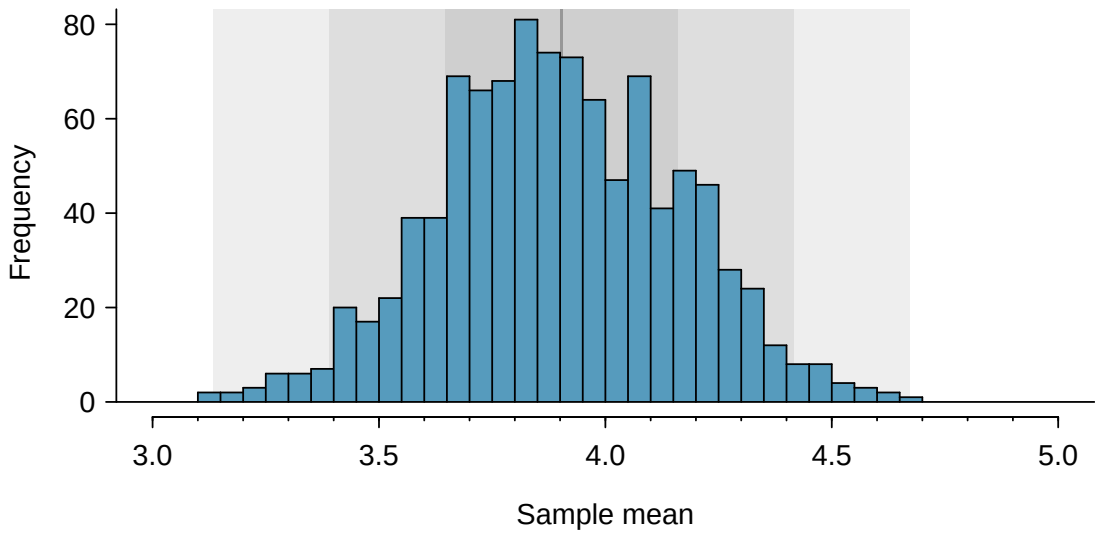


Figure 9.3: A histogram of 1000 sample means for number of days physically active per week, where the samples are of size $n = 100$.

Sampling distribution

The sampling distribution represents 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 such a distribution. Understanding the concept of a sampling distribution is central to understanding statistical inference.

The sampling distribution shown in Figure 9.3 is unimodal and approximately symmetric. It is also centred exactly at the true population mean: $\mu = 3.90$. Intuitively, this makes sense. The sample means should tend to "fall around" the population mean.

We can see that the sample mean has some variability around the population mean, which can be quantified using the standard deviation of this distribution of sample means: $\sigma_{\bar{x}} = 0.26$. The standard deviation of the sample mean tells us how far the typical estimate is away from the actual population mean, 3.90 days. It also describes the typical **error** of the point estimate, and for this reason we usually call this standard deviation the **standard error (SE)** of the estimate.

SE
standard error

Standard error of an estimate

The standard deviation associated with an estimate is called the *standard error*. It describes the typical error or uncertainty associated with the estimate.

When considering the case of the point estimate $\bar{x}$, there is one problem: there is no obvious way to estimate its standard error from a single sample. However, statistical theory provides a helpful tool to address this issue.

⊙ Guided Practice 9.1 (a) Would you rather use a small sample or a large sample when estimating a parameter? Why? (b) Using your reasoning from (a), would you expect a point estimate based on a small sample to have smaller or larger standard error than a point estimate based on a larger sample?[96]

In the sample of 100 students, the standard error of the sample mean is equal to the population standard deviation divided by the square root of the sample size:

$$SE_{\bar{x}} = \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{2.6}{\sqrt{100}} = 0.26$$

where σ_x is the standard deviation of the individual observations. This is no coincidence. We can show mathematically that this equation is correct when the observations are independent.

Computing SE for the sample mean

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

$$SE = \frac{\sigma}{\sqrt{n}} \tag{2}$$

A reliable method to ensure sample observations are independent is to conduct a simple random sample consisting of less than 10% of the population.

There is one subtle issue in Equation (2): the population standard deviation is typically unknown. You might have already guessed how to resolve this problem: we can use the point estimate of the standard deviation from the sample. This estimate tends to be sufficiently good when the sample size is at least 30 and the population distribution is not strongly skewed. Thus, we often just use the sample standard deviation s instead of σ. When the sample size is smaller than 30, we will need to use a method to account for extra uncertainty in the standard error. If the skew condition is not met, a larger sample is needed to compensate for the extra skew. These topics are further discussed in Section 9.4.

⊙ Guided Practice 9.3 In the sample of 100 students, the standard deviation of student heights is $s_{height} = 0.088$ meters. In this case, we can confirm that the observations are independent by checking that the data come from a simple random sample consisting of less than 10% of the population. (a) What is the standard error of the sample mean, $\bar{x}_{height} = 1.70$ meters? (b) Would you be surprised if someone told you the average height of all YRBSS respondents was actually 1.69 meters?[97]

[96](a) Consider two random samples: one of size 10 and one of size 1000. Individual observations in the small sample are highly influential on the estimate while in larger samples these individual observations would more often average each other out. The larger sample would tend to provide a more accurate estimate. (b) If we think an estimate is better, we probably mean it typically has less error. Based on (a), our intuition suggests that a larger sample size corresponds to a smaller standard error.

[97](a) Use Equation (2) with the sample standard deviation to compute the standard error: $SE_{\bar{y}} = 0.088/\sqrt{100} = 0.0088$ meters. (b) It would not be surprising. Our sample is about 1 standard error from

⊙ **Guided Practice 9.4** (a) Would you be more trusting of a sample that has 100 observations or 400 observations? (b) We want to show mathematically that our estimate tends to be better when the sample size is larger. If the standard deviation of the individual observations is 10, what is our estimate of the standard error when the sample size is 100? What about when it is 400? (c) Explain how your answer to part (b) mathematically justifies your intuition in part (a).[98]

Basic properties of point estimates

We achieved three goals in this section. 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 (2). While we could also quantify the standard error for other estimates – such as the median, standard deviation, or any other number of statistics – these are beyond the scope of the current course.

9.2 Confidence Intervals

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.

Capturing the population 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.

⊙ **Guided Practice 9.5** If we want to be very certain we capture the population parameter, should we use a wider interval or a smaller interval?[99]

1.69m. In other words, 1.69m does not seem to be implausible given that our sample was relatively close to it. (We use the standard error to identify what is close.)

[98](a) Extra observations are usually helpful in understanding the population, so a point estimate with 400 observations seems more trustworthy. (b) The standard error when the sample size is 100 is given by $SE_{100} = 10/\sqrt{100} = 1$. For 400: $SE_{400} = 10/\sqrt{400} = 0.5$. The larger sample has a smaller standard error. (c) The standard error of the sample with 400 observations is lower than that of the sample with 100 observations. The standard error describes the typical error, and since it is lower for the larger sample, this mathematically shows the estimate from the larger sample tends to be better – though it does not guarantee that every large sample will provide a better estimate than a particular small sample.

[99]If we want to be more certain we will capture the fish, we might use a wider net. Likewise, we use a

An approximate 95% confidence interval

Our point estimate is the most plausible value of the parameter, so it makes sense to build the confidence interval around the point estimate. The standard error, which is a measure of the uncertainty associated with the point estimate, provides a guide for how large we should make the confidence interval.

The standard error represents the standard deviation associated with the estimate, and roughly 95% of the time the estimate will be within 2 standard errors of the parameter. If the interval spreads out 2 standard errors from the point estimate, we can be roughly 95% **confident** that we have captured the true parameter:

$$\text{point estimate} \ \pm \ 2 \times SE \tag{6}$$

But what does "95% confident" mean? Suppose we took many samples and built a confidence interval from each sample using Equation (6). Then about 95% of those intervals would contain the actual mean, μ. Figure 9.4 shows this process with 25 samples, where 24 of the resulting confidence intervals contain the average number of days per week that YRBSS students are physically active, $\mu = 3.90$ days, and one interval does not.

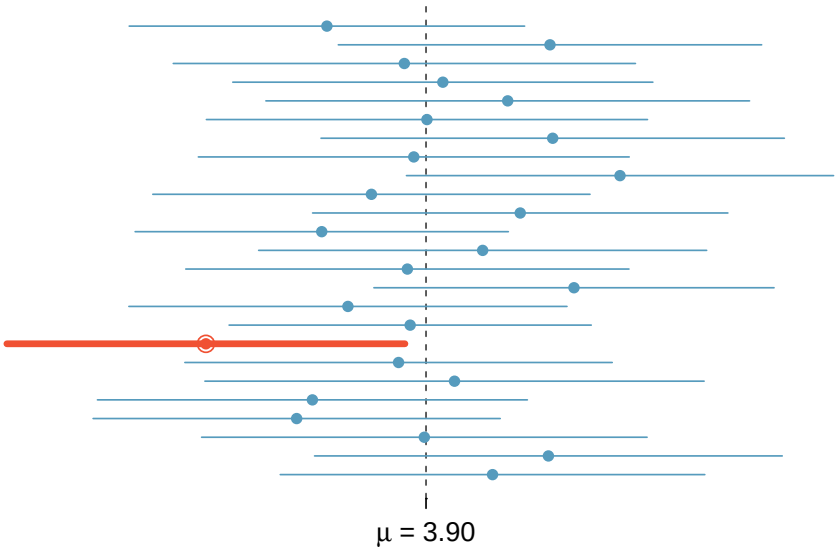


Figure 9.4: Twenty-five samples of size $n = 100$ were taken from `yrbss`. For each sample, a confidence interval was created to try to capture the average number of days per week that students are physically active. Only 1 of these 25 intervals did not capture the true mean, $\mu = 3.90$ days.

Guided Practice 9.7 In Figure 9.4, one interval does not contain 3.90 minutes. Does this imply that the mean cannot be 3.90?[100]

wider confidence interval if we want to be more certain that we capture the parameter.

[100] Just as some observations occur more than 2 standard deviations from the mean, some point estimates will be more than 2 standard errors from the parameter. A confidence interval only provides a plausible range of values for a parameter. While we might say other values are implausible based on the data, this does not mean they are impossible.

The rule where about 95% of observations are within 2 standard deviations of the mean is only approximately true. However, it holds very well for the normal distribution. As we will soon see, the mean tends to be normally distributed when the sample size is sufficiently large.

Example 9.8 The sample mean of days active per week from `yrbss_samp` is 3.75 days. The standard error, as estimated using the sample standard deviation, is $SE = \frac{2.6}{\sqrt{100}} = 0.26$ days. (The population SD is unknown in most applications, so we use the sample SD here.) Calculate an approximate 95% confidence interval for the average days active per week for all YRBSS students.

We apply Equation (6):

$$3.75 \ \pm \ 2 \times 0.26 \quad \rightarrow \quad (3.23, 4.27)$$

Based on these data, we are about 95% confident that the average days active per week for all YRBSS students was larger than 3.23 but less than 4.27 days. Our interval extends out 2 standard errors from the point estimate, $\bar{x}_{active}$.

Guided Practice 9.9 The sample data suggest the average YRBSS student height is $\bar{x}_{height} = 1.697$ meters with a standard error of 0.0088 meters (estimated using the sample standard deviation, 0.088 meters). What is an approximate 95% confidence interval for the average height of all of the YRBSS students?[101]

The sampling distribution for the mean

In Section 9.1, we introduced a sampling distribution for $\bar{x}$, the average days physically active per week for samples of size 100. We examined this distribution earlier in Figure 9.3. Now we'll take 100,000 samples, calculate the mean of each, and plot them in a histogram to get an especially accurate depiction of the sampling distribution. This histogram is shown in the left panel of Figure 9.5.

Does this distribution look familiar? Hopefully so! The distribution of sample means closely resembles the normal distribution. A normal probability plot of these sample means is shown in the right panel of Figure 9.5. Because all of the points closely fall around a straight line, we can conclude the distribution of sample means is nearly normal. This result can be explained by 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.

[101] Apply Equation (6): $1.697 \ \pm \ 2 \times 0.0088 \rightarrow (1.6794, 1.7146)$. We interpret this interval as follows: We are about 95% confident the average height of all YRBSS students was between 1.6794 and 1.7146 meters (5.51 to 5.62 feet).

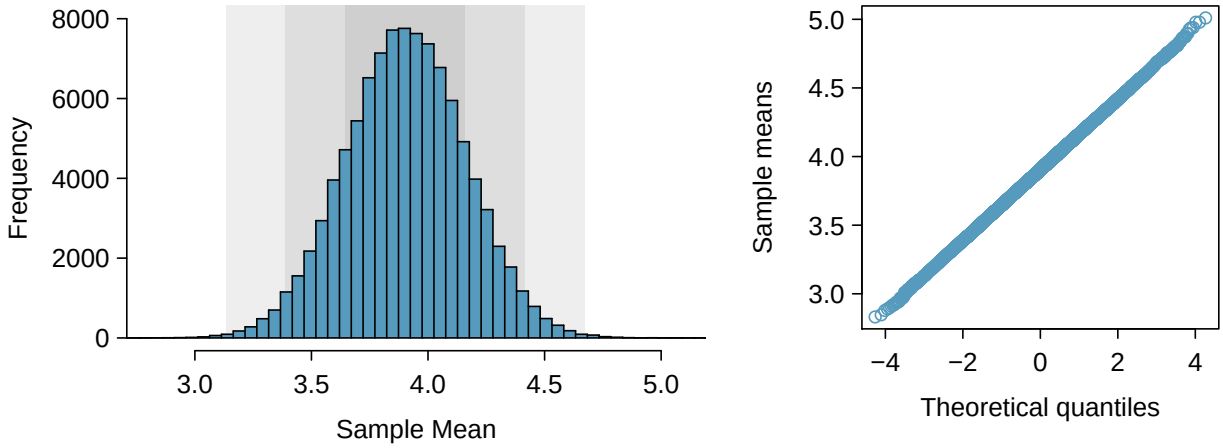


Figure 9.5: The left panel shows a histogram of the sample means for 100,000 different random samples. The right panel shows a normal probability plot of those sample means.

We will apply this informal version of the Central Limit Theorem for now, and discuss its details further in Section 9.4.

The choice of using 2 standard errors in Equation (6) was based on our general guideline that roughly 95% of the time, observations are within two standard deviations of the mean. Under the normal model, we can make this more accurate by using 1.96 in place of 2.

$$\text{point estimate} \ \pm \ 1.96 \times SE \tag{10}$$

If a point estimate, such as $\bar{x}$, is associated with a normal model and standard error SE, then we use this more precise 95% confidence interval.

Interpreting confidence intervals

A careful eye might have observed the somewhat awkward language used to describe confidence intervals. Correct interpretation:

> We are XX% confident that the population parameter is between...

Incorrect language might try to describe the confidence interval as capturing the population parameter with a certain probability. This is a common error: while it might be useful to think of it as a probability, the confidence level only quantifies how plausible it is that the parameter is in the interval.

Another important consideration of confidence intervals is that they *only try to capture the population parameter*. A confidence interval says nothing about the confidence of capturing individual observations, a proportion of the observations, or about capturing point estimates. Confidence intervals only attempt to capture population parameters.

9.3 Hypothesis Testing

Are students lifting weights or performing other strength training exercises more or less often than they have in the past? We'll compare data from students from the 2011 YRBSS survey to our sample of 100 students from the 2013 YRBSS survey.

We'll also consider sleep behavior. A recent study found that college students average about 7 hours of sleep per night.[102] However, researchers at a rural college are interested in showing that their students sleep longer than seven hours on average. We investigate this topic in Section 9.3.

Hypothesis testing framework

Students from the 2011 YRBSS lifted weights (or performed other strength training exercises) 3.09 days per week on average. We want to determine if the `yrbss_samp` data set provides strong evidence that YRBSS students selected in 2013 are lifting more or less than the 2011 YRBSS students, versus the other possibility that there has been no change.[103] We simplify these three options into two competing **hypotheses**:

H_0: The average days per week that YRBSS students lifted weights was the same for 2011 and 2013.

H_A: The average days per week that YRBSS students lifted weights was *different* for 2013 than in 2011.

We call H_0 the null hypothesis and H_A the alternative hypothesis.

Null and alternative hypotheses

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

The null hypothesis often represents a sceptical position or a perspective of no difference. The alternative hypothesis often represents a new perspective, such as the possibility that there has been a change.

TIP: Hypothesis testing framework

The sceptic will not reject the null hypothesis (H_0), unless the evidence in favour of the alternative hypothesis (H_A) is so strong that she rejects H_0 in favour of H_A.

[102] *Poll shows college students get least amount of sleep.* theloquitur.com/?p=1161

[103] While we could answer this question by examining the entire YRBSS data set from 2013 (`yrbss`), we only consider the sample data (`yrbss_samp`), which is more realistic since we rarely have access to population data.

The hypothesis testing framework is a very general tool, and we often use it without a second thought. If a person makes a somewhat unbelievable claim, we are initially sceptical. However, if there is sufficient evidence that supports the claim, we set aside our scepticism and reject the null hypothesis in favour of the alternative. The hallmarks of hypothesis testing are also found in the court system.

⊙ **Guided Practice 9.11** A court considers two possible claims about a defendant: she is either innocent or guilty. If we set these claims up in a hypothesis framework, which would be the null hypothesis and which the alternative?[104]

Jurors examine the evidence to see whether it convincingly shows a defendant is guilty. Even if the jurors leave unconvinced of guilt beyond a reasonable doubt, this does not mean they believe the defendant is innocent. This is also the case with hypothesis testing: *even if we fail to reject the null hypothesis, we typically do not accept the null hypothesis as true.* Failing to find strong evidence for the alternative hypothesis is not equivalent to accepting the null hypothesis.

In the example with the YRBSS, the null hypothesis represents no difference in the average days per week of weight lifting in 2011 and 2013. The alternative hypothesis represents something new or more interesting: there was a difference, either an increase or a decrease. These hypotheses can be described in mathematical notation using μ_{13} as the average days of weight lifting for 2013:

H_0: $\mu_{13} = 3.09$

H_A: $\mu_{13} \neq 3.09$

where 3.09 is the average number of days per week that students from the 2011 YRBSS lifted weights. Using the mathematical notation, the hypotheses can more easily be evaluated using statistical tools. We call 3.09 the **null value** since it represents the value of the parameter if the null hypothesis is true.

Decision errors

Hypothesis tests are not flawless, since we can make a wrong decision in statistical hypothesis tests based on the data. For example, in the court system innocent people are sometimes wrongly convicted and the guilty sometimes walk free. However, the difference is that in statistical hypothesis tests, we have the tools necessary to quantify how often we make such errors.

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a statement about which one might be true, but we might choose incorrectly. There are four possible scenarios, which are summarized in Table 9.5.

A **Type 1 Error** is rejecting the null hypothesis when H_0 is actually true. A **Type 2 Error** is failing to reject the null hypothesis when the alternative is actually true.

[104]The jury considers whether the evidence is so convincing (strong) that there is no reasonable doubt regarding the person's guilt; in such a case, the jury rejects innocence (the null hypothesis) and concludes the defendant is guilty (alternative hypothesis).

Table 9.5: Four different scenarios for hypothesis tests.

		Test conclusion	
		do not reject H_0	reject H_0 in favour of H_A
Truth	H_0 true	okay	Type 1 Error
	H_A true	Type 2 Error	okay

⊙ **Guided Practice 9.12** In a court, the defendant is either innocent (H_0) or guilty (H_A). What does a Type 1 Error represent in this context? What does a Type 2 Error represent? Table 9.5 may be useful.[105]

⊙ **Guided Practice 9.13** How could we reduce the Type 1 Error rate in courts? What influence would this have on the Type 2 Error rate?[106]

⊙ **Guided Practice 9.14** How could we reduce the Type 2 Error rate in courts? What influence would this have on the Type 1 Error rate?[107]

Exercises 12-14 provide an important lesson: if we reduce how often we make one type of error, we generally make more of the other type.

Hypothesis testing is built around rejecting or failing to reject the null hypothesis. That is, we do not reject H_0 unless we have strong evidence. But what precisely does *strong evidence* mean? As a general rule of thumb, for those cases where the null hypothesis is actually true, we do not want to incorrectly reject H_0 more than 5% of the time. This corresponds to a **significance level** of 0.05. We often write the significance level using α (the Greek letter *alpha*): $\alpha = 0.05$.

α significance level of a hypothesis test

If we use a 95% confidence interval to evaluate a hypothesis test where the null hypothesis is true, we will make an error whenever the point estimate is at least 1.96 standard errors away from the population parameter. This happens about 5% of the time (2.5% in each tail). Similarly, using a 99% confidence interval to evaluate a hypothesis is equivalent to a significance level of $\alpha = 0.01$.

A confidence interval is, in one sense, simplistic in the world of hypothesis tests. Consider the following two scenarios:

- The null value (the parameter value under the null hypothesis) is in the 95% confidence interval but just barely, so we would not reject H_0. However, we might like to somehow

[105] If the court makes a Type 1 Error, this means the defendant is innocent (H_0 true) but wrongly convicted. A Type 2 Error means the court failed to reject H_0 (i.e. failed to convict the person) when she was in fact guilty (H_A true).

[106] To lower the Type 1 Error rate, we might raise our standard for conviction from "beyond a reasonable doubt" to "beyond a conceivable doubt" so fewer people would be wrongly convicted. However, this would also make it more difficult to convict the people who are actually guilty, so we would make more Type 2 Errors.

[107] To lower the Type 2 Error rate, we want to convict more guilty people. We could lower the standards for conviction from "beyond a reasonable doubt" to "beyond a little doubt". Lowering the bar for guilt will also result in more wrongful convictions, raising the Type 1 Error rate.

say, quantitatively, that it was a close decision.

- The null value is very far outside of the interval, so we reject H_0. However, we want to communicate that, not only did we reject the null hypothesis, but it wasn't even close. Such a case is depicted in Figure 9.6.

In Section 9.3, we introduce a tool called the *p-value* that will be helpful in these cases. The p-value method also extends to hypothesis tests where confidence intervals cannot be easily constructed or applied.

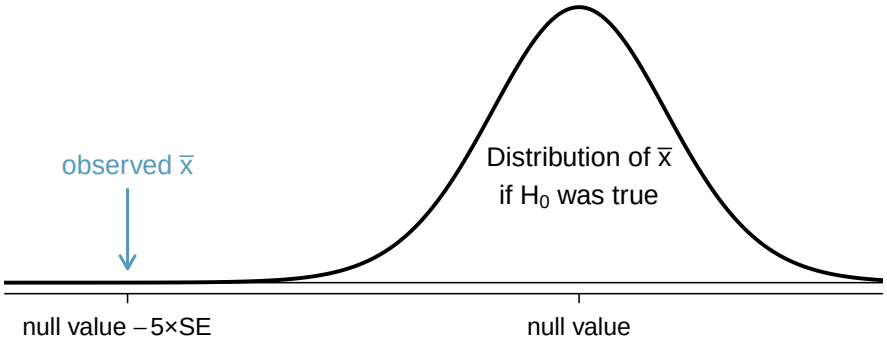


Figure 9.6: It would be helpful to quantify the strength of the evidence against the null hypothesis. In this case, the evidence is extremely strong.

Formal testing using p-values

The p-value is a way of quantifying the strength of the evidence against the null hypothesis and in favour of the alternative. Formally the *p-value* is a conditional probability.

p-value

The **p-value** is the probability of observing data at least as favourable to the alternative hypothesis as our current data set, if the null hypothesis is true. We typically use a summary statistic of the data, in this chapter the sample mean, to help compute the p-value and evaluate the hypotheses.

⊙ **Guided Practice 9.15** A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. Researchers at a rural school are interested in showing that students at their school sleep longer than seven hours on average, and they would like to demonstrate this using a sample of students. What would be an appropriate sceptical position for this research?[108]

[108] A sceptic would have no reason to believe that sleep patterns at this school are different than the sleep patterns at another school.

We can set up the null hypothesis for this test as a sceptical perspective: the students at this school average 7 hours of sleep per night. The alternative hypothesis takes a new form reflecting the interests of the research: the students average more than 7 hours of sleep. We can write these hypotheses as

H_0: $\mu = 7$.

H_A: $\mu > 7$.

Using $\mu > 7$ as the alternative is an example of a **one-sided** hypothesis test. In this investigation, there is no apparent interest in learning whether the mean is less than 7 hours.[109] Earlier we encountered a **two-sided** hypothesis where we looked for any clear difference, greater than or less than the null value.

Always use a two-sided test unless it was made clear prior to data collection that the test should be one-sided. Switching a two-sided test to a one-sided test after observing the data is dangerous because it can inflate the Type 1 Error rate.

TIP: One-sided and two-sided tests

When you are interested in checking for an increase or a decrease, but not both, use a one-sided test. When you are interested in any difference from the null value – an increase or decrease – then the test should be two-sided.

TIP: Always write the null hypothesis as an equality

We will find it most useful if we always list the null hypothesis as an equality (e.g. $\mu = 7$) while the alternative always uses an inequality (e.g. $\mu \neq 7$, $\mu > 7$, or $\mu < 7$).

The researchers at the rural school conducted a simple random sample of $n = 110$ students on campus. They found that these students averaged 7.42 hours of sleep and the standard deviation of the amount of sleep for the students was 1.75 hours. A histogram of the sample is shown in Figure 9.7.

Before we can use a normal model for the sample mean or compute the standard error of the sample mean, we must verify conditions. (1) Because this is a simple random sample from less than 10% of the student body, the observations are independent. (2) The sample size in the sleep study is sufficiently large since it is greater than 30. (3) The data show strong skew in Figure 9.7 and the presence of a couple of outliers. This skew and the outliers are acceptable for a sample size of $n = 110$. With these conditions verified, the normal model can be safely applied to $\bar{x}$ and we can reasonably calculate the standard error.

⊙ **Guided Practice 9.16** In the sleep study, the sample standard deviation was 1.75 hours and the sample size is 110. Calculate the standard error of $\bar{x}$.[110]

[109] This is entirely based on the interests of the researchers. Had they been only interested in the opposite case – showing that their students were actually averaging fewer than seven hours of sleep but not interested in showing more than 7 hours – then our setup would have set the alternative as $\mu < 7$.

[110] The standard error can be estimated from the sample standard deviation and the sample size: $SE_{\bar{x}} = \frac{s_x}{\sqrt{n}} = \frac{1.75}{\sqrt{110}} = 0.17$.

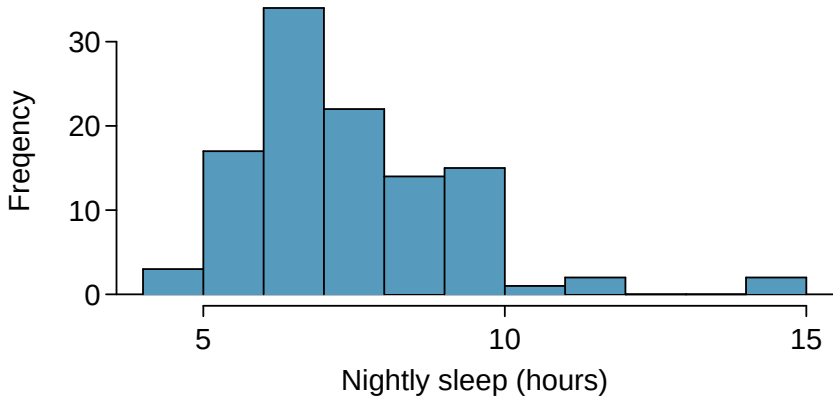


Figure 9.7: Distribution of a night of sleep for 110 college students. These data are strongly skewed.

The hypothesis test for the sleep study will be evaluated using a significance level of $\alpha = 0.05$. We want to consider the data under the scenario that the null hypothesis is true. In this case, the sample mean is from a distribution that is nearly normal and has mean 7 and standard deviation of about $SE_{\bar{x}} = 0.17$. Such a distribution is shown in Figure 9.8.

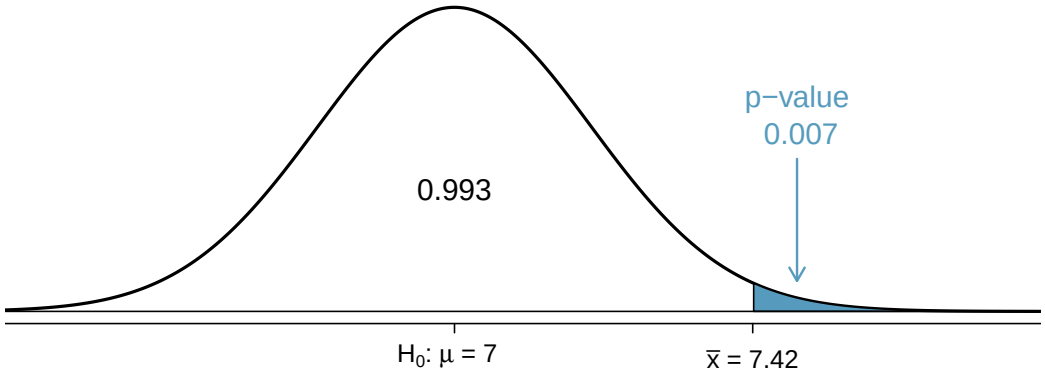


Figure 9.8: If the null hypothesis is true, then the sample mean $\bar{x}$ came from this nearly normal distribution. The right tail describes the probability of observing such a large sample mean if the null hypothesis is true.

The shaded tail in Figure 9.8 represents the chance of observing such a large mean, conditional on the null hypothesis being true. That is, the shaded tail represents the p-value. We shade all means larger than our sample mean, $\bar{x} = 7.42$, because they are more favourable to the alternative hypothesis than the observed mean.

We compute the p-value by finding the tail area of this normal distribution. First compute the Z-score of the sample mean, $\bar{x} = 7.42$:

$$Z = \frac{\bar{x} - \text{null value}}{SE_{\bar{x}}} = \frac{7.42 - 7}{0.17} = 2.47$$

Using the normal probability table, the lower unshaded area is found to be 0.993. Thus the

shaded area is $1 - 0.993 = 0.007$. *If the null hypothesis is true, the probability of observing a sample mean at least as large as 7.42 hours for a sample of 110 students is only 0.007.* That is, if the null hypothesis is true, we would not often see such a large mean.

We evaluate the hypotheses by comparing the p-value to the significance level. Because the p-value is less than the significance level (p-value $= 0.007 < 0.05 = \alpha$), we reject the null hypothesis. What we observed is so unusual with respect to the null hypothesis that it casts serious doubt on H_0 and provides strong evidence favouring H_A.

p-value as a tool in hypothesis testing

The smaller the p-value, the stronger the data favour H_A over H_0. A small p-value (usually < 0.05) corresponds to sufficient evidence to reject H_0 in favour of H_A.

TIP: It is useful to first draw a picture to find the p-value

It is useful to draw a picture of the distribution of $\bar{x}$ as though H_0 was true (i.e. μ equals the null value), and shade the region (or regions) of sample means that are at least as favourable to the alternative hypothesis. These shaded regions represent the p-value.

The ideas below review the process of evaluating hypothesis tests with p-values:

- The null hypothesis represents a sceptic's position or a position of no difference. We reject this position only if the evidence strongly favours H_A.
- A small p-value means that if the null hypothesis is true, there is a low probability of seeing a point estimate at least as extreme as the one we saw. We interpret this as strong evidence in favour of the alternative.
- We reject the null hypothesis if the p-value is smaller than the significance level, α, which is usually 0.05. Otherwise, we fail to reject H_0.
- We should always state the conclusion of the hypothesis test in plain language so non-statisticians can also understand the results.

The p-value is constructed in such a way that we can directly compare it to the significance level (α) to determine whether or not to reject H_0. This method ensures that the Type 1 Error rate does not exceed the significance level standard.

⊙ **Guided Practice 9.17** If the null hypothesis is true, how often should the p-value be less than 0.05?[111]

⊙ **Guided Practice 9.18** Suppose we had used a significance level of 0.01 in the sleep study. Would the evidence have been strong enough to reject the null hypothesis? (The p-value was 0.007.) What if the significance level was $\alpha = 0.001$? [112]

[111] About 5% of the time. If the null hypothesis is true, then the data only has a 5% chance of being in the 5% of data most favourable to H_A.

[112] We reject the null hypothesis whenever $p\text{-}value < \alpha$. Thus, we would still reject the null hypothesis if $\alpha = 0.01$ but not if the significance level had been $\alpha = 0.001$.

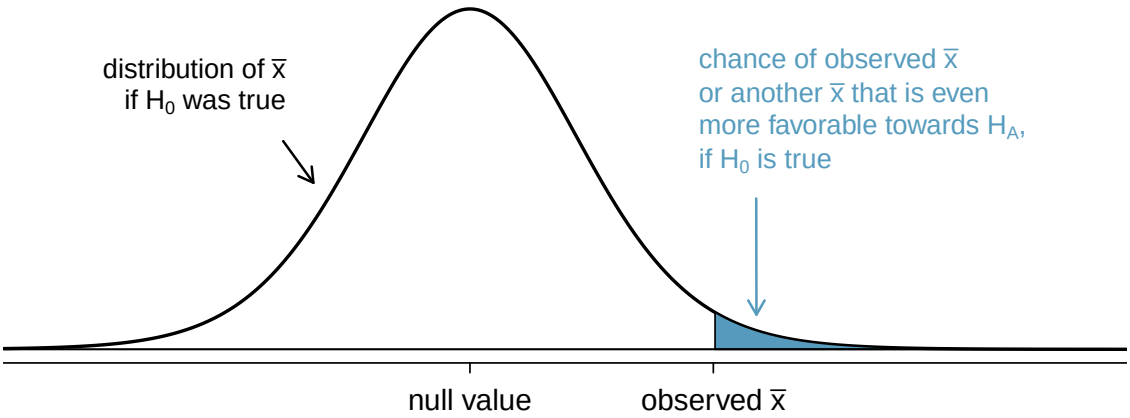


Figure 9.9: To identify the p-value, the distribution of the sample mean is considered as if the null hypothesis was true. Then the p-value is defined and computed as the probability of the observed $\bar{x}$ or an $\bar{x}$ even more favourable to H_A under this distribution.

Guided Practice 9.19 Ebay might be interested in showing that buyers on its site tend to pay less than they would for the corresponding new item on Amazon. We'll research this topic for one particular product: a video game called *Mario Kart* for the Nintendo Wii. During early October 2009, Amazon sold this game for $46.99. Set up an appropriate (one-sided!) hypothesis test to check the claim that Ebay buyers pay less during auctions at this same time.[113]

Guided Practice 9.20 During early October 2009, 52 Ebay auctions were recorded for *Mario Kart*.[114] The total prices for the auctions are presented using a histogram in Figure 9.10, and we may like to apply the normal model to the sample mean. Check the three conditions required for applying the normal model: (1) independence, (2) at least 30 observations, and (3) the data are not strongly skewed.[115]

Example 9.21 The average sale price of the 52 Ebay auctions for *Wii Mario Kart* was $44.17 with a standard deviation of $4.15. Does this provide sufficient evidence to reject the null hypothesis in Guided Practice 19? Use a significance level of $\alpha = 0.01$.

The hypotheses were set up and the conditions were checked in Exercises 19 and 20. The next step is to find the standard error of the sample mean and produce a sketch

[113]The sceptic would say the average is the same on Ebay, and we are interested in showing the average price is lower.

H_0: The average auction price on Ebay is equal to (or more than) the price on Amazon. We write only the equality in the statistical notation: $\mu_{ebay} = 46.99$.

H_A: The average price on Ebay is less than the price on Amazon, $\mu_{ebay} < 46.99$.

[114]These data were collected by OpenIntro staff.

[115](1) The independence condition is unclear. *We will make the assumption that the observations are independent, which we should report with any final results.* (2) The sample size is sufficiently large: $n = 52 \geq 30$. (3) The data distribution is not strongly skewed; it is approximately symmetric.

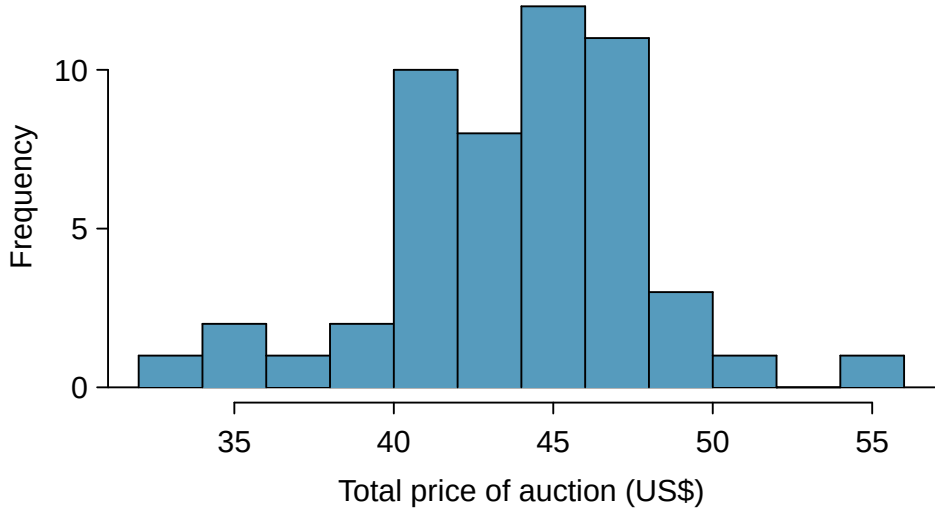
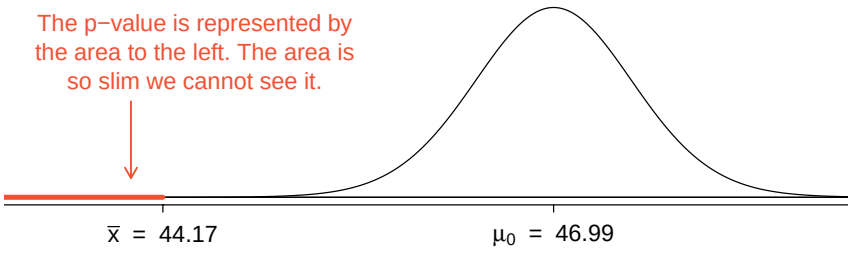


Figure 9.10: A histogram of the total auction prices for 52 Ebay auctions.

to help find the p-value.

$$SE_{\bar{x}} = s/\sqrt{n} = 4.15/\sqrt{52} = 0.5755$$



Because the alternative hypothesis says we are looking for a smaller mean, we shade the lower tail. We find this shaded area by using the Z-score and normal probability table: $Z = \frac{44.17-46.99}{0.5755} = -4.90$, which has area less than 0.0002. The area is so small we cannot really see it on the picture. This lower tail area corresponds to the p-value.

Because the p-value is so small – specifically, smaller than $\alpha = 0.01$ – this provides sufficiently strong evidence to reject the null hypothesis in favour of the alternative. The data provide statistically significant evidence that the average price on Ebay is lower than Amazon's asking price.

What's so special about 0.05?

It's common to use a threshold of 0.05 to determine whether a result is statistically significant, but why is the most common value 0.05? Maybe the standard significance level should be bigger, or maybe it should be smaller. If you're a little puzzled, that probably means you're reading with a critical eye – good job!

9.4 Examining the Central Limit Theorem

The normal model for the sample mean tends to be very good when the sample consists of at least 30 independent observations and the population data are not strongly skewed. The Central Limit Theorem provides the theory that allows us to make this assumption.

Central Limit Theorem, informal definition

The distribution of $\bar{x}$ is approximately normal. The approximation can be poor if the sample size is small, but it improves with larger sample sizes.

The Central Limit Theorem states that when the sample size is small, the normal approximation may not be very good. However, as the sample size becomes large, the normal approximation improves. We will investigate three cases to see roughly when the approximation is reasonable.

We consider three data sets: one from a *uniform* distribution, one from an *exponential* distribution, and the other from a *log-normal* distribution. These distributions are shown in the top panels of Figure 9.11. The uniform distribution is symmetric, the exponential distribution may be considered as having moderate skew since its right tail is relatively short (few outliers), and the log-normal distribution is strongly skewed and will tend to produce more apparent outliers.

The left panel in the $n = 2$ row represents the sampling distribution of $\bar{x}$ if it is the sample mean of two observations from the uniform distribution shown. The dashed line represents the closest approximation of the normal distribution. Similarly, the centre and right panels of the $n = 2$ row represent the respective distributions of $\bar{x}$ for data from exponential and log-normal distributions.

Guided Practice 9.22 Examine the distributions in each row of Figure 9.11. What do you notice about the normal approximation for each sampling distribution as the sample size becomes larger?[116]

Example 9.23 Would the normal approximation be good in all applications where the sample size is at least 30?

Not necessarily. For example, the normal approximation for the log-normal example is questionable for a sample size of 30. Generally, the more skewed a population distribution or the more common the frequency of outliers, the larger the sample required to guarantee the distribution of the sample mean is nearly normal.

TIP: With larger n, the sampling distribution of $\bar{x}$ becomes more normal

As the sample size increases, the normal model for $\bar{x}$ becomes more reasonable. We can also relax our condition on skew when the sample size is very large.

[116] The normal approximation becomes better as larger samples are used.

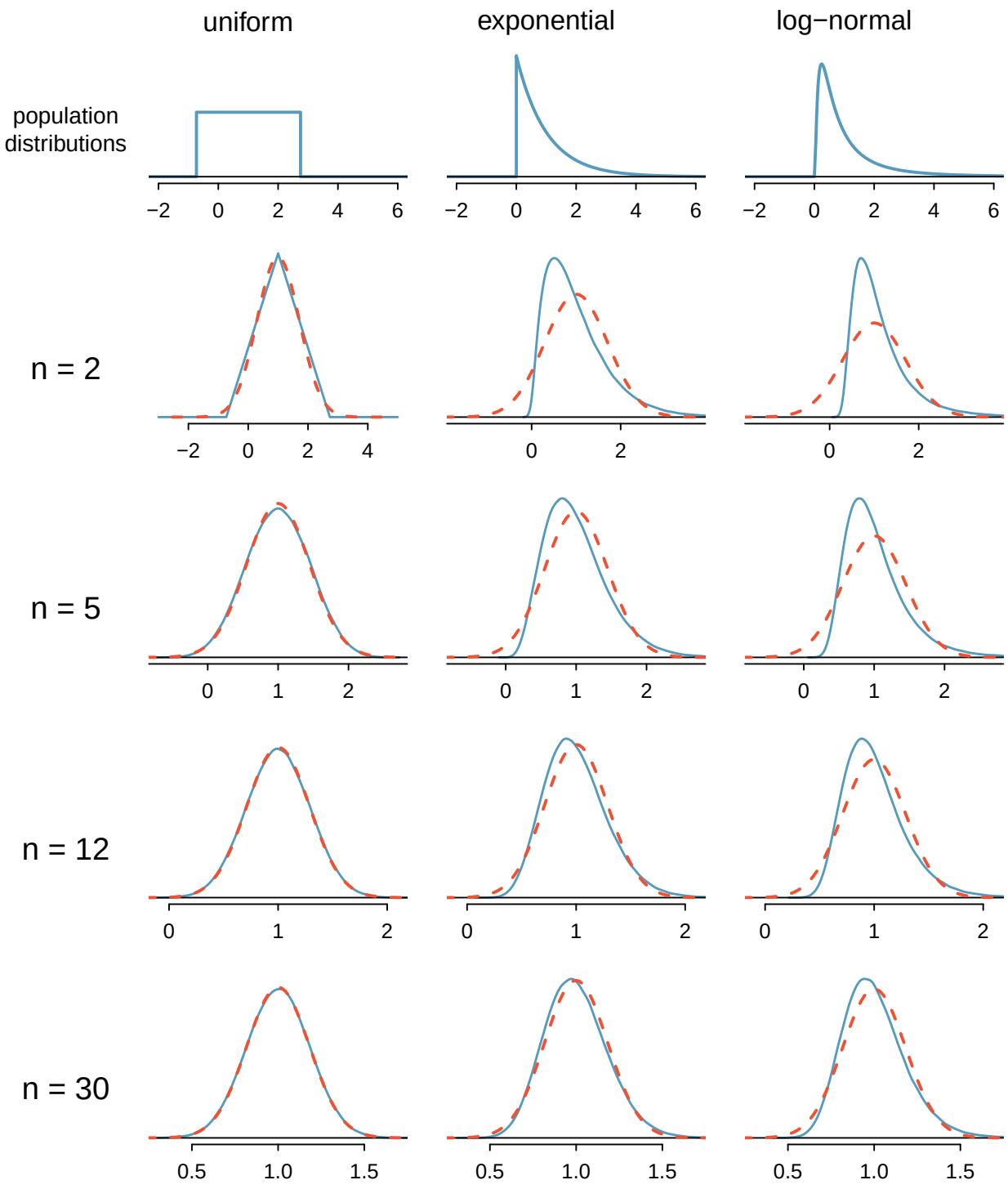


Figure 9.11: Sampling distributions for the mean at different sample sizes and for three different distributions. The dashed red lines show normal distributions.

We discussed in Section 9.1 that the sample standard deviation, s, could be used as a substitute of the population standard deviation, σ, when computing the standard error. This estimate tends to be reasonable when $n \geq 30$.

Example 9.24 Figure 9.12 shows a histogram of 50 observations. These represent winnings and losses from 50 consecutive days of a professional poker player. Can the normal approximation be applied to the sample mean, 90.69?

We should consider each of the required conditions.

(1) These are referred to as **time series data**, because the data arrived in a particular sequence. If the player wins on one day, it may influence how she plays the next. To make the assumption of independence we should perform careful checks on such data. While the supporting analysis is not shown, no evidence was found to indicate the observations are not independent.

(2) The sample size is 50, satisfying the sample size condition.

(3) There are two outliers, one very extreme, which suggests the data are very strongly skewed or very distant outliers may be common for this type of data. Outliers can play an important role and affect the distribution of the sample mean and the estimate of the standard error.

Since we should be sceptical of the independence of observations and the very extreme upper outlier poses a challenge, we should not use the normal model for the sample mean of these 50 observations. If we can obtain a much larger sample, perhaps several hundred observations, then the concerns about skew and outliers would no longer apply.

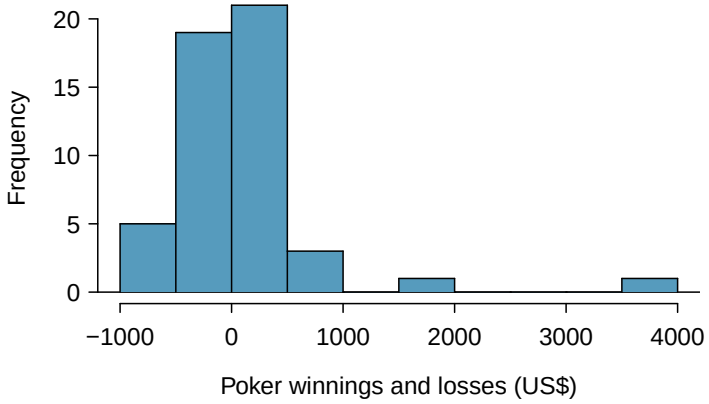


Figure 9.12: Sample distribution of poker winnings. These data include some very clear outliers. These are problematic when considering the normality of the sample mean. For example, outliers are often an indicator of very strong skew.

Caution: Examine data structure when considering independence

Some data sets are collected in such a way that they have a natural underlying structure between observations, e.g. when observations occur consecutively. Be especially cautious about independence assumptions regarding such data sets.

Caution: Watch out for strong skew and outliers

Strong skew is often identified by the presence of clear outliers. If a data set has prominent outliers, or such observations are somewhat common for the type of data under study, then it is useful to collect a sample with many more than 30 observations if the normal model will be used for $\bar{x}$.

You won't be a pro at assessing skew by the end of this book, so just use your best judgement and continue learning. As you develop your statistics skills and encounter tough situations, also consider learning about better ways to analyse skewed data, such as the studentized bootstrap (bootstrap-t), or consult a more experienced statistician.

Statistical significance versus practical significance

When the sample size becomes larger, point estimates become more precise and any real differences in the mean and null value become easier to detect and recognize. Even a very small difference would likely be detected if we took a large enough sample. Sometimes researchers will take such large samples that even the slightest difference is detected. While we still say that difference is **statistically significant**, it might not be **practically significant**.

Statistically significant differences are sometimes so minor that they are not practically relevant. This is especially important to research: if we conduct a study, we want to focus on finding a meaningful result. We don't want to spend lots of money finding results that hold no practical value.

The role of a statistician in conducting a study often includes planning the size of the study. The statistician might first consult experts or scientific literature to learn what would be the smallest meaningful difference from the null value. She also would obtain some reasonable estimate for the standard deviation. With these important pieces of information, she would choose a sufficiently large sample size so that the power for the meaningful difference is perhaps 80% or 90%. While larger sample sizes may still be used, she might advise against using them in some cases, especially in sensitive areas of research.

10 Single Sample t-test

The t-test is the most common statistical test in science, and also one of the most important. The test focus is the sample mean. Specifically, the test seeks to establish whether the sample mean is different from a specified value. The test provides information on whether it is relatively likely or relatively unlikely the specific sample we have could be a sample drawn from a population that has a mean equal to the hypothesis test value we specify. The t-test approach was devised by a beer brewer (W.S. Gosset, aka A. Student) at the Guinness brewery as he tried to control the alcohol content of beer. If the alcohol content of the beer was too high the brewery had to pay a lot of tax, and so made no profit. If the alcohol content of the beer was too low consumers switched to other products, and with low sales the company also made no profit. So, it turns out that beer is responsible for modern statistics. Go figure.

Provided the data sample is 'reasonably' large the t-test does not require that the sample data follow a normal distribution. If the data sample is small then it does matter whether the data sample is normally distributed or not. How to deal with small data samples that do not look normal is something that is covered in a subsequent handout.

There are many high quality text books and online resources that explain the theory of the t-test. The focus of this resource is on how the conceptual structure of a t-test and how to implement a t-test in **R**; not an explanation of the theory behind the t-test.

The test statistic for the t-test is calculated as the difference between the sample mean and the hypothesis test value (numerator), divided by the sample standard error (denominator). So we have ([sample mean - null hypothesis value]/standard error). As such, the test statistic will tend to be large when the difference between the sample mean and the null hypothesis test value is large and the standard error is small.

If the test statistic value is far from zero (large in absolute value terms) the conclusion drawn is that the sample mean is different to the null hypothesis value. If the test statistic is close to zero the conclusion drawn is that the sample mean is not different to the null hypothesis value. To determine what constitutes close to zero and what constitutes far from zero we use the p-value decision rule. For the t-test, the null hypothesis that the sample mean is not different to the null hypothesis test value is rejected when the p-value is small.

10.1 Preliminary Data Exploration

For this example we will use the Mauna Loa Atmospheric CO2 Concentration data set, **co2**. The data is a time series of monthly carbon dioxide concentrations at Mauna Loa, Hawaii, from 1959 to 1997. The data set is already installed with base **R**.

We want to know two things:

1. Is the average CO2 concentration at Mauna Loa for the period 1959 to 1997 different

from 338 ppm? At Mauna Loa, the average CO2 concentration for the entire 20th century as a whole was 338 ppm, so there is a logical reason for considering this test value.

2. Is the average CO2 concentration at Mauna Loa for the period 1959 to 1997 different from 321 ppm? At Mauna Loa, the average CO2 concentration for the entire 19th century as a whole was 321 ppm; so again there is some logic in selecting this as a test value.

We will answer these questions by conducting single sample t-tests; but first, as will always be the case,we need to conduct an informal investigation of our data set to make sure we understand the nature of the data we are working with.

Let's first look at the summary and structure of the data, and then create a histogram. The code used here has been covered in previous guides, but note that for a univariate time series object you only have to specify the data name (co2) rather than also having to use the *with()* command to first specify the data set name.

Note that the data structure information looks a little different to what you have previously seen. When we have data observations recorded through time there is a natural order to the data. The observations for 1960 should be listed after the observations for 1959, and the information we get with the *str()* command confirms this.

```
# get summary and structure
summary(co2) # gives us the 6 num. summary (mean, min, max, etc.)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   313.2   323.5   335.2   337.1   350.3   366.8

str(co2)     # we have a time-series; [1 var and 468 obs] from 1959 to 1998

##  Time-Series [1:468] from 1959 to 1998: 315 316 316 318 318 ...

# plot histogram
hist(co2, col= "lightgrey", border= "black",      # set optional parameters
      main= "Histogram of Mauna Loa CO2 Concentrations",
      xlab= "CO2 (ppm)", xlim= c(300,380), ylim= c(0,80))
```

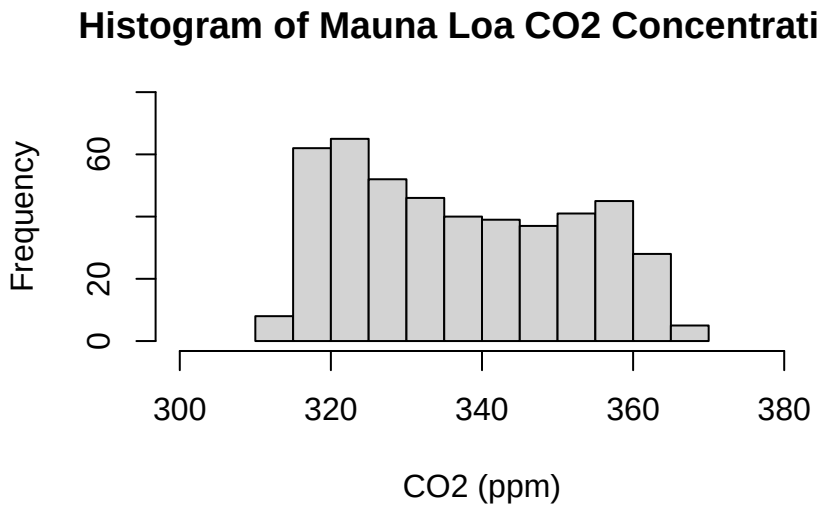


Figure 10.1: Histogram of Mauna Loa carbon dioxide concentrations between 1959 and 1997.

What shape does the histogram suggest? Does the distribution look normal, or maybe more like a uniform distribution? Remember, the t-test doesn't rely on the sample data distribution looking like a normal distribution, if we have a reasonably large sample. Here we have lots of observations. Although there is no fixed rule > 30 observations is generally seen as meeting the requirement of a reasonably large sample. Here we have 468 observations.

Let's go back to our two questions - is the sample mean statistically different from the mean carbon dioxide concentrations of the 19th and 20th centuries (321 ppm; 338 ppm). Where do these values fall in our distribution when looking at our histogram? Try and predict the outcomes of our tests based only on looking at the histogram and where these values lie. Recall that the numerator for the test statistic is the difference between the null hypothesis and the sample mean, which here is (337-338). If the difference between these two values is small then it is unlikely the test statistic will be far from zero.

If you have appropriately explored the data before conducting a t-test using visual techniques such as creating a histogram or a boxplot, you are unlikely to be surprised by the formal t-test result.

10.2 Conducting a One Sample t-test

Example: Fail to Reject the Null

From the data summary we know the sample mean is 337 ppm. We want to test whether 337 ppm is statistically different to the 20th century mean of 338 ppm. For this example we will test the claim that:

- The average CO2 concentration for the period 1959 and 1997 is different to the century average.

Let's go through the formal steps:

Step 1: Set the Null and Alternate Hypotheses
- Null hypothesis: The mean is 338 ppm
- Alternate hypothesis: The mean is not 338 ppm

Step 2: Set the Alpha Level
The alpha level is the probability of rejecting the null hypothesis when the null hypothesis is true (Type 1 error)

- We will always set this to 0.05 (5% chance of a Type 1 error) which is the default setting in almost all software programs. This is a topic that comes up in later guides, for the moment we just use the default for alpha.

Step 3: Implement the Single Sample t-test
To implement the test in **R** we use the function *t.test()*. The default for alpha is 0.05 so we only need to set the parameter `mu`, (for mean) to our Null Hypothesis value, which is 338 ppm.

Note that when working with your own data file you may need to also use the *with* command to specify which data file you are using before you apply the *t.test()* command.

```
t.test(co2, mu = 338)

##
##  One Sample t-test
##
## data:  co2
## t = -1.3681, df = 467, p-value = 0.1719
## alternative hypothesis: true mean is not equal to 338
## 95 percent confidence interval:
##  335.6941 338.4130
## sample estimates:
## mean of x
##  337.0535
```

Step 4: Interpret the Results
The first thing to check in the output is the test name. The output conforms we have the results for a single sample t-test. So we know we have implemented the correct test. Now consider the actual test output.

The p-value
From the output we see the `p-value = 0.1719`. Since our alpha is 0.05 and 0.172 is greater than 0.05, we: - **Fail to reject** the null hypothesis that the mean is 338 ppm.

What does this mean? It means, we did **not** have **sufficient evidence** to say the mean

carbon dioxide level at Mauna Loa, between 1959 and 1997, is different to the mean for the whole 20th century. Is this what you expected from the histogram plot? What about the rest of the output? While it is possible to make a decision based on the p-value decision rule, the test output includes a range of other values, such as the t-statistic and the 95% confidence interval. In addition to the p-value, it is worth making clear the relationship between the different values reported in the test output.

The t-value
The reported t-value is the test statistic. The t-value for the test is -1.37. Formally, our test looks at whether or not the value -1.37 is close to or far from zero. We used the p-value information (p-value =0.17) to interpret the reported test statistic value and conclude that this value is not far from zero (do not reject the null). We can also generate this value manually.

Recall the test static formula is: ([sample mean - null hypothesis value]/standard error). We know the difference between sample mean - null hypothesis value is (337.05-338.00) = -0.95, which is the numerator in the test statistic. We also know that the formula for the standard error (of the mean) is the sample standard deviation divided by the square root of the sample size. Using *sd(co2)* we find the sample standard deviation is 14.97. The sample size is 486, and $\sqrt{468} = 21.63$. The standard error is then 14.97/21.63 = 0.69, and the test statistic is found as = -0.95/ 0.69 = -1.37, which is the value reported in the output. Thankfully, we do not to do these calculations by hand, but it is nice to know how the reported values are derived.

The 95% confidence interval
The 95% confidence interval is (`335.7, 338.4`). The 95% confidence interval can be thought of as defining all values for a null hypothesis that will not be rejected. The 95% confidence interval can be derived as the sample mean value plus/minus the critical t-value multiplied by the sample standard error. However, for the moment we will ignore the specifics of the 95% confidence interval calculation and just focus on interpreting the 95% confidence interval as the values for a null hypothesis that will not be rejected. Using our p-value decision rule we concluded: do not reject the null. As the 95% confidence interval includes 338 ppm, our decision is consistent with the 95% Confidence Interval information.

Sample mean
We are also presented with the mean of our sample (`337.1 ppm`), which, unsurprisingly is the same value as when calculated using the *summary()* command.

The logic check
Now, go back and have a look at the histogram. Do the results of the test make sense? The visual plot is part of the analysis, and you should expect the plot and your formal tests to be in agreement. If they are not, it is worth checking to see if you have made a mistake somewhere. This structured approach is a way to develop good habits.

Example: Reject the Null

Now we'll test whether our sample mean of 337 ppm is different to the 19th century mean of 321 ppm. From the histogram we know that 321 is within our sample range of observations, but quite far from the mean. For this example we will test the claim that:

- The average CO2 concentration between 1959 and 1997 is not different from the average of the 19th century.

Let's go through the formal steps:

Step 1: Set the Null and Alternate Hypotheses

- Null hypothesis: The mean is 321 ppm
- Alternate hypothesis: The mean is not 321 ppm

Step 2: Set the Alpha Level

- We will always set this to 0.05 (5% chance of Type 1 error), and it is the default value anyway.

Step 3: Implement the Single Sample t-test

Here we set the parameter `mu` to 321 ppm, which is our new Null Hypothesis value.

```
t.test(co2, mu = 321)

##
##  One Sample t-test
##
## data:  co2
## t = 23.205, df = 467, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 321
## 95 percent confidence interval:
##  335.6941 338.4130
## sample estimates:
## mean of x
##  337.0535
```

Step 4: Interpret the Results

Again, first check that the output relates to the test you wanted to implement. Here, we can see that yes, the results are for a single sample t-test.

The p-value

From the output we see the `p-value = < 2.2e-16`, or < 0.001, which is smaller than 0.05. So, we:

- **Reject** the null hypothesis that the mean is 321 ppm.

A p-value less than 0.05 means we have **sufficient evidence** to conclude that the mean carbon dioxide level at Mauna Loa between 1959 and 1997 is different to the mean for the 19th century as a whole.

The t-value
The t-value for the test is 23.2, and we have concluded that this value is far from zero, hence the reject the null decision. Recall, to manually calculate the test statistic we use the formula: ([sample mean - null hypothesis value]/standard error). So we have: (337.05-321.00) = 16.05 as the numerator, and as with our earlier test we have $14.97/\sqrt{468}$= 0.69 as the denominator. The test statistic is then: 16.05/0.69= 23.2, which is the value reported in the test output.

The 95% confidence interval
We can also see that the 95% confidence interval (`335.7, 338.4`) does not include 321 ppm. The 95% confidence interval defines potential null hypothesis values that we would not reject using a p-value decision rule threshold value of 0.05. As 321 ppm falls outside the 95% confidence interval we should expect that a p-value decision rule would lead us to reject the null.

The logic check
Go back and have a look at the histogram. In this instance we see that 321 ppm is towards the edge of the plot and so there seems to be consistency between the 'feel' we have for the data based on the plot and the formal test result.

10.3 Advanced: Superscripts, Lines, Text, and Legends

This section will cover how to add text and vertical lines to your histogram to better illustrate the core elements you are testing. Adding vertical lines using *abline()* was first introduced in the guide Creating Histograms in R, and the function to add text, *text()*, works in a way that is similar to the way the *abline()* function work. Within *text()* we will specify: our position for the text with axis coordinates using parameters `x` and `y`; the text with parameter `labels`, position relative to the x, y coordinates with `pos` (1=below, 2=left, 3=above, 4=right), and text colour with `col`. The same method for entering coordinates for our text will also be used to specify our legend's position, instead of using the general position options as before (e.g. "topright").

We will also cover how to include superscripts in plot and legend text. These features require complex formatting instructions and syntax, and use an additional function *expression()*. All text for the title or legend item is included within this function's (parentheses); a superscript is denoted by a caret or exponent character '^' and the characters/numbers to be raised enclosed in {braces}. Normal text within the expression is enclosed in "quotes" and a tilde '~' is used to add spaces between superscripts and normal text. So the syntax will look something like this:

```
expression("text"^{superscript} ~ "text")
```

Let's create our new histogram:

```
# create histogram
hist(co2, col= "lightgrey", border= "black",
     # add labels using expressions() to allow for superscript formatting
     main= expression("Histogram of Mauna Loa CO"^{2} ~ "Concentrations"),
     xlab= expression("CO"^{2} ~ "(ppm)"),
     xlim= c(300,400), ylim= c(0,80))

# save mean as object, and add vertical lines
mean.co2 <- mean(co2)
abline(v= mean.co2, lty= 1, lwd= 3, col= "black")
abline(v= 338, lty= 2, lwd= 2, col= "red")
abline(v= 321, lty= 2, lwd= 2, col= "blue")

# add text giving value at each vertical line added
# list coordinates, text, position to coordinates, and colour
text(x= 338, y= 80, labels= "338 ppm", pos= 4, col= "red")
text(x= 321, y= 80, labels= "321 ppm", pos= 2, col= "blue")
text(x= mean.co2, y= 70, labels= "337 ppm", pos= 4)

# add legend for three lines
legend(x= 360, y= 70,                    # give coordinates for legend position
       legend= c("Sample Mean",
                 # specify objects using expression() to allow for superscripts
                  expression("20"^{th} ~ "Century Mean"),
                  expression("19"^{th} ~  "Century Mean")),
       col= c("black","red","blue"),     # specify colours
       lwd= c(3,2,2), lty= c(1,2,2),     # specify line widths and types
       bty= "n", cex= 0.8)               # supress outline / decrease font size
```

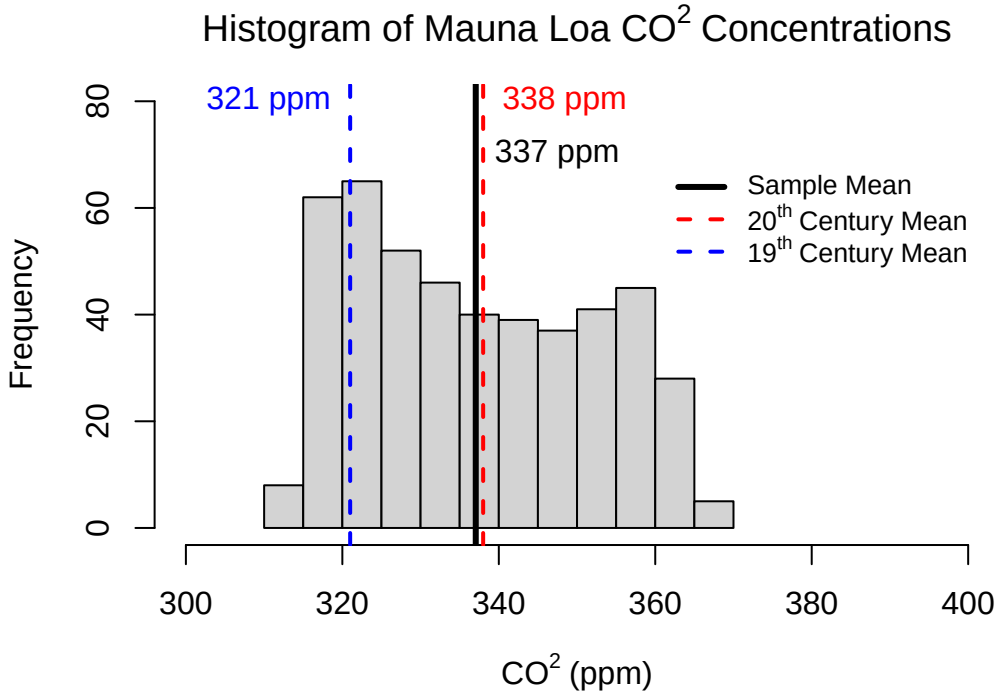


Figure 10.2: Histogram of Mauna Loa carbon dioxide concentrations between 1959 and 1997, also displaying the reported means for the 19th and 20th centuries.

10.4 Advanced: The p-value

P-value decision rules are controversial. The architect of the t-test – W.S. Gosset – had no time for such ideas; yet R. Fisher, another giant of the field, thought p-value decision rules a useful approach. The issues involved with these arguments are complex, and preferences for reporting standards vary across disciplines. The approach advocated in this handout is to combine a preliminary visual inspection of the data with a t-test that relies on a p-value decision rule. This might be thought of as a weight of evidence approach. If both the data visualisation and the formal t-test results (based on a p-value decision rule) agree, most reasonable people will be convinced of the result.

If you are a student in the latter part of your studies you may wish to consult the below references. The paper provides a relatively accessible treatment of the strengths and weaknesses of p-value decision rules. There are also many other papers that discuss these issues, and if you are going on to advanced study, it might be worth consulting some of this literature.

Reference

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. The American Statistician. Vol. 70, No. 2, pp. 129-133.

10.5 Advanced: Technical Details on Single-Sample t-test

We generally require a large sample for two reasons:

1. The sampling distribution of $\bar{x}$ tends to be more normal when the sample is large.
2. The calculated standard error is typically very accurate when using a large sample.

So what should we do when the sample size is small? If the population data are nearly normal, then $\bar{x}$ will also follow a normal distribution, which addresses the first problem. The accuracy of the standard error is trickier, and for this challenge we'll introduce a new distribution called the t-distribution.

While we emphasize the use of the t-distribution for small samples, this distribution is also generally used for large samples, where it produces similar results to those from the normal distribution.

The normality condition

A special case of the Central Limit Theorem ensures the distribution of sample means will be nearly normal, regardless of sample size, when the data come from a nearly normal distribution.

Central Limit Theorem for normal data

The sampling distribution of the mean is nearly normal when the sample observations are independent and come from a nearly normal distribution. This is true for any sample size.

While this seems like a very helpful special case, there is one small problem. It is inherently difficult to verify normality in small data sets.

Caution: Checking the normality condition

We should exercise caution when verifying the normality condition for small samples. It is important to not only examine the data but also think about where the data come from. For example, ask: would I expect this distribution to be symmetric, and am I confident that outliers are rare?

You may relax the normality condition as the sample size goes up. If the sample size is 10 or more, slight skew is not problematic. Once the sample size hits about 30, then moderate skew is reasonable. Data with strong skew or outliers require a more cautious analysis.

Introducing the t-distribution

In the cases where we will use a small sample to calculate the standard error, it will be useful to rely on a new distribution for inference calculations: the t-distribution. A t-distribution, shown as a solid line in Figure 10.3, has a bell shape. However, its tails are thicker than

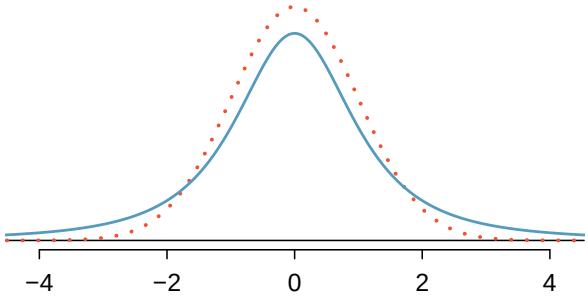


Figure 10.3: Comparison of a t-distribution (solid line) and a normal distribution (dotted line).

the normal model's. This means observations are more likely to fall beyond two standard deviations from the mean than under the normal distribution.[117] While our estimate of the standard error will be a little less accurate when we are analyzing a small data set, these extra thick tails of the t-distribution are exactly the correction we need to resolve the problem of a poorly estimated standard error.

The t-distribution, always centered at zero, has a single parameter: degrees of freedom. The **degrees of freedom (df)** describe the precise form of the bell-shaped t-distribution. Several t-distributions are shown in Figure 10.4. When there are more degrees of freedom, the t-distribution looks very much like the standard normal distribution.

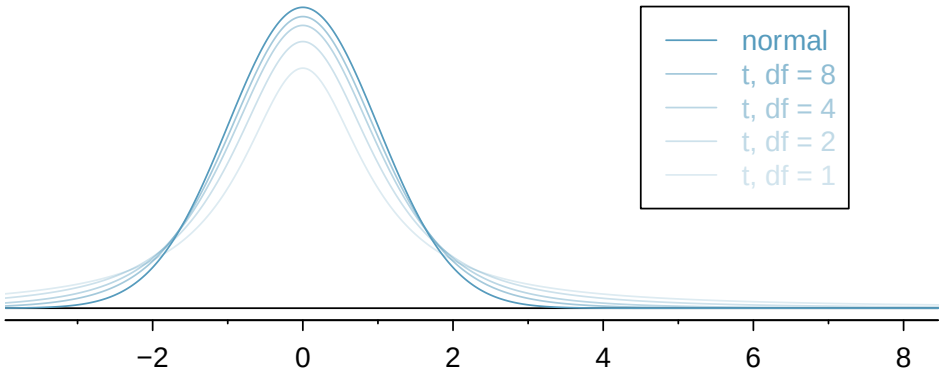


Figure 10.4: The larger the degrees of freedom, the more closely the t-distribution resembles the standard normal model.

Degrees of freedom (df)

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

[117] The standard deviation of the t-distribution is actually a little more than 1. However, it is useful to always think of the t-distribution as having a standard deviation of 1 in all of our applications.

When the degrees of freedom is about 30 or more, the t-distribution is nearly indistinguishable from the normal distribution. In Section 10.5, we relate degrees of freedom to sample size.

It's very useful to become familiar with the t-distribution, because it allows us greater flexibility than the normal distribution when analyzing numerical data. We use a **t-table**, partially shown in Table 10.1, in place of the normal probability table. In practice, it's more common to use statistical software instead of a table.

	one tail	0.100	0.050	0.025	0.010	0.005
	two tails	0.200	0.100	0.050	0.020	0.010
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

Table 10.1: 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 tail probabilities. For instance, if we know we are working with the t-distribution with $df = 18$, we can examine row 18, which is highlighted in Table 10.1. If we want the value in this row that identifies the cutoff for an upper tail of 10%, we can look in the column where *one tail* is 0.100. This cutoff is 1.33. If we had wanted the cutoff for the lower 10%, we would use -1.33. Just like the normal distribution, all t-distributions are symmetric.

Example 10.1 What proportion of the t-distribution with 18 degrees of freedom falls below -2.10?

Just like a normal probability problem, we first draw the picture in Figure 10.5 and shade the area below -2.10. To find this area, we identify the appropriate row: $df = 18$. Then we identify the column containing the absolute value of -2.10; it is the third column. Because we are looking for just one tail, we examine the top line of the table, which shows that a one tail area for a value in the third row corresponds to 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.

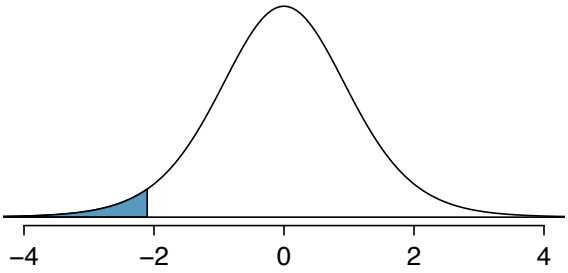


Figure 10.5: The t-distribution with 18 degrees of freedom. The area below -2.10 has been shaded.

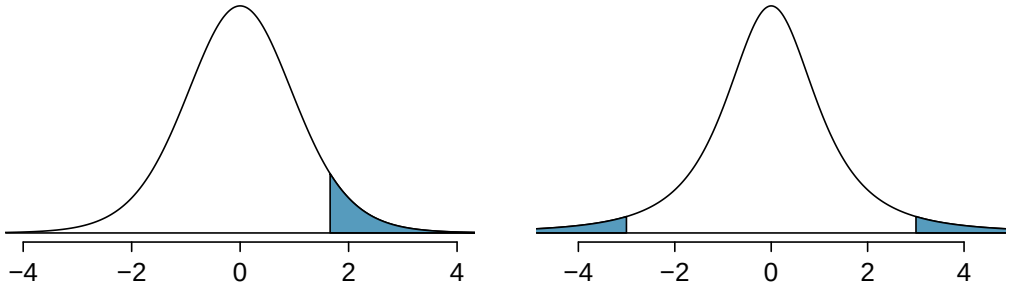


Figure 10.6: Left: The t-distribution with 20 degrees of freedom, with the area above 1.65 shaded. Right: The t-distribution with 2 degrees of freedom, with the area further than 3 units from 0 shaded.

Example 10.2 A t-distribution with 20 degrees of freedom is shown in the left panel of Figure 10.6. Estimate the proportion of the distribution falling above 1.65.

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 first and second columns. Since these values bound 1.65, their tail areas will bound the tail area corresponding to 1.65. We identify the one tail area of the first and second columns, 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.

Example 10.3 A t-distribution with 2 degrees of freedom is shown in the right panel of Figure 10.6. Estimate the proportion of the distribution falling more than 3 units from the mean (above or below).

As before, first identify the appropriate row: $df = 2$. Next, find the columns that capture 3; because $2.92 < 3 < 4.30$, we use the second and third columns. Finally, we find bounds for the tail areas by looking at the two tail values: 0.05 and 0.10. We use the two tail values because we are looking for two (symmetric) tails.

⊙ **Guided Practice 10.4** What proportion of the t-distribution with 19 degrees of freedom falls above -1.79 units?[118]

Conditions for using the t-distribution for inference on a sample mean

To proceed with the t-distribution for inference about a single mean, we first 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 the data are from an experiment or random process, we 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.

When examining a sample mean and estimated standard error from a sample of n independent and nearly normal observations, we use a t-distribution with $n - 1$ degrees of freedom (df). For example, if the sample size was 19, then we would use the t-distribution with $df = 19 - 1 = 18$ degrees of freedom and proceed exactly as we did with the Z-distribution, except that *now we use the t-distribution.*

TIP: When to use the t-distribution

Use the t-distribution for inference of the sample mean when observations are independent and nearly normal. You may relax the nearly normal condition as the sample size increases. For example, the data distribution may be moderately skewed when the sample size is at least 30.

One sample t-confidence intervals

Dolphins are at the top of the oceanic food chain, which causes dangerous substances such as mercury to concentrate in their organs and muscles. This is an important problem for both dolphins and other animals, like humans, who occasionally eat them. For instance, this is particularly relevant in Japan where school meals have included dolphin at times.

Here we identify a confidence interval for the average mercury content in dolphin muscle using a sample of 19 Risso's dolphins from the Taiji area in Japan.[119] The data are summa-

[118] We find the shaded area *above* -1.79 (we leave the picture to you). The small left tail is between 0.025 and 0.05, so the larger upper region must have an area between 0.95 and 0.975.

[119] Taiji was featured in the movie *The Cove*, and it is a significant source of dolphin and whale meat in Japan. Thousands of dolphins pass through the Taiji area annually, and we will assume these 19 dolphins represent a simple random sample from those dolphins. 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.

Figure 10.7: A Risso's dolphin.

Photo by Mike Baird (www.bairdphotos.com). CC BY 2.0 license.

rized in Table 10.2. The minimum and maximum observed values can be used to evaluate whether or not there are obvious outliers or skew.

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

Table 10.2: Summary of mercury content in the muscle of 19 Risso's dolphins from the Taiji area. Measurements are in μg/wet g (micrograms of mercury per wet gram of muscle).

Example 10.5 Are the independence and normality conditions satisfied for this data set?

The observations are a simple random sample and consist of less than 10% of the population, therefore independence is reasonable. The summary statistics in Table 10.2 do not suggest any skew or outliers; all observations are within 2.5 standard deviations of the mean. Based on this evidence, the normality assumption seems reasonable.

In the normal model, we used $z^\star$ and the standard error to determine the width of a confidence interval. We revise the confidence interval formula slightly when using the t-distribution:

$$\bar{x} \pm t^\star_{df} SE$$

$t^\star_{df}$ Multiplication factor for t conf. interval

The sample mean and estimated standard error are computed just as before ($\bar{x} = 4.4$ and $SE = s/\sqrt{n} = 0.528$). The value $t^\star_{df}$ is a cutoff we obtain based on the confidence level and the t-distribution with df degrees of freedom. Before determining this cutoff, we will first need the degrees of freedom.

Degrees of freedom for a single sample

If the sample has n observations and we are examining a single mean, then we use the t-distribution with $df = n - 1$ degrees of freedom.

In our current example, we should use the t-distribution with $df = 19 - 1 = 18$ degrees of freedom. Then identifying $t^\star_{18}$ is similar to how we found $z^\star$.

- For a 95% confidence interval, we want to find the cutoff $t^\star_{18}$ such that 95% of the t-distribution is between -$t^\star_{18}$ and $t^\star_{18}$.
- We look in the t-table on page 136, find the column with area totalling 0.05 in the two tails (third column), and then the row with 18 degrees of freedom: $t^\star_{18} = 2.10$.

Generally the value of $t^\star_{df}$ is slightly larger than what we would get under the normal model with $z^\star$.

Finally, we can substitute all our values into the confidence interval equation to create the 95% confidence interval for the average mercury content in muscles from Risso's dolphins that pass through the Taiji area:

$$\bar{x} \pm t^\star_{18} SE \quad \rightarrow \quad 4.4 \pm 2.10 \times 0.528 \quad \rightarrow \quad (3.29, 5.51)$$

We are 95% confident the average mercury content of muscles in Risso's dolphins is between 3.29 and 5.51 μg/wet gram, which is considered extremely high.

Finding a t-confidence interval for the mean

Based on a sample of n independent and nearly normal observations, a confidence interval for the population mean is

$$\bar{x} \pm t^\star_{df} SE$$

where $\bar{x}$ is the sample mean, $t^\star_{df}$ corresponds to the confidence level and degrees of freedom, and SE is the standard error as estimated by the sample.

Guided Practice 10.6 The FDA's webpage provides some data on mercury content of fish.[120] Based on a sample of 15 croaker white fish (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. We will assume these observations are independent. Based on the summary statistics of the data, do you have any objections to the normality condition of the individual observations?[121]

Example 10.7 Estimate the standard error of $\bar{x} = 0.287$ ppm using the data summaries in Guided Practice 6. If we are to use the t-distribution to create a 90% confidence interval for the actual mean of the mercury content, identify the degrees of freedom we should use and also find $t^\star_{df}$.

The standard error: $SE = \frac{0.069}{\sqrt{15}} = 0.0178$. Degrees of freedom: $df = n - 1 = 14$.

Looking in the column where two tails is 0.100 (for a 90% confidence interval) and row $df = 14$, we identify $t^\star_{14} = 1.76$.

Guided Practice 10.8 Using the results of Guided Practice 6 and Example 7, compute a 90% confidence interval for the average mercury content of croaker white fish (Pacific).[122]

One sample t-tests

Is the typical US runner getting faster or slower over time? We consider this question in the context of the Cherry Blossom Race, which is a 10-mile race in Washington, DC each spring.[123]

The average time for all runners who finished the Cherry Blossom Race in 2006 was 93.29 minutes (93 minutes and about 17 seconds). We want to determine using data from 100 participants in the 2012 Cherry Blossom Race whether runners in this race are getting faster or slower, versus the other possibility that there has been no change.

Guided Practice 10.9 What are appropriate hypotheses for this context?[124]

Guided Practice 10.10 The data come from a simple random sample from less than 10% of all participants, so the observations are independent. However, should we be worried about skew in the data? See Figure 10.8 for a histogram of the differences.[125]

---

[120] www.fda.gov/food/foodborneillnesscontaminants/metals/ucm115644.htm

[121] There 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, and we do not have reason to believe the mercury content is not nearly normal in this type of fish.

[122] $\bar{x} \pm t^\star_{14} SE \rightarrow 0.287 \pm 1.76 \times 0.0178 \rightarrow (0.256, 0.318)$. We are 90% confident that the average mercury content of croaker white fish (Pacific) is between 0.256 and 0.318 ppm.

[123] www.cherryblossom.org

[124] H_0: The average 10 mile run time was the same for 2006 and 2012. $\mu = 93.29$ minutes. H_A: The average 10 mile run time for 2012 was *different* than that of 2006. $\mu \neq 93.29$ minutes.

[125] With a sample of 100, we should only be concerned if there is very strong skew. The histogram of the data suggests, at worst, slight skew.

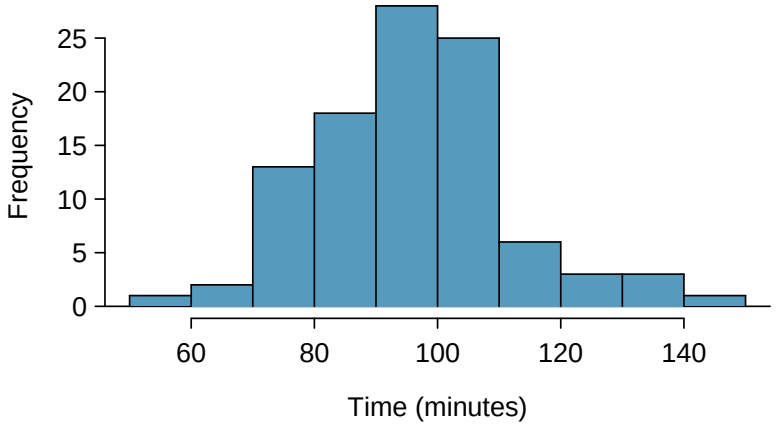


Figure 10.8: A histogram of `time` for the sample Cherry Blossom Race data.

With independence satisfied and slight skew not a concern for this large of a sample, we can proceed with performing a hypothesis test using the t-distribution.

⊙ **Guided Practice 10.11** The sample mean and sample standard deviation of the sample of 100 runners from the 2012 Cherry Blossom Race are 95.61 and 15.78 minutes, respectively. Recall that the sample size is 100. What is the p-value for the test, and what is your conclusion?[126]

When using a t-distribution, we use a T-score (same as Z-score)

To help us remember to use the t-distribution, we use a T to represent the test statistic, and we often call this a **T-score**. The Z-score and T-score are computed in the exact same way and are conceptually identical: each represents how many standard errors the observed value is from the null value.

[126]With the conditions satisfied for the t-distribution, we can compute the standard error ($SE = 15.78/\sqrt{100} = 1.58$ and the *T-score*: $T = \frac{95.61-93.29}{1.58} = 1.47$. (There is more on this after the guided practice, but a T-score and Z-score are calculated in the same way.) For $df = 100 - 1 = 99$, we would find $T = 1.47$ to fall between the first and second column, which means the p-value is between 0.10 and 0.20 (use $df = 90$ and consider two tails since the test is two-sided). The p-value could also have been calculated more precisely with statistical software: 0.1447. Because the p-value is greater than 0.05, we do not reject the null hypothesis. That is, the data do not provide strong evidence that the average run time for the Cherry Blossom Run in 2012 is any different than the 2006 average.

11 Variance Ratio Test

Generally we are interested in testing whether or not there is a difference in the group means. When testing for differences in group means the specific test statistic formula to use depends on whether or not the group variances are equal. There is one formula for the case when the groups have equal variance, and a second formula to use when the group variances are not equal. Sometimes we are just interested in knowing whether or not the variance of two groups is different. For example, if we invest in a glasshouse to grow plants we would expect the variation in plant development to be less than for plants grown in the field.

When there are only two groups the test we use to determine if the variance is the same is called a variance ratio test. The test involves dividing the variance of group one by the variance of group two. If this ratio is close to one the conclusion drawn is that the variance of each group is the same. If the ratio is far from one the conclusion drawn is that the variances are not the same.

How far from one would this ratio need to be before we are convinced that the two population variances are not equal? In the case where the variable that we are interested in follows a normal distribution in the two populations, and these normal distributions have the same variance, then the ratio of sample variances follows another known distribution, known as the F-distribution. In the standard notation it S_1^2 and S_2^2 are the sample variances from population 1 and population 2 respectively, then the ratio

$$F = S_1^2 / S_2^2$$

follows an F-distribution when then population variances are equal. The only two parameters of the F-distribution are the *degrees of freedom*. These are related to the sizes of the samples taken from the two populations, m and n. The way that the shape of the F-distribution changes with different degrees of freedom is shown in Figure 11.1.

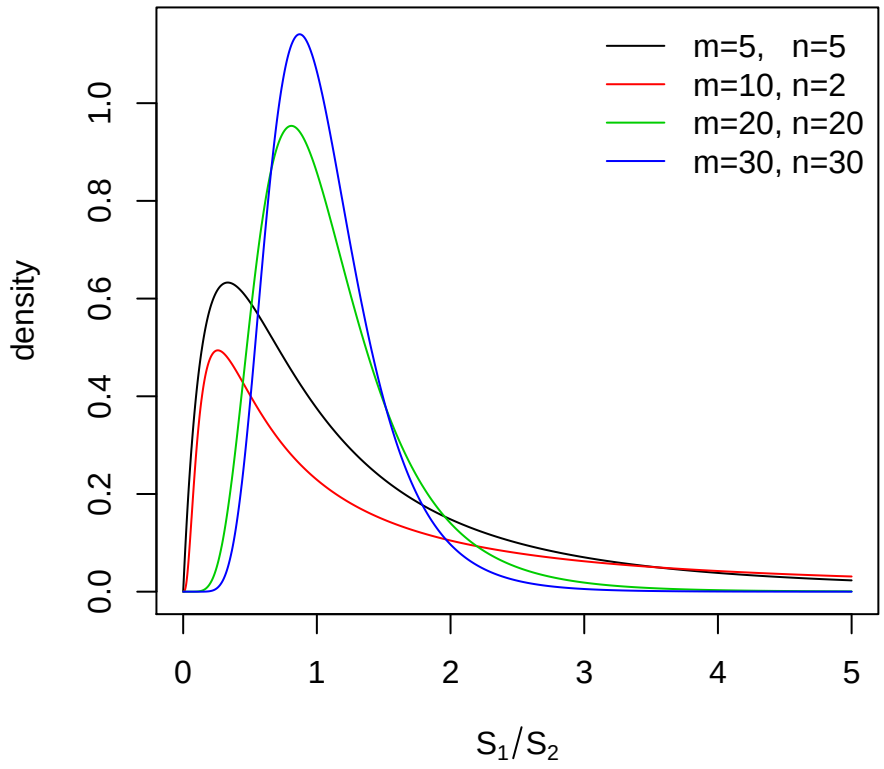


Figure 11.1: F-distribution for varying degrees of freedom

You'll notice that as the sample sizes increase, the shape of the distribution gets more closely centred around one. This means that as our sample variances are based on more observations, the ratio of sample variances doesn't need to stray as far from one in order for us to declare that the population variance are not equal. To put this another way, if we only take small samples from each population, our sample variance estimates won't be very good and so we could belive that the ratio of the two sample variances might be quite different to one, just by chance. As we increase the sample taken from each population we are more certain that the sample variances are close to the population variances and so if the ratio was only a small distance from one, we would no longer believe that the population variances were equal.

To determine what values are close to one and what values are far from one we use the p-value decision rule. For the variance ratio test the null hypothesis that the group variances are equal is rejected when the p-value is small.

11.1 Equal Variance Testing - Two Groups

For this example we have data from an aquaculture farm rearing rainbow trout. It has been suggested that reducing the number of fish in each pen can decrease the size variation of market ready fish. The farm owner believes more consistently sized fish will sell for a bigger profit than fish that are more varied in size, so wants to look into this further.

Here we will be using the function `var.test()` to test whether a pen holding 250 fish results in less size variation, than a pen holding 300 fish. We will not test whether the means are different, but this could be done with a two sample t-test. To run our test, we will first read in our data: a sample of 10 weight measurements from each pen. Generally this data will be in an MS Excel file that we read into `R`, but here we enter the data directly into R and bind the two samples together as one data frame using the functions `data.frame()` and `cbind()` ('column' bind) - this will result in wide format data.

```
# read in our data and get it in the correct format
Trout.250 <- c(508, 479, 545, 531, 559, 422, 547, 525, 420, 491, 508, 511, 569,
    453, 533, 460, 523, 540, 463, 502)
Trout.300 <- c(461, 464, 344, 559, 445, 617, 402, 531, 535, 413, 456, 479, 393,
    504, 416, 468, 368, 519, 523, 531)
Farmed.Trout <- data.frame(cbind(Trout.250, Trout.300))  # combine as data frame
```

Once we have our data in `R`, before formally conducting our variance ratio test we use the `str()` and `summary()` commands to look at the data, and then create a boxplot to visually compare the two distributions.

Let's get started.

```
str(Farmed.Trout)  # wide data - weights of 2 groups in separate columns

## 'data.frame': 20 obs. of  2 variables:
##  $ Trout.250: num  508 479 545 531 559 422 547 525 420 491 ...
##  $ Trout.300: num  461 464 344 559 445 617 402 531 535 413 ...

summary(Farmed.Trout)  # compare values of 2 groups

##    Trout.250       Trout.300
##  Min.   :420.0   Min.   :344.0
##  1st Qu.:475.0   1st Qu.:415.2
##  Median :509.5   Median :466.0
##  Mean   :504.4   Mean   :471.4
##  3rd Qu.:534.8   3rd Qu.:525.0
##  Max.   :569.0   Max.   :617.0
```

Although the summary has given us the 5-number summary (plus the mean) a boxplot makes it easier to comprehend the information.

```
# wide data format: use , not ~
with(Farmed.Trout,boxplot(Trout.250,Trout.300,
     col= "lightgray",
     main= "Weights of Aquaculture Raised Rainbow Trout",
     xlab= "Pen density (No. fish)",
     ylab= "Weight (g)",
     ylim= c(300,650),
     names= c("250 per pen","300 per pen"), # group names
     las= 1,
     boxwex =0.6))
```

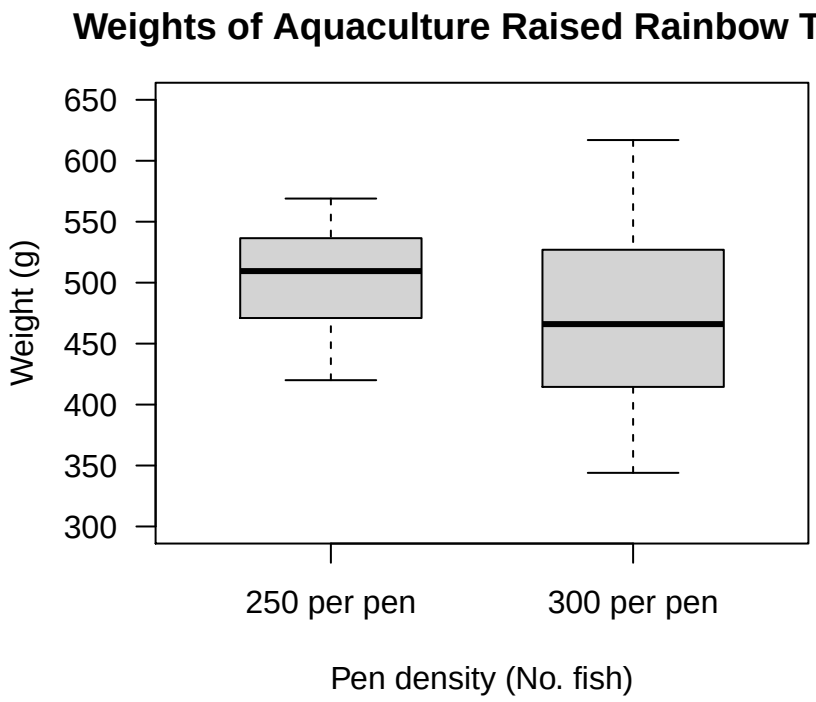


Figure 11.2: Box plot of rainbow trout weight distributions from two aquaculture pen densities (250 fish/pen and 300 fish/pen)

What do you see? It looks like the mean weight is higher for the lower density pen, but the largest fish are from the higher density pen. What about the variances? There looks like less variation in the lower density pen, but does the difference look significant? Let's see what our test says:

Step 1: Set the Null and Alternate Hypotheses

- Null hypothesis: The variance ratio is equal to one
- Null hypothesis: The group variances are equal

- Alternate hypothesis: The variance ratio is not equal to one
- Alternate hypothesis: The group variances are not equal

Step 2: Implement the Variance Ratio Test

Here we provide the data, and two groups separated by a ',' because the data is in wide format. Our alpha value will be set as the default: 0.05.

```
with(Farmed.Trout, var.test(Trout.250, Trout.300))

##
##  F test to compare two variances
##
## data:  Trout.250 and Trout.300
## F = 0.38583, num df = 19, denom df = 19, p-value = 0.04421
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1527172 0.9747866
## sample estimates:
## ratio of variances
##          0.3858324
```

Step 3: Interpret the Results

From the output we see the `p-value = 0.04421`. Since 0.044 is less than 0.05 (alpha), we:

- **Reject** the null hypothesis that the group variances are the same

What does this mean? It means, we have **sufficient evidence** to say the variance for fish weight for the two aquaculture pen densities are different. The variance ratio (`0.39`) is, in this context, far enough from 1 for us to reject the null.

We also get a 95% confidence interval for the variance ratio of (`0.15, 0.97`). This range does not include 1. This aligns with the p-value information. If the p-value is less than 0.05 the 95% confidence interval does not contain 1. So, although our 95% upper confidence level is close to 1 our estimate is still quite far away and we reject the null. If we were to move on to test the means of the two groups we would set our t-test parameter `var.equal` to FALSE.

11.2 Advanced: Technical Note

Recall that the ratio of sample variances follows an F-distribution when the variable of interest is normally distributed in each of the populations. The F-test for equal variances is

only valid when this is true and in fact this test is quite sensitive to departures from normality.

Whether or not a variance ratio test should ever be conducted is an open question. Simulation studies have shown that for many cases where the standard t-test performs poorly, and hence we want to use the unequal variance t-test formula, the variance ratio test does not have good ability to accurately determine that the variances of the two groups are different. Conversely, for scenarios where the variance ratio test has good ability to accurately determine that the variance of the two groups are different, just using the standard t-test that incorrectly assumes the variances of the groups are equal still works very well. Combined these two results have led some to conclude that it is better to not ever use a variance ratio test.

To more fully understand the issues involved requires an understanding of what is meant by a type I error; what is meant by the expression power of the test; and the definitions of both the chi-square distribution and the F-distribution. Consideration of such issues is something that is only relevant for students in their 4th or 5th year at university. A relatively accessible reference for such students is provided below. If you are taking an introductory class, for example SCIE1104, these issues are beyond the scope of the course and before you undertake a two sample t-test you are expected to conduct a variance ratio test.

Reference

Markowski, C. A., & Markowski, E. P. (1990). Conditions for the effectiveness of a preliminary test of variance. The American Statistician, 44(4), 322-326.

12 Two Sample t-test

Comparing two groups and testing whether the means of the two groups are different is a common process in science. However, rather than jump straight into the t-test it is necessary to follow a structured process of investigation. The structured process involves (i) a visual inspection of the data; (ii) a variance ratio pre-test, and (iii) the actual two sample t-test.

12.1 Example with Wide Format Data

For this example we have some phytoremediation data. Phytoremediation is the use of plants for remediating contaminated soil and water. Plant species are selected based on their ability to uptake or stabilize specific contaminants at a site, and this method of remediation is often preferred because it is low cost and relatively non-invasive.

Here we will look at the efficiency of two crop plants (redbeet and barley) at removing cadmium from the top 20 cm of soil at a contaminated site. The data we have is the percent reduction of cadmium after one harvest. Let's read in our data and have a look at it before we conduct our test. Generally you will start by reading in the data from a .csv file or a .txt file. So that the work is reproducible, here we enter the data directly.

Note: As our data is in wide format we use a comma between groups, rather than a tilde '~' when creating the boxplot. In addition to the *str()* and *summary()* commands, here we also use the *head()* command to look at the data we have. The *head()* command shows the first six rows of the data, and is an easy way to see the data structure.

```
# read in our data (wide format)
Cd.BeetBarley<- data.frame(
  redbeet= c(18, 5, 10, 8, 16, 12, 8, 8, 11, 5, 6, 8, 9, 21, 9),
  barley= c(8, 5, 10, 19, 15, 18, 11, 8, 9, 4, 5, 13, 7, 5, 7))

# three ways to look at the data structure
str(Cd.BeetBarley)

## 'data.frame': 15 obs. of  2 variables:
##  $ redbeet: num  18 5 10 8 16 12 8 8 11 5 ...
##  $ barley : num  8 5 10 19 15 18 11 8 9 4 ...

summary(Cd.BeetBarley)

##     redbeet          barley
##  Min.   : 5.00   Min.   : 4.0
##  1st Qu.: 8.00   1st Qu.: 6.0
##  Median : 9.00   Median : 8.0
##  Mean   :10.27   Mean   : 9.6
```

```
##  3rd Qu.:11.50   3rd Qu.:12.0
##  Max.   :21.00   Max.   :19.0

head(Cd.BeetBarley)

##   redbeet barley
## 1      18      8
## 2       5      5
## 3      10     10
## 4       8     19
## 5      16     15
## 6      12     18
```

Once we understand the data structure, and are confident the data has been read into **R** correctly, we then create a boxplot to look at the data. Remember, the boxplot is essentially a visual representation of the information we get with the *summary()* command.

```
with(Cd.BeetBarley,boxplot(redbeet,barley,
     col= "lightgrey",
     main= "Phytoremediation Efficiency of Crop Plants",
     xlab= "Crop type", ylab= "Cadmium reduction (%)",
     names= c("Redbeet","Barley"),
     ylim=  c(0,25), las= 1,
     boxwex=0.6))
```

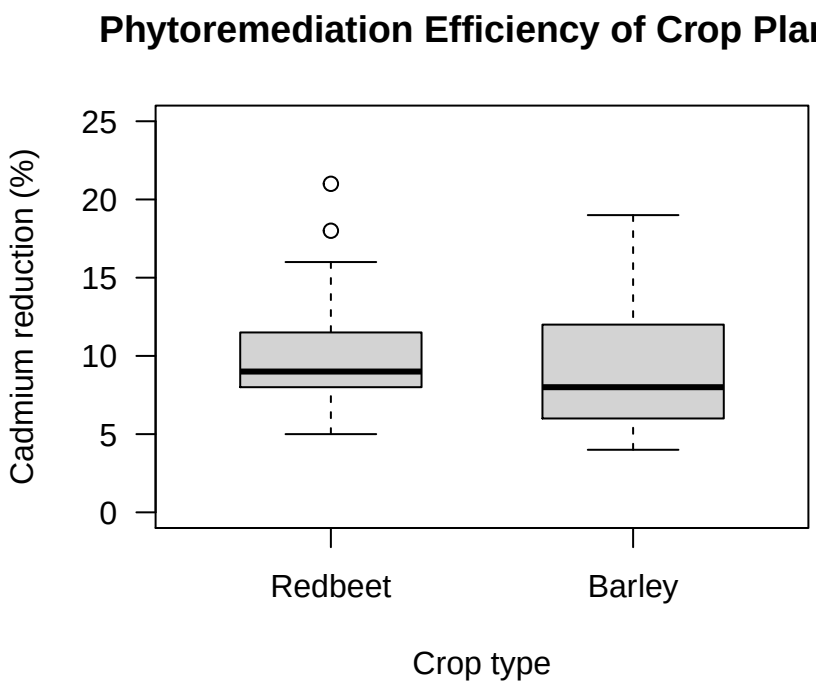


Figure 12.1: Box plot comparing the phytoremediation efficiency of redbeet and barley crop plants for removing cadmium from contaminated soil at depths of 0 to 20 cm

What do you see? Do the means looks different; does it look like we have equal variances? Looking at the data is a very important step in the process.

To determine which Two Sample t-test formula to use – equal variance formula or unequal variance formula – we need to compare the group variances. The **Null hypothesis** for the test is that the ratio of the two group variances is equal to one (they are equal); the **Alternate hypothesis** is that the ratio of the two group variances is not equal to one (they are not equal). For a more in-depth explanation of the variance ratio test see the reference guide for Variance Ratio Testing in R.

Before we conduct the actual variance ratio test, it is worth thinking through what we know about the data samples based on the data summary and the visual plot. From the data summary we know that we only have 15 observations in each group. On average, the more observations we have the more likely we are to reject the null. So, with few observations, if we are to reject the null the difference between the variance of each group will need to be substantial. From the boxplot we see that the 'box' for redbeet looks a bit smaller than for barley. That suggests the variance for the redbeet sample might be smaller than for barley. On the other hand, we can also see that there are two 'outliers' for the redbeet sample. These 'outliers' values work to increase the variance, and so work to increase the variance

estimate for redbeet. So based on a general understanding of the data samples we would not be surprised if, at the end of the formal test we conclude: do not reject the null.

```
with(Cd.BeetBarley, var.test(redbeet, barley))

##
##  F test to compare two variances
##
## data:  redbeet and barley
## F = 0.97888, num df = 14, denom df = 14, p-value = 0.9687
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3286377 2.9156685
## sample estimates:
## ratio of variances
##          0.9788762
```

From the output we see the `p-value = 0.969` is greater than 0.05, so we: **Fail to reject** the null hypothesis that the variance ratio is equal to one. If we look at the actual ratio of the variances we see that it is (`0.98`). As such, are you surprised that we do not reject the null that the ratio is one? The practical implication of this for us is that we need to set our t-test parameter `var.equal` to TRUE.

Now let's run our formal Two Sample t-test:

Step 1: Set the Null and Alternate Hypotheses

- Null hypothesis: The means of both groups are equal
- Alternate hypothesis: The means of both groups are not equal

Step 2: Implement the Two Sample t-test
Here we set `var.equal` to TRUE, and leave alpha as the default, 0.05.

Note: it is worth paying attention to the structure of the t-test when you have data in wide format.

```
with(Cd.BeetBarley, t.test(redbeet, barley, var.equal = TRUE))

##
##  Two Sample t-test
##
## data:  redbeet and barley
## t = 0.38658, df = 28, p-value = 0.702
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.865852  4.199185
```

```
## sample estimates:
## mean of x mean of y
##  10.26667   9.60000
```

Step 3: Interpret the Results

From the output we see the `p-value = 0.702`. Since 0.702 is greater than 0.05 (alpha), we:

- **Fail to reject** the null hypothesis that the means are equal

In other words: We do **not** have **sufficient evidence** to say the mean percent reduction of cadmium in soil at our site is different for our two crops (redbeet and barley). If you were trying to determine which plant to use to remediate a particular site, you would likely look to other factors, such as cost, environmental conditions or additional contaminants to be removed, to assist you in making your decision. Our output also provides the 95% confidence interval (`-2.9, 4.2`) for the actual difference between the means. Because the interval includes zero we get a non-significant p-value. Also note that the test output reports our group means (`10.3, 9.6`), which are the same as were given when we used the *str()* command.

12.2 Example with Long Format Data

In this example, we will use a similar data set, but compare two different crops (maize and cabbage) which are known for their ability to remove cadmium from greater soil depths (20 to 40 cm). We will follow the same process as before, but this time our data will be in long format. This means when specifying our data we list our continuous variable (% Cd reduction) and factor variable (crop type) separated by a tilde, '~', rather than a comma.

Note: Although creating a data.frame in long format looks complex, this process is really just to reproduce the way data typically looks in a .csv file in the first place. The data creation step is not something you would usually need to do. Here the information is included so that you can reproduce the results if you want to.

```
# read in our data (long format)
Cd.CabbageMaize <- data.frame(remed.pcnt = c(46, 50, 44, 44, 43, 52, 48, 24,
    51, 29, 53, 32, 61, 59, 35, 34, 26, 44, 17, 34, 19, 34, 34, 43, 18, 34,
   27, 27, 53, 30), plt.typ = c(rep("cabbage", times = 15), rep("maize", times = 15)))
```

So, we will be working with data that is in two columns. But unlike the wide data format, the first column holds the numerical values for each group, and the second column holds the grouping information (what type of plant).

```
# get summary & check data structure
str(Cd.CabbageMaize)

## 'data.frame': 30 obs. of  2 variables:
```

```
##  $ remed.pcnt: num  46 50 44 44 43 52 48 24 51 29 ...
##  $ plt.typ   : Factor w/ 2 levels "cabbage","maize": 1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(Cd.CabbageMaize)
```

```
##    remed.pcnt       plt.typ
##  Min.   :17.00   cabbage:15
##  1st Qu.:29.25   maize  :15
##  Median :34.50
##  Mean   :38.17
##  3rd Qu.:47.50
##  Max.   :61.00
```

```
head(Cd.CabbageMaize)
```

```
##   remed.pcnt plt.typ
## 1         46 cabbage
## 2         50 cabbage
## 3         44 cabbage
## 4         44 cabbage
## 5         43 cabbage
## 6         52 cabbage
```

So we will be working with data that is in two columns. But unlike the wide data format, the first column holds the numerical values for each group, and the second column holds the grouping information (what type of plant). The way the data is structured has implications for the commands we use to create plots and conduct tests.

```
# Note we don't NEED to give names for boxes if the data is in long format
# This is an advantage of the long format approach

with(Cd.CabbageMaize,boxplot(remed.pcnt~plt.typ,
     col= "lightgrey",
     main= "Phytoremediation Efficiency of Crop Plants",
     xlab= "Crop type", ylab= "Cadmium reduction (%)",
     ylim=  c(10,70), las= 1, boxwex=.6))
```

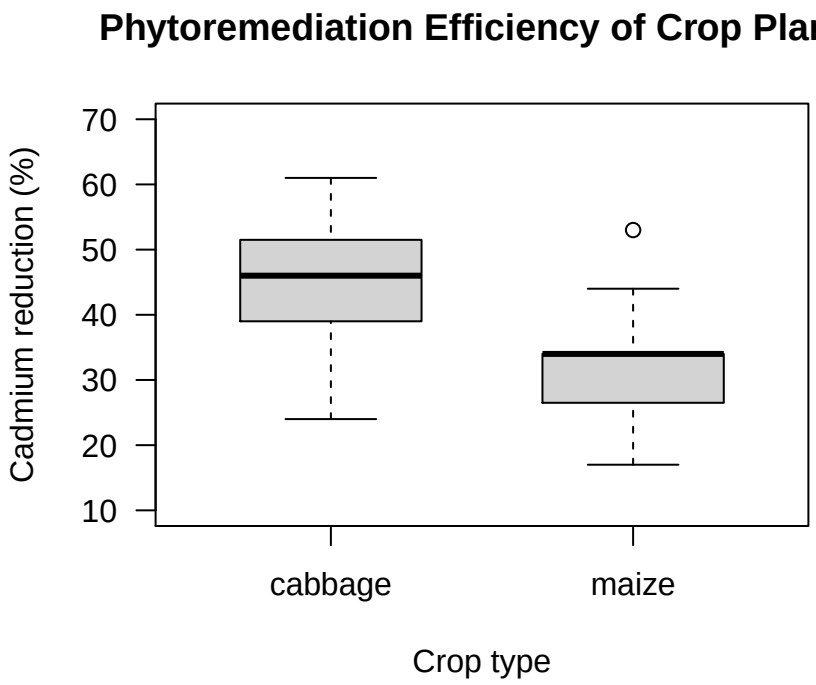


Figure 12.2: Box plot comparing the phytoremediation efficiency of cabbage and maize crop plants for removing cadmium from contaminated soil at depths of 20 to 40 cm.

Here the variances do not look that different. The box part of the plot for maize is smaller than for cabbage, but with the outlier for the maize crop working to increase the variance estimate for this group, I would not be surprised if the formal test says it is safe to assume the group variances are the same. As before, our **Null hypothesis** is that the variance ratio is equal to one (they are equal), and our **Alternate hypothesis** is that the variance ratio is not equal to one (they are not equal).

```
with(Cd.CabbageMaize, var.test(remed.pcnt ~ plt.typ))  # long format, use ~ not ,
```

```
##
##  F test to compare two variances
##
## data:  remed.pcnt by plt.typ
## F = 1.1449, num df = 14, denom df = 14, p-value = 0.8037
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3843653 3.4100823
## sample estimates:
## ratio of variances
##           1.144866
```

From the output we see the `p-value = 0.804`, is greater than 0.05, so again we: **Fail to reject** the null hypothesis that the variance ratio is equal to one and set our t-test parameter `var.equal` to TRUE. Again, if we look at the actual variance ratio number (`1.14`) we should not be surprised that we fail to reject the null: the value is quite close to one.

Now let's run our formal Two Sample t-test:

Step 1: Set the Null and Alternate Hypotheses

- Null hypothesis: The group means are equal
- Alternate hypothesis: The group means are not equal

Step 2: Implement the Two Sample t-test

Here we set `var.equal` to TRUE, and leave alpha as the default, 0.05.

```
with(Cd.CabbageMaize, t.test(remed.pcnt ~ plt.typ, var.equal = TRUE))

##
##  Two Sample t-test
##
## data:  remed.pcnt by plt.typ
## t = 3.4687, df = 28, p-value = 0.00171
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   5.377502 20.889165
## sample estimates:
## mean in group cabbage   mean in group maize
##              44.73333              31.60000
```

Step 3: Interpret the Results

From the output we see the `p-value = 0.00171`. Since 0.00171 is less than 0.05 (alpha), we:

- **Reject** the null hypothesis that the means are equal

In other words: We have **sufficient evidence** to say the mean percent reduction of cadmium in soil at our site is higher for cabbage than for maize. This is an important piece of information, but it is not the only piece of relevant information. You will also need other information, such as whether or not there are any differences in establishment costs for each plant. You would also always refer back to the group means (`44.7, 31.6` %) and your confidence interval (`5.4, 20.9` %). Does a difference of about 13 percentage points matter in your situation? What about if the difference is only 5 percentage points (lower end of your 95% C.I.)? Your statistical test results are just numbers until you logically apply them to your research question, and the accepted methods of your research discipline.

12.3 Advanced: Technical Details on Two-Sample t-test

In this section we consider a difference in two population means, $\mu_1 - \mu_2$, under the condition that the data are not paired. Just as with a single sample, we identify conditions to ensure we can use the t-distribution with a point estimate of the difference, $\bar{x}_1 - \bar{x}_2$.

We apply these methods in three contexts: determining whether stem cells can improve heart function, exploring the impact of pregnant women's smoking habits on birth weights of newborns, and exploring whether there is statistically significant evidence that one variations of an exam is harder than another variation. This section is motivated by questions like "Is there convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?"

Confidence interval for a difference of means

Does treatment using embryonic stem cells (ESCs) help improve heart function following a heart attack? Table 12.1 contains summary statistics for an experiment to test ESCs in sheep that had a heart attack. Each of these sheep was randomly assigned to the ESC or control group, and the change in their hearts' pumping capacity was measured in the study. A positive value corresponds to increased pumping capacity, which generally suggests a stronger recovery. Our goal will be to identify a 95% confidence interval for the effect of ESCs on the change in heart pumping capacity relative to the control group.

A point estimate of the difference in the heart pumping variable can be found using the difference in the sample means:

$$\bar{x}_{esc} - \bar{x}_{control} = 3.50 - (-4.33) = 7.83$$

	n	$\bar{x}$	s
ESCs	9	3.50	5.17
control	9	-4.33	2.76

Table 12.1: Summary statistics of the embryonic stem cell study.

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.

Example 12.1 Can the t-distribution be used to make inference using the point estimate, $\bar{x}_{esc} - \bar{x}_{control} = 7.83$?

We check the two required conditions:

1. In this study, the sheep were independent of each other. Additionally, the distributions in Figure 12.3 don't show any clear deviations from normality, where we watch for prominent outliers in particular for such small samples. These findings imply each sample mean could itself be modelled using a t-distribution.
2. The sheep in each group were also independent of each other.

Because both conditions are met, we can use the t-distribution to model the difference of the two sample means.

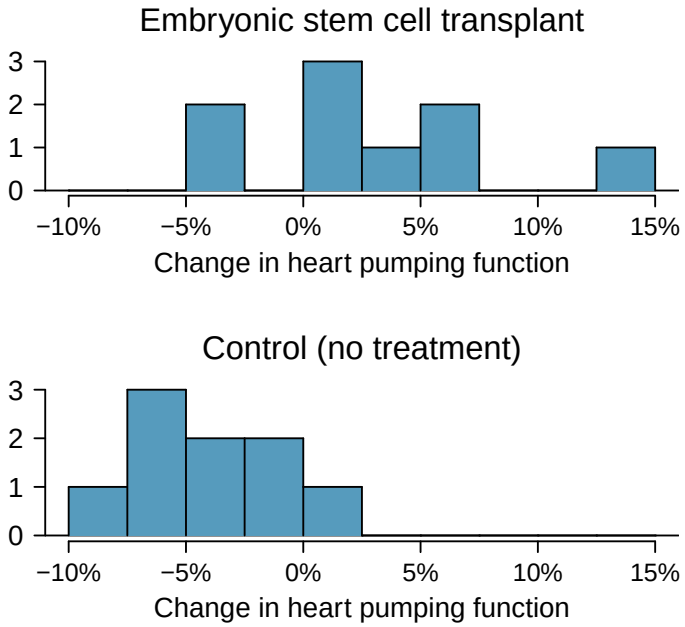


Figure 12.3: Histograms for both the embryonic stem cell group and the control group. Higher values are associated with greater improvement. We don't see any evidence of skew in these data; however, it is worth noting that skew would be difficult to detect with such a small sample.

When not assuming that the variances are equal, we can quantify the variability in the point estimate, $\bar{x}_{esc} - \bar{x}_{control}$, using the following formula for its standard error:

$$SE_{\bar{x}_{esc}-\bar{x}_{control}} = \sqrt{\frac{\sigma_{esc}^2}{n_{esc}} + \frac{\sigma_{control}^2}{n_{control}}}$$

We usually estimate this standard error using standard deviation estimates based on the

samples:

$$SE_{\bar{x}_{esc}-\bar{x}_{control}} = \sqrt{\frac{\sigma_{esc}^2}{n_{esc}} + \frac{\sigma_{control}^2}{n_{control}}}$$
$$\approx \sqrt{\frac{s_{esc}^2}{n_{esc}} + \frac{s_{control}^2}{n_{control}}} = \sqrt{\frac{5.17^2}{9} + \frac{2.76^2}{9}} = 1.95$$

Because we will use the t-distribution, we also must identify the appropriate degrees of freedom. This can be done using computer software. An alternative technique is to use the smaller of $n_1 - 1$ and $n_2 - 1$, which is the method we will typically apply in the examples and guided practice.[127]

Distribution of a difference of sample means for unequal variances

The sample difference of two means, $\bar{x}_1 - \bar{x}_2$, can be modelled using the t-distribution and the standard error

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{2}$$

when each sample mean can itself be modelled using a t-distribution and the samples are independent. To calculate the degrees of freedom, use statistical software or the smaller of $n_1 - 1$ and $n_2 - 1$.

Example 12.3 Calculate a 95% confidence interval for the effect of ESCs on the change in heart pumping capacity of sheep after they've suffered a heart attack.

We will use the sample difference and the standard error for that point estimate from our earlier calculations:

$$\bar{x}_{esc} - \bar{x}_{control} = 7.83$$
$$SE = \sqrt{\frac{5.17^2}{9} + \frac{2.76^2}{9}} = 1.95$$

Using $df = 8$, we can identify the appropriate $t^\star_{df} = t^\star_8$ for a 95% confidence interval as 2.31. Finally, we can enter the values into the confidence interval formula:

$$\text{point estimate} \pm t^\star SE \quad \rightarrow \quad 7.83 \pm 2.31 \times 1.95 \quad \rightarrow \quad (3.32, 12.34)$$

We are 95% confident that embryonic stem cells improve the heart's pumping function in sheep that have suffered a heart attack by 3.32% to 12.34%.

[127] This 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 df method. In this example, computer software would have provided us a more precise degrees of freedom of $df = 12.225$.

Hypothesis tests based on a difference in means

A data set called `baby_smoke` represents a random sample of 150 cases of mothers and their newborns in North Carolina over a year. Four cases from this data set are represented in Table 12.2. We are particularly interested in two variables: `weight` and `smoke`. The `weight` variable represents the weights of the newborns and the `smoke` variable describes which mothers smoked during pregnancy. We would like to know, is there convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke? We will use the North Carolina sample to try to answer this question. The smoking group includes 50 cases and the nonsmoking group contains 100 cases, represented in Figure 12.4.

	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
⋮	⋮	⋮	⋮	⋮	⋮	
150	45	50	36	9.25	female	nonsmoker

Table 12.2: Four cases from the `baby_smoke` data set. The value "NA", shown for the first two entries of the first variable, indicates that piece of data is missing.-2mm

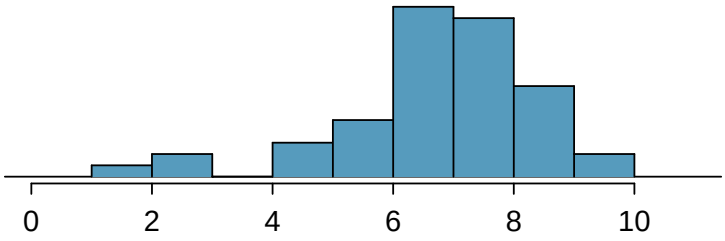
Example 12.4 Set up appropriate hypotheses to evaluate whether there is a relationship between a mother smoking and average birth weight.

The null hypothesis represents the case of no difference between the groups.

H_0: There is no difference in average birth weight for newborns from mothers who did and did not smoke. In statistical notation: $\mu_n - \mu_s = 0$, where μ_n represents non-smoking mothers and μ_s represents mothers who smoked.

H_A: There is some difference in average newborn weights from mothers who did and did not smoke ($\mu_n - \mu_s \neq 0$).

We check the two conditions necessary to apply the t-distribution to the difference in sample means. (1) Because the data come from a simple random sample and consist of less than 10% of all such cases, the observations are independent. Additionally, while each distribution is strongly skewed, the sample sizes of 50 and 100 would make it reasonable to model each mean separately using a t-distribution. The skew is reasonable for these sample sizes of 50 and 100. (2) The independence reasoning applied in (1) also ensures the observations in each sample are independent. Since both conditions are satisfied, the difference in sample means may be modelled using a t-distribution.



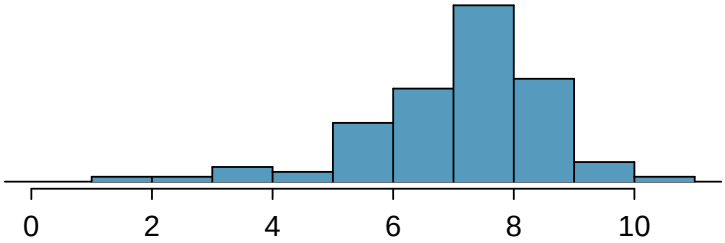


Figure 12.4: The top panel represents birth weights for infants whose mothers smoked. The bottom panel represents the birth weights for infants whose mothers who did not smoke. The distributions exhibit moderate-to-strong and strong skew, respectively.

	`smoker`	`nonsmoker`
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

Table 12.3: Summary statistics for the `baby_smoke` data set.

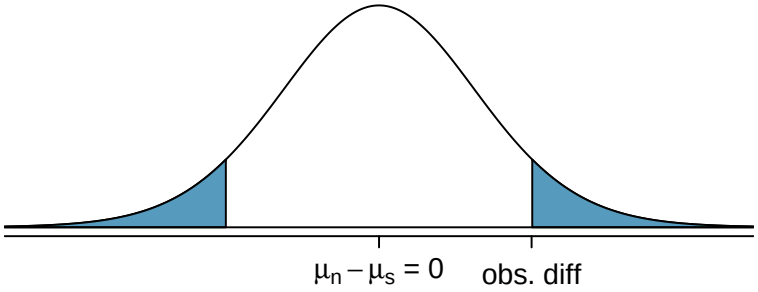
⊙ Guided Practice 12.5 The summary statistics in Table 12.3 may be useful for this exercise. (a) What is the point estimate of the population difference, $\mu_n - \mu_s$? (b) Compute the standard error of the point estimate from part (a).[128]

[128](a) The difference in sample means is an appropriate point estimate: $\bar{x}_n - \bar{x}_s = 0.40$. (b) The standard error of the estimate can be estimated using Equation (2):

$$SE = \sqrt{\frac{\sigma_n^2}{n_n} + \frac{\sigma_s^2}{n_s}} \approx \sqrt{\frac{s_n^2}{n_n} + \frac{s_s^2}{n_s}} = \sqrt{\frac{1.60^2}{100} + \frac{1.43^2}{50}} = 0.26$$

Example 12.6 Draw a picture to represent the p-value for the hypothesis test from Example 4.

To depict the p-value, we draw the distribution of the point estimate as though H_0 were true and shade areas representing at least as much evidence against H_0 as what was observed. Both tails are shaded because it is a two-sided test.



Guided Practice 12.7 If we made a Type 2 Error and there is a difference, what could we have done differently in data collection to be more likely to detect the difference?[129]

[129] We could have collected more data. If the sample sizes are larger, we tend to have a better shot at finding a difference if one exists.

Public service announcement: while we have used this relatively small data set as an example, larger data sets show that women who smoke tend to have smaller newborns. In fact, some in the tobacco industry actually had the audacity to tout that as a *benefit* of smoking:

> *It's true. The babies born from women who smoke are smaller, but they're just as healthy as the babies born from women who do not smoke. And some women would prefer having smaller babies.*
>
> - Joseph Cullman, Philip Morris' Chairman of the Board on CBS' *Face the Nation*, Jan 3, 1971

Fact check: the babies from women who smoke are not actually as healthy as the babies from women who do not smoke.[130]

Case study: two versions of a course exam

An instructor decided to run two slight variations of the same exam. Prior to passing out the exams, she shuffled the exams together to ensure each student received a random version. Summary statistics for how students performed on these two exams are shown in Table 12.4. Anticipating complaints from students who took Version B, she would like to evaluate whether the difference observed in the groups is so large that it provides convincing evidence that Version B was more difficult (on average) than Version A.

Version	n	$\bar{x}$	s	min	max
A	30	79.4	14	45	100
B	27	74.1	20	32	100

Table 12.4: Summary statistics of scores for each exam version.

Guided Practice 12.8 Construct a hypotheses to evaluate whether the observed difference in sample means, $\bar{x}_A - \bar{x}_B = 5.3$, is due to chance.[131]

Guided Practice 12.9 To evaluate the hypotheses in Guided Practice 8 using the t-distribution, we must first verify assumptions. (a) Does it seem reasonable that the scores are independent within each group? (b) What about the normality / skew condition for observations in each group? (c) Do you think scores from the two groups would be independent of each other, i.e. the two samples are independent?[132]

[130] You can watch an episode of John Oliver on *This Week Tonight* to explore the present day offenses of the tobacco industry. Please be aware that there is some adult language: youtu.be/6UsHHOCH4q8.

[131] Because the teacher did not expect one exam to be more difficult prior to examining the test results, she should use a two-sided hypothesis test. H_0: the exams are equally difficult, on average. $\mu_A - \mu_B = 0$. H_A: one exam was more difficult than the other, on average. $\mu_A - \mu_B \neq 0$.

[132] (a) It is probably reasonable to conclude the scores are independent, provided there was no cheating. (b) The summary statistics suggest the data are roughly symmetric about the mean, and it doesn't seem unreasonable to suggest the data might be normal. Note that since these samples are each nearing 30, moderate skew in the data would be acceptable. (c) It seems reasonable to suppose that the samples are independent since the exams were handed out randomly.

After verifying the conditions for each sample and confirming the samples are independent of each other, we are ready to conduct the test using the t-distribution. In this case, we are estimating the true difference in average test scores using the sample data, so the point estimate is $\bar{x}_A - \bar{x}_B = 5.3$. The standard error of the estimate can be calculated as

$$SE = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} = \sqrt{\frac{14^2}{30} + \frac{20^2}{27}} = 4.62$$

Finally, we construct the test statistic:

$$T = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{(79.4 - 74.1) - 0}{4.62} = 1.15$$

If we have a computer handy, we can identify the degrees of freedom as 45.97. Otherwise we use the smaller of $n_1 - 1$ and $n_2 - 1$: $df = 26$.

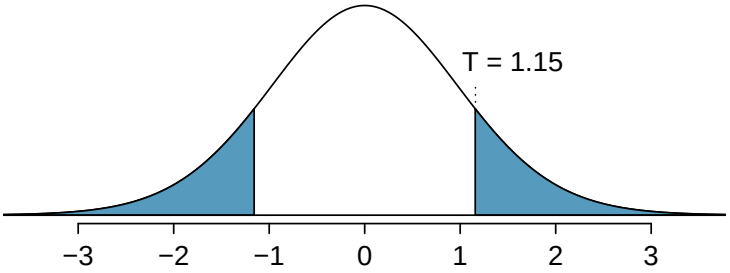


Figure 12.5: The t-distribution with 26 degrees of freedom. The shaded right tail represents values with $T \geq 1.15$. Because it is a two-sided test, we also shade the corresponding lower tail.

Example 12.10 Identify the p-value using $df = 26$ and provide a conclusion in the context of the case study.

We examine row $df = 26$ in the t-table. Because this value is smaller than the value in the left column, the p-value is larger than 0.200 (two tails!). Because the p-value is so large, we do not reject the null hypothesis. That is, the data do not convincingly show that one exam version is more difficult than the other, and the teacher should not be convinced that she should add points to the Version B exam scores.

Summary for inference using the t-distribution

Hypothesis tests. When applying the t-distribution for a hypothesis test, we proceed as follows:

- Write appropriate hypotheses.
- Verify conditions for using the t-distribution.

- One-sample or differences from paired data: the observations (or differences) must be independent and nearly normal. For larger sample sizes, we can relax the nearly normal requirement, e.g. slight skew is okay for sample sizes of 15, moderate skew for sample sizes of 30, and strong skew for sample sizes of 60.
- For a difference of means when the data are not paired: each sample mean must separately satisfy the one-sample conditions for the t-distribution, and the data in the groups must also be independent.

- Compute the point estimate of interest, the standard error, and the degrees of freedom. For df, use $n - 1$ for one sample, and for two samples use either statistical software or the smaller of $n_1 - 1$ and $n_2 - 1$.
- Compute the T-score and p-value.
- Make a conclusion based on the p-value, and write a conclusion in context and in plain language so anyone can understand the result.

Confidence intervals. Similarly, the following is how we generally computed a confidence interval using a t-distribution:

- Verify conditions for using the t-distribution. (See above.)
- Compute the point estimate of interest, the standard error, the degrees of freedom, and $t^\star_{df}$.
- Calculate the confidence interval using the general formula, point estimate $\pm\ t^\star_{df} SE$.
- Put the conclusions in context and in plain language so even non-statisticians can understand the results.

Examining the standard error formula (special topic)

The formula for the standard error of the difference in two means is similar to the formula for other standard errors. Recall that the standard error of a single mean, $\bar{x}_1$, can be approximated by

$$SE_{\bar{x}_1} = \frac{s_1}{\sqrt{n_1}}$$

where s_1 and n_1 represent the sample standard deviation and sample size.

The standard error of the difference of two sample means can be constructed from the standard errors of the separate sample means:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{SE^2_{\bar{x}_1} + SE^2_{\bar{x}_2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{11}$$

This special relationship follows from probability theory.

Pooled standard deviation estimate (special topic)

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 the 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:

$$s^2_{pooled} = \frac{s_1^2 \times (n_1 - 1) + s_2^2 \times (n_2 - 1)}{n_1 + n_2 - 2}$$

where n_1 and n_2 are the sample sizes, as before. To use this new statistic, we substitute s^2_{pooled} 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 $\bar{x}_1 - \bar{x}_2$, if the standard deviations of the two groups are equal.

Caution: Pool standard deviations only after careful consideration

A pooled standard deviation is used when a variance ratio test indicates it is appropriate. When the sample size is large and the condition may be adequately checked with data, the benefits of pooling the standard deviations greatly diminishes.

13 Paired t-test

When we conduct a statistical test looking at group means we want to detect a difference in the group means if there really is a difference. One of the things we can to do to increase our ability to detect a difference if there really is a difference is to take advantage of extra information in the data structure. One example of this type of extra information arises when we have so called 'before and after' measurements, repeat measurements from specific locations, or measurements that are in someway linked.

For example, say we are interested in understanding the effectiveness of a blood thinning drug. To test drug effectiveness we could obtain one set of measurements from 20 people that have not taken the drug and then another 20 measurements from 20 different people that have taken the drug and use a two-sample t-test to look for differences in mean blood clotting time. An alternative approach would be to take 20 measurements on blood clotting time from 20 people; then administer the drug to these same 20 people and take a second measurement on blood clotting time after the drug has been administered. Such a data set does not represent two random samples. Rather, the data set consists of observation pairs.

Although the context of a measurement before and after a treatment is the clearest example, there are many scenarios where it is possible to obtain data pairs. For example, if we send the same mineral sample to two different labs we end up with a 'pair' of measurements on the sample from different labs, not two independent observations. Similarly, if we ask husbands and wives a common set of question about happiness it might be appropriate to treat the husband and wife observations as pairs rather than independent observations.

Conceptually, the approach of obtaining observations pairs works to reduce sampling variability, or uncontrolled variation in the data sample. A reduction in uncontrolled variation (sampling variability) in the data set increases our ability to detect differences when differences exist.

13.1 Example with Wide Format Data

In this example we will be looking at before and after data from a contaminated site. Groundwater at the site was remediated using a Pump & Treat method with the intention of removing unacceptable levels of Total Petroleum Hydrocarbons (TPC). TCP is a general term encompassing hundreds of hydrocarbon based compounds. The paired measurements, recorded in micrograms per litre, were taken at observation wells around the contaminated site before and after the treatment.

We would like to determine whether the treatment has been effective - whether the differences between measurements before and after are different. To do this we will need to read in our data and calculate the differences between the two paired measurements so we can then create a boxplot, scatter plot, or histogram of these differences. From the plot we should have an idea of what the results of our formal statistical test will be. We will then run a formal paired t-test on the data. To reinforce how the paired t-test works we will also

show how a paired t-test is the same as a One Sample t-test on the difference series.

We will show how to create a plot title with two lines. This is done either by adding '\n' where you want the separation between the lines, or by pressing 'enter/return' for a new line in your script (see below). We will illustrate the '\n' approach. As you become more confident with plotting, how to put text on two lines is a good trick to know.

Generally you will read the data in from an MS Excel .csv file or similar, but here we input the data directly to provide a complete record of values.

```
#test
# read in our data (wide format)
TPH.remediation<- data.frame(
  before= c(1475.7, 1292.2, 1575.9, 1440.8, 1606.1, 1425.1, 1502.3, 1327.4,
            1526.4, 1422.4, 1540.4, 1550.2, 1544.7, 1630.1, 1454.4, 1398.0,
            1428.1, 1421.8),
  after= c(695.1, 706.1, 675.5, 706.6, 717.8, 729.4, 722.3, 668.2, 714.4,
           672.5, 665.8, 658.7, 694.6, 684.9, 704.2, 690.0, 702.0, 710.4))

# Look at the raw data
str(TPH.remediation)

## 'data.frame': 18 obs. of  2 variables:
##  $ before: num  1476 1292 1576 1441 1606 ...
##  $ after : num  695 706 676 707 718 ...

summary(TPH.remediation)

##      before         after
##  Min.   :1292   Min.   :658.7
##  1st Qu.:1423   1st Qu.:677.9
##  Median :1465   Median :698.5
##  Mean   :1476   Mean   :695.5
##  3rd Qu.:1544   3rd Qu.:709.5
##  Max.   :1630   Max.   :729.4
```

First, we want to create a plot. However, because we have paired measurements, we look at the difference between the series, and plot this difference. It is possible to specifiy the difference directly into the boxplot command. Also note that because we only have one group we use a horizontal boxplot.

```
# use \n for fine level control of where to split the heading text

with(TPH.remediation,boxplot(before-after,
col= "lightgray", boxwex= 1.2,
```

```
main= "Pump and Treat Groundwater Remediation of
Total Petroleum Hydrocarbons (TPH)",
xlab= "Differences in TPH (ug/L)", ylim= c(500,1000),
horizontal= TRUE))
```

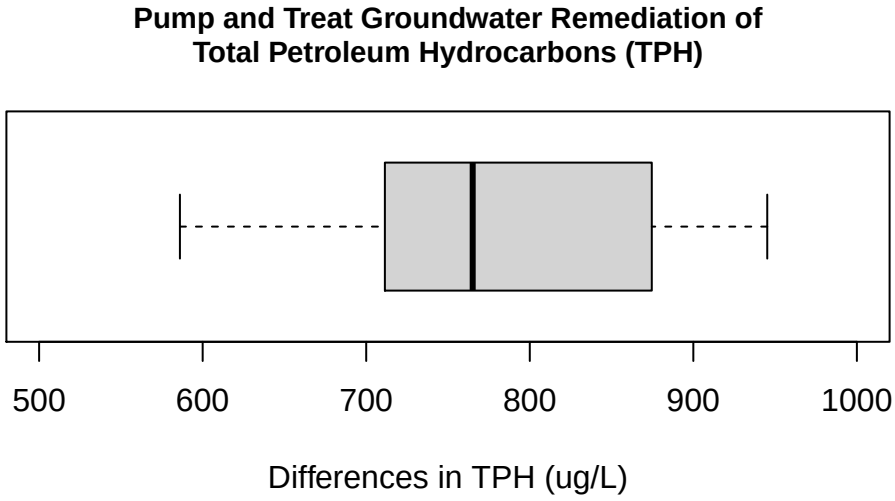


Figure 13.1: Box plot showing the differences in groundwater measurements of Total Petroleum Hydrocarbons (ug/L) at monitoring wells before and after pump and treat remediation.

Look at the way the difference has been calculated. We have subtracted the after measurement from the before measurement. When we look at the plot we can see that all the values are positive. This means that at every observation point the measurement after the intervention is lower. This is a good thing. For paired data, an alternative way to investigate the relationship is a scatter plot. The scatter plot approach is demonstrated in the advanced material.

From the boxplot (and our summary output) it looks like our remediation was successful. All the differences are positive, and if you have a look at the actual numbers again, all the measurements have dropped by around 50% from the original levels. We now conduct the formal paired t-test. For a paired t-test we do not conduct a variance ratio test. This is because the test works on the difference series, so there is only one series not two. As there is only one series we do not have two group variance values to compare.

Paired t-test using the paired t-test function

Step 1: Set the Null and Alternate Hypotheses

- Null hypothesis: The mean of the difference series is zero
- Alternate hypothesis: The mean of the difference series is not zero

Step 2: Implement the Paired t-test
Here we set a new parameter: `paired` to TRUE, as part of the t-test formula, and leave alpha at the default 0.05 level.

```
with(TPH.remediation, t.test(before, after, paired = TRUE))

##
##  Paired t-test
##
## data:  before and after
## t = 34.464, df = 17, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  732.4325 827.9564
## sample estimates:
## mean of the differences
##                780.1944
```

Step 3: Interpret the Results
From the output we see the `p-value = < 2.2e-16`, which we would generally report as < 0.001, which is smaller than 0.05. So, we:

- **Reject** the null hypothesis that the means are equal.

Similar to other tests the output also provides the 95% confidence interval (`732.4, 828 ug/L`) for the actual difference, and reports the mean difference (`780.2 ug/L`).

In words we say we have **sufficient evidence** to say the concentration of Total Petroleum Hydrocarbons is lower after remediation was conducted (we know it's lower from looking at the data values and boxplot).

That the remediation has had an impact is great news; but we still needs to be compare actual pollution levels to acceptable levels for the relevant land use. What if TPC levels needed to be, on average, below 800 ug/L? Our test does not answer this question. To answer that question we would need to conduct a one sample t-test on the 'after' measurements, setting `mu` to 800. To find a statistically significant difference is one thing. To find a difference that is practically important is something else.

subsubsection*Paired t-test using the one sample t-test on the difference series When we selected a paired t-test, **R** will automatically create a difference series to use, and apply the one sample t-test to the difference series. To check this we can manually create the difference series and then apply the one sample t-test directly to the difference series that

we create. To create the difference series we will create the variable `diff` in the dataframe `TPH.remediation` by subtracting the after values from the before values.

```
# calculate difference between measurements
# this creates a new variable 'diff' in our data frame
TPH.remediation$diff<- TPH.remediation$before - TPH.remediation$after

# check what the data looks like after we have added the new variable
str(TPH.remediation)

## 'data.frame': 18 obs. of  3 variables:
##  $ before: num  1476 1292 1576 1441 1606 ...
##  $ after : num  695 706 676 707 718 ...
##  $ diff  : num  781 586 900 734 888 ...

summary(TPH.remediation)

##      before         after            diff
##  Min.   :1292   Min.   :658.7   Min.   :586.1
##  1st Qu.:1423   1st Qu.:677.9   1st Qu.:715.1
##  Median :1465   Median :698.5   Median :765.1
##  Mean   :1476   Mean   :695.5   Mean   :780.2
##  3rd Qu.:1544   3rd Qu.:709.5   3rd Qu.:868.5
##  Max.   :1630   Max.   :729.4   Max.   :945.2
```

As we have created a difference series we can use this new variable directly to create a boxplot.

```
with(TPH.remediation,boxplot(diff,
     col= "lightgray", boxwex= 1.2,
     main= "Pump and Treat Groundwater Remediation of
      Total Petroleum Hydrocarbons (TPH)",  # second line of title
     xlab= "Differences in TPH (ug/L)", ylim= c(500,1000),
      horizontal= TRUE))
```

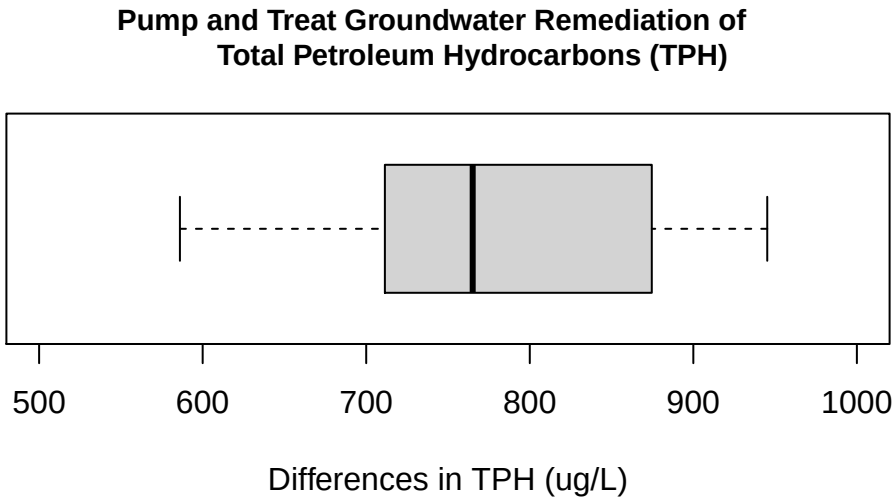


Figure 13.2: Box plot showing the differences in groundwater measurements of Total Petroleum Hydrocarbons (ug/L) at monitoring wells before and after pump and treat remediation.

Our **Null hypothesis** for the t-test is that the mean of the difference series is zero; our **Alternate hypothesis** is that the mean of the difference series is not zero. To implement this test we set `mu`=0 and leave alpha (`conf.level`) at the default value of 0.05.

```
with(TPH.remediation, t.test(diff))

##
##  One Sample t-test
##
## data:  diff
## t = 34.464, df = 17, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  732.4325 827.9564
## sample estimates:
## mean of x
##  780.1944
```

Note that the one sample t-value output is the same as for the paired t-test. This is because we are conducting exactly the same test. We are also given the same confidence interval, and the mean difference between the paired values (`780.2 ug/L`) matches the mean difference value for the paired t-test. So, we have shown that a paired t-test is the same as a one-sample t-test on the difference series.

13.2 Advanced: Example with Long Format Data

For the paired t-test it is easier to have the data in wide format. Although the process for a paired t-test is nearly the same for long format data, it is not that easy to plot the data when it is in long format. Here we work through the process and introduce a new command, *subset()*. We use the *subset()* command to extract the data for each of our groups by specifying parameter `obs` with the character string (i.e. name) of the factor level (i.e. group) to select. We also use a double equal sign '==' to tell R to select for something exactly equal to the name we provide. Alternatively you can run a paired t-test directly on long format data (see below), but the most natural way to illustrate paired data is by plotting the differences in a histogram or boxplot, or the actual data via a scatter plot.

Our data here is paired (before/after measurements at the same location), but remember paired data can also be before and after measurements from the same subject, or a comparison of treatments or measurement methods applied to the same subject or site. Here we have measurements of percent coral cover from a reef before and after a marine heat wave. Prolonged and extreme heat stress are known to impact the symbiotic relationship between coral and algae living within its tissues. The algae are expelled causing coral bleaching and, in more severe cases, coral death. We will be testing whether coral cover has significantly changed since the heat wave.

```
# read in our data (wide format)
coral<- data.frame(
  cover= c(19, 20, 35, 13, 22, 26, 19, 31, 35, 29, 10, 31, 8, 20, 19, 15, 22, 19,
           10, 8), obs= c(rep("before", times =10), rep("after", times= 10)))

# subset to extract values for each group
coral.before <- subset(coral, obs== "before")
coral.after <- subset(coral, obs== "after")
# calculate differences (creates a new dataframe of just the differences)
coral.diff<- coral.before$cover - coral.after$cover

# get summary & structure and create a histogram of our differences
str(coral.diff)
```

```
##  num [1:10] 9 -11 27 -7 3 11 -3 12 25 21
```

```
summary(coral.diff)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -11.00   -1.50   10.00    8.70   18.75   27.00
```

```
hist(coral.diff, breaks = 'fd',
     col= "lightgray",
     # write title on two lines (use \n or 'return' for new line)
     main= "Differences in Coral Cover
     Before and After a Marine Heat Wave",
     xlab= "Differences in percent cover", las= 1)
```

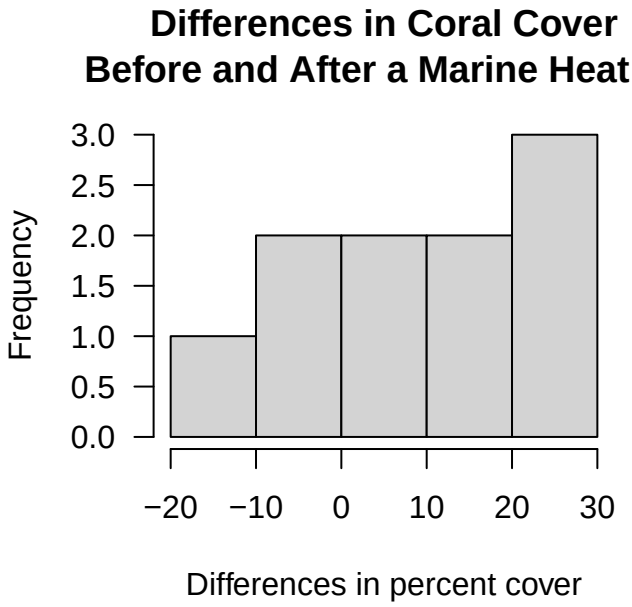


Figure 13.3: Histogram of the differences in paired measurements of percent coral cover from before and after a marine heat wave.

In this instance the histogram does not provide a strong visual cue. There are few data points and we have both positive and negative values. An alternative way of looking at paired data is to create a scatter plot with a 45° line. If the data set is in wide format the plot is easy to create. However, here we have to pull the observations from the two objects we created. If all the observations fall to one side of the 45° line it tells you something.

```
plot(coral.before$cover,coral.after$cover,
     col= "lightgray",
     pch=19,
     main= "Marine Heat Wave Effect: \n Coral Cover Before and After",
     xlab= "Before measurement",
     ylab= "After measurement",
     xlim = c(5,40),
     ylim = c(5,40),
    # cex.main=2,
     #cex.lab=1.5,
     #cex.axis=1.5,
     las= 1)
abline(0,1) #  45 degress line
```

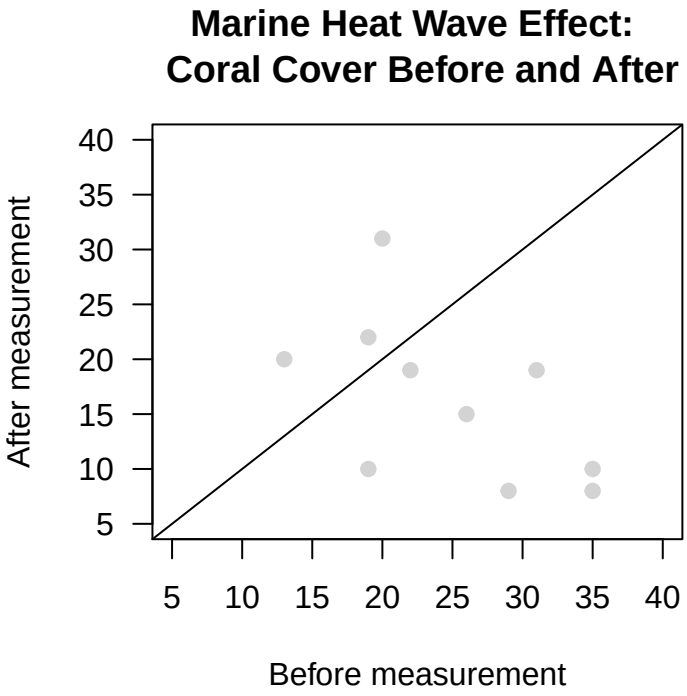


Figure 13.4: Scatter plot of paired measurements of percent coral cover from before and after a marine heat wave.

Here we see that seven out of ten of the dots fall in the 'before' triangle. This means that in seven out of ten cases the coral cover was greater before the heatwave than after the heatwave. If the dots cluster around the 45° line the differences are likely to be due to random sampling effects. The further away the dots are from the 45° line the stronger the evidence of a difference. Here the dots in the 'after' triangle are not that far from the 45° line, while many of the dots in the 'before' triangle are quite far from the 45° line. This suggests that the heatwave has had a negative effect, but it is something that we still need to check with a formal test.

Step 1: Set the Null and Alternate Hypotheses

- Null hypothesis: The mean of the difference series is zero
- Alternate hypothesis: The mean of the difference series is not zero

Step 2: Implement the Paired T-Test

Here we set `paired` to TRUE, and leave alpha at the default value of 0.05. Note that for the t-test, because the data are in long format we use '~' rather than ','.

```
with(coral, t.test(cover ~ obs, paired = TRUE))

##
```

```
##  Paired t-test
##
## data:  cover by obs
## t = -2.0816, df = 9, p-value = 0.06709
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.1545661   0.7545661
## sample estimates:
## mean of the differences
##                    -8.7
```

Step 3: Interpret the Results

From the output we see the `p-value = 0.06709`, or 0.0671 is larger than 0.05. So, we:

- **Fail to Reject** the null hypothesis that the means are equal.

Is this what you expected based on the plot?

For completness we can also apply the one sample t-test directly to the difference series. Our **Null hypothesis** is that the mean of the difference series is zero; our **Alternate hypothesis** is that the mean of the difference series is not equal to zero. As always alpha is 0.05.

```
t.test(coral.diff)
```

```
##
##  One Sample t-test
##
## data:  coral.diff
## t = 2.0816, df = 9, p-value = 0.06709
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.7545661 18.1545661
## sample estimates:
## mean of x
##       8.7
```

If you compare the information in the one sample t-test to that in the paired t-test you will see that they are the same (you can ignore the difference in the sign on the t-statistic and the . This is because you are testing the same thing. But what about the result - what does it mean for the reef? Under our default alpha and 95% confidence interval the p-value is deemed not statistically significant. However, what if our alpha was 0.1 (90% confidence)? All of a sudden we have a significant result!

If the outcome of this research was to determine whether to implement a policy to better protect the reef, what would you decide? This is why there is much controversy surrounding

the standard 0.05 p-value decision rule, and why it is absolutely essential that the context of your data and implications of research outcomes are taken into consideration.

The other thing to be discussed are the negative differences we noticed. Where could these have come from? A heat wave is very unlikely to have increased coral cover. By critically thinking about the results and data we might suggest that this is more likely due to other factors, such as sampling error or possibly some sort of measurement error. Seeing these results in this context might lead you to look deeper into sources of error and put less confidence in the results of your analysis. In other contexts, this result might be perfectly normal.

14 Power

In section 9.3 we displayed four scenarios that cover how our conclusions from hypothesis testing can be correct and incorrect. This table is reproduced here as table 14.1. The notion of a p-value was introduced and this is the probability of rejecting the null hypothesis when in fact the null hypothesis is true. That is, the probability of making a Type 1 error.

What about the the probability of making a Type 2 error? We are also interested in this - when the alternative hypothesis is true, we want our hypothesis test to pick this up and lead us to rejecting the null hypothesis in favour of the alternative. The probability of making a Type 2 error is denoted by the Greek letter β (beta). A hypothesis test that has a high probability of alerting us to when we should reject the null hypothesis is called a test of high *power*. We quantify this by defining the *power* of a hypothesis test as

$$\begin{aligned} \text{Power} &= 1 - P(\text{Type 2 Error}) \\ &= 1 - \beta. \end{aligned}$$

Table 14.1: Four different scenarios for hypothesis tests.

		Test conclusion	
		do not reject H_0	reject H_0 in favour of H_A
Truth	H_0 true	okay $(1-\alpha)$	Type 1 Error (α)
	H_A true	Type 2 Error (β)	okay $(1-\beta)$

When we say that power is the probability of rejecting H_0 when H_A is true, that might (hopefully!) lead you to wonder: *what is the distribution of the variable of interest under the alternative hypothesis?* The null hypothesis in our examples for the population mean had a specific value for the population mean e.g. $\mu = 7$ in the example in Guided Practice 15 on hours of sleep per night. The alternative hypothesis was just that the true average was greater than 7.

H_0: $\mu = 7$.

H_A: $\mu > 7$.

The alternative hypothesis doesn't have a single reference parameter value so we need to define one. We do this by defining the size of the shift from the null value that we wish to detect, the *effect size*. Thinking of the sleep study, the alternative hypothesis of $\mu > 7$ would be true if the actual mean was 7.01 or 7.25 hours of sleep. However, if the true average was 7.01 hours of sleep it is much less likely that our test would pick up this difference than if the true average was 7.25 hours. Defining this effect size requires an understanding of the subject matter to which the research relates. For example, in a clinical sense, 7.01 hours of sleep is not significantly more than 7 hours, but 7.25 hours might be.

TIP: Statistical versus practical significance

This is another good place to remind ourselves of the difference between **statistical** significance and **scientific/clinical/practical** significance. As a thought experiment, just think that if you have a large enough sample size, and test

H_0: $\mu = 0$.

H_A: $\mu \neq 0$.

you will reject the null hypothesis every time because very few natural phenomena have a true mean of *exactly* zero. Even if the true mean is 0.00001 you will reject the hypothesis that $\mu = 0$. To put it bluntly, there is growing frustration in the scientific and statistical community that an over-emphasis on the p-value in presenting scientific results has distracted some researchers from the relevance of the magnitude of the effect that they are observing. For the viewpoint of an international professional body on this matter, see https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf

14.1 Power calculation in R

Example 14.1 Power calculation to determine sample size for one-sample t-test

Using data from an aquaculture farm rearing rainbow trout, we will illustrate power calculation in a t-test. If it is important to the farmer to ascertain whether the mean weight of the trout is at least 500g, will a one-sided t-test be able to detect this? That depends on several factors

1. how much great than 500g the true mean weight is
2. the size of the sample
3. the power of the test, i.e. the probability we reject the null hypothesis when the alternative is true
4. how variable the weight of trout is
5. the significance level used for the statistical test.

To start with, let's see how large a sample needs to be taken to reject at the 0.05 significance level the null hypothesis with probability 0.9 when the true mean is 510g. We estimate the variation of the trout weights using an existing available sample. The R function `power.t.test` gives the sample size required.

```
# read in our data and get it in the correct format
Trout <- c(508, 479, 545, 531, 559, 422, 547, 525, 420, 491, 508, 511, 569,
           453, 533, 460, 523, 540, 463, 502)
sd(Trout)
```

```
## [1] 43.11609

power.t.test(n = NULL,          # want function to solve for n
             delta = 10,        # effect 10g > null hypothesis
             sd = sd(Trout),    # estimate std dev from existing data
             sig.level = 0.05, # the usual significance level
             power = 0.9,       # prob reject null if true mean 510g
             type = "one.sample",
             alternative = "one.sided")


##
##      One-sample t test power calculation
##
##               n = 160.5642
##           delta = 10
##              sd = 43.11609
##       sig.level = 0.05
##           power = 0.9
##     alternative = one.sided
```

It turns out that we would need a sample of 161 trout in order to have a one-sided t-test with the required power. If the true mean were higher, or the standard deviation were lower then a smaller sample would be sufficient. We won't show this mathematically, but it makes sense intuitively that detecting a difference is easier if the true difference (delta) is greater or if the parameter (mean) is easier to estimate accurately (smaller standard deviation).

To see this for the current example, firstly increase the effect size so that we are interested in detecting an effect that is 20g above the mean of the null hypothesis.

```
power.t.test(n = NULL,
             delta = 20,        # effect 20g > null hypothesis
             sd = sd(Trout),
             sig.level = 0.05,
             power = 0.9,
             type = "one.sample",
             alternative = "one.sided")


##
##      One-sample t test power calculation
##
##               n = 41.19163
##           delta = 20
##              sd = 43.11609
```

```
##       sig.level = 0.05
##           power = 0.9
##     alternative = one.sided
```

This increase in the effect size from 10g to 20g decreases the required sample size from 161 trout to 42. Now let's look at the effect on the required sample size of reduced variance. Here we will decrease the standard deviation in the example by 20%.

```
power.t.test(n = NULL,
             delta = 10,
             sd = sd(Trout)*0.8,  #sd 20% less than previously
             sig.level = 0.05,
             power = 0.9,
             type = "one.sample",
             alternative = "one.sided")


##
##      One-sample t test power calculation
##
##               n = 103.257
##           delta = 10
##              sd = 34.49287
##       sig.level = 0.05
##           power = 0.9
##     alternative = one.sided
```

This decrease in the standard deviation from 43.12 to 34.49 decreases the required sample size from 161 trout to 104.

This example shows the most common type of power calculation performed as part of experimental design. That is, given what we know (or can estimate) about the variation in the population, and given the magnitude of the effect that we wish to be able to detect, what size sample is required for the experiment. Another way to look at the power calculation is: based on a fixed sample size, what magnitude effect could be detected with given power?

Example 14.2 Power calculation to determine effect size for one-sample t-test

The current sample consists of 20 measurements of trout weight. For this sample size, and the estimate of variation from it, what size effect will be we able to detect with 90% power and 5% significance level?

```
power.t.test(n = 20,           # current sample size is 20
             delta = NULL,     # want to calculate effect we can detect
             sd = sd(Trout),   # estimate std dev from existing data
```

```
             sig.level = 0.05, # the usual significance level
             power = 0.9,      # prob reject null if true mean 510g
             type = "one.sample",
             alternative = "one.sided")

##
##      One-sample t test power calculation
##
##               n = 20
##           delta = 29.28238
##              sd = 43.11609
##       sig.level = 0.05
##           power = 0.9
##     alternative = one.sided
```

With the current sample we would be able to detect when the mean weight rose to 530 with the required power and signifiance level of the one-sided t-test.

Any statistical hypothesis test has an associated power. We will stick to t-tests in this chapter, however power calculations can be applied to more complicated hypothesis tests, such as whether the slope parameter in a linear regression is different to zero, or whether a time series of monthly data displays a seasonal effect.

The next variant of the t-test we will examine is a two-sample t-test. Suppoe that we wish to test whether a particular treatment given to the farmed trout, let's say a new type of fish feed, results in an increased mean weight of the fish. We could test this by randomly assigning fish to two separate groups, one group to be given the usual feed and another group to be given the new type. After a given period of time the fish can be weighed and we can test whether there is a statistically significant increase in mean weight using a two-sample t-test. One question we need to ask before conducting such an experiment is: how many trout do we need to have in each group to make the experiement worthwhile and cost-effective?

As in the single-sample t-test, we need to specify the magnitude of the effect that we consider to be of practical significance. In the one-sample scenario, this was the difference from the mean under the null hypothesis. In the two-sample scenario, we need to specify the difference between the two groups that would be of practical significance. Let's say for now that an increase in mean weight of 20g would be of interest to us from a commercial point of view. Because we are only interested in whether or not there is an *increase* in weight, the test is a one-sided test.

Example 14.3 Power calculation to determine sample size for two-sample t-test

```
power.t.test(n = NULL,          # want function to solve for n
             delta = 20,        # want to calculate effect we can detect
             sd = sd(Trout),    # estimate std dev from existing data
             sig.level = 0.05, # the usual significance level
             power = 0.9,       # prob reject null if true mean 510g
             type = "two.sample",
             alternative = "one.sided")


##
##      Two-sample t test power calculation
##
##               n = 80.2864
##           delta = 20
##              sd = 43.11609
##       sig.level = 0.05
##           power = 0.9
##     alternative = one.sided
##
## NOTE: n is number in *each* group
```

The sample size required for a two-sided t-test with the required significance level and power is 81 for a difference of 20g. As the helpful R output reminds us, this is the required sample size for each group.

14.2 Power calculations in experimental design

Power calculations form an important part of **experimental design**. These calculations are done *before* any data are collected. Imagine completing a time-consuming and costly experiment, only to examine the data and discover that

1. given the variability in the data, your sample size would only be able to detect an effect if it was enormous, or

2. you collected a sample that was much bigger than was actually necessary and wasted a lot of money/time/plants/animals/good-will etc.

Power calculations are therefore required as part of the ethics approval process for experiments requiring human or animal subjects.

We have seen in the examples above that the power calculations depend on the type of hypothesis that is to be tested with the data. We have chosen simple examples above, but there are corresponding power calculations for more complex hypotheses. For example, these could be testing hypotheses of longitudinal effects in repeated measures data, or testing hypotheses about the parameters in regression models.

If power calculations are smart practice to avoid waste, and power calculations require the hypotheses to be defined, then that means that researchers need to have the analyes and hypothesis tests determined *before* data are collected. It's tempting to jump right in and collect data, however without proper experimental design a lot of painstakingly collected data can be worthless. The need for careful experimental design might seem obvious (hopefully!) but it does still happen that major problems in a study are only picked up during the analysis phase, when it can be too late to rectify them.

15 Variance Homogeneity: many groups

Generally we are interested in testing whether or not there is a difference in the group means. When testing for differences in group means the specific test statistic formula to use depends on whether or not the group variances are equal. There is one formula for the case when the groups have equal variance, and a second formula to use when the group variances are not equal. When we are comparing two groups we use a variance ratio test. When there are multiple groups a possible test of whether the group variances are all equal is the Bartlett test.

15.1 The Bartlett Test

For the Bartlett test, the null hypothesis is that all the group variances are the same. The alternate hypothesis is that at least one of the group variances is different. The formula to calculate the test statistic is complicated, but it is still possible to obtain an intuitive understanding of the way the test works. The test works by comparing the variance calculated when the data are pooled together to form a single group to the variance calculated separately for each group. If the differences between the variance for each group and the pooled variance are large the conclusion drawn is that the group variances are not the same. The test statistic follows the chi-square distribution, but to determine what constitutes a 'large' value for the differences we use the the p-value decision rule. The null hypothesis that all the group variances are equal is rejected when the p-value is small.

15.2 Equal Variance Testing - Multiple Groups

To illustrate the test We will use the base **R** data set **chickwts**. This data set comes installed with **R**, which means we do not have to read in the data first. The data set has the weights of chickens that are on six different diets (feed regimes). Here we will be testing whether or not is appropriate to assume the variance for all six groups is the same using the Bartlett test function `bartlett.test()`.

```
# Normally we would need to read the data in first but in this instance it
# is loaded already
str(chickwts)  # the data is in long format

## 'data.frame': 71 obs. of  2 variables:
##  $ weight: num  179 160 136 227 217 168 108 124 143 140 ...
##  $ feed  : Factor w/ 6 levels "casein","horsebean",..: 2 2 2 2 2 2 2 2 2 2 ...

summary(chickwts)
```

```
##      weight             feed
##  Min.   :108.0   casein   :12
##  1st Qu.:204.5   horsebean:10
##  Median :258.0   linseed  :12
##  Mean   :261.3   meatmeal :11
##  3rd Qu.:323.5   soybean  :14
##  Max.   :423.0   sunflower:12
```

Note the data structure. Can you see that there are six different feed types? Can you see how many observations there are per group? When the data are in long format the 5-number summary information ignores the grouping structure, so until we create a boxplot it is difficult to understand the nature of the data from just the summary information. One of the advantages of using the long format for data management with many groups is that it makes it easier to create a boxplot.

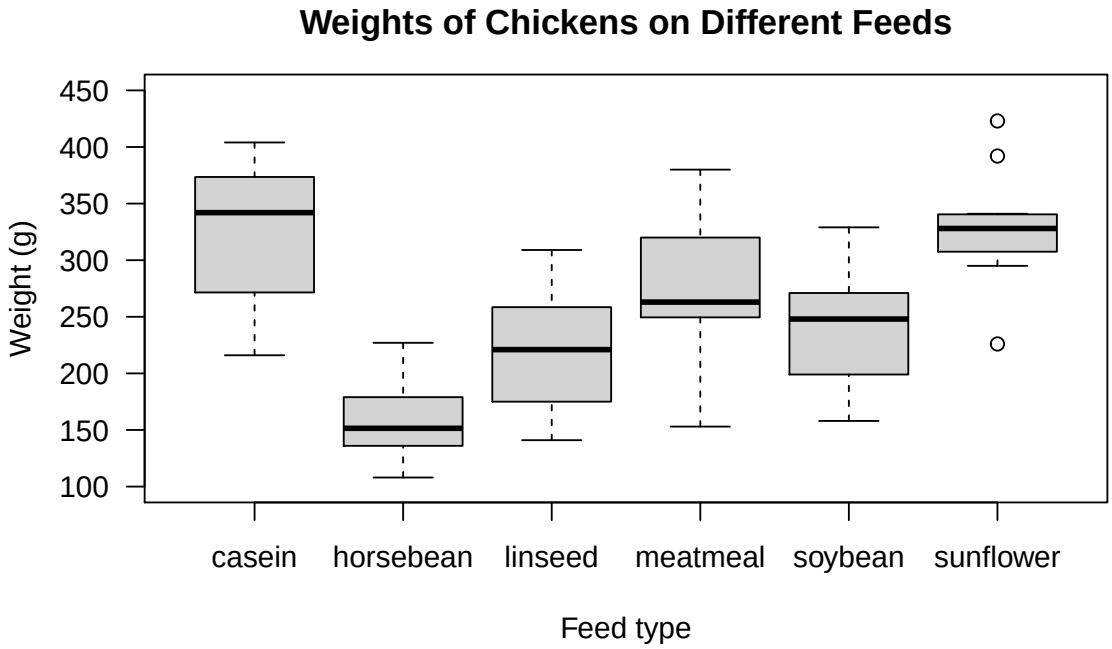


Figure 15.1: Boxplot of chicken weight distributions from six different feed types

Based on the boxplot, do the variances look equal? There definitely looks to be some difference in the means, but we cannot appropriately test the means before knowing if all the group variances are equal.

There appears to be some difference between the group variances, particularly with the sunflower feed, but remember we also have a fairly small number of observations in each group. Also note that for the sunflower group several 'outliers' have been identified. These

values will work to increase the variance estimate. Overall, the difference in the spread of observations between groups does not look extreme. Based on this plot I would not be surprised if the formal test said it was safe to assume the group variances are equal.

Let's run our test.

Step 1: Set the Null and Alternate Hypotheses

- Null hypothesis: The variance is the same for all groups
- Alternate hypothesis: The variance is not the same for all groups

Step 2: Implement Bartlett Test

For the test we specify the continuous and factor variable separated by a '~', in this order, because the data is in long format. Our alpha will be set as the default: 0.05.

```
with(chickwts, bartlett.test(weight ~ feed))  # long data format, use ~ not ,

##
##  Bartlett test of homogeneity of variances
##
## data:  weight by feed
## Bartlett's K-squared = 3.2597, df = 5, p-value = 0.66
```

Step 3: Interpret the Results

From the output we see the `p-value = 0.66`. Since 0.66 is greater than 0.05 (alpha), we:

- **Fail to reject** the null hypothesis that the group variances are the same

What does this mean? It means, we have do **not** have **sufficient evidence** to say the variance is different across the groups. The test output also reports a K-squared value (`3.26`). Technically this is the test statistic - and it is this value we have used the p-value decision rule to conclude is not large. The df (`5`) represents our degrees of freedom, or the number of levels we have minus 1 (6 feeds - 1 = 5). As large values for the test statistic lead to a rejection of the null, we know that in this instance (`3.26`) is not large.

If we were to move on to conduct an ANOVA test to look for differences in the group means, we would conduct the test using the assumption that the group variances are equal. Using the function *oneway.test()*, this means we set the parameter `var.equal` to TRUE. If we reject the ANOVA test null hypothesis and move on to pair-wise t-tests then we would set the parameter `pool.sd` to TRUE when using the *pairwise.t.test()* function.

15.3 Advanced: Technical Note

As with the variance ratio test there are some known issues with the Bartlett test. For example, when the distributions are more peaked than a normal distribution (leptokurtic)

the true alpha level is greater than the stated alpha level of the test. The chance of making a type I error is therefore higher than we think. Conversely, with distributions that are less peaked than a normal distribution (platykurtic) the true alpha level is less than the stated alpha level of the test, so the chance of making a type II error is increased. No relatively accessible reference for this issue has been identified, but a classic (relatively technical) reference that covers the issues is provided below.

For those interested in robust tests there are alternative tests that can be used. For example, there is Levene's test, which is available in the `car` package; and also the non-parametric Fligner-Killeen test. Details on these tests can be obtained from the `R` help pages.

Reference

Box, G. E. (1953). Non-normality and tests on variances. Biometrika, 40(3/4), 318-335.

16 ANOVA

The ANOVA test is a classical statistical test associated with R. Fisher. The test is suited to situations where we have observations grouped by a categorical variable and we want to understand whether or not there are differences in the group means. The mathematical notation for ANOVA is horrible, but the intuition is relatively easy to understand.

The ANOVA test statistic follows the F-distribution, and large values are evidence against the null hypothesis that the group means are all the same. In practice we can think of the test as dividing the observed variation in the data into: (i) the variation between the group means and the grand mean (the average across all the groups); and (ii) the variation within each group.

If the variation between the group means and the grand mean is high, and the variation within groups is small, this suggests there are true differences in the group means. On the other hand, if the variation between the group means and the grand mean is small, and the variation within groups is large, this suggests there are no real differences in the group means i.e. the variations we observe is just sampling variation.

For the ANOVA test, if we observe a p-value of less than 0.05 we say we have sufficient evidence to reject the null hypothesis that all the group means are the same. If we reject the null we then move on to pair-wise t-tests to try and determine which specific group means are different. For the pair-wise t-tests we use an adjustment method to control the Type I error rate. The need to adjust p-values to control the Type I error rate is one of the reasons we need specialist software.

If we observe a p-value that is greater than 0.05 we say we have insufficient evidence to reject the null hypothesis that all the group means are the same. If we do not reject the null hypothesis we do not proceed to pair-wise t-tests. We just stop out analysis at the ANOVA test. We have concluded that there are no significant differences so we stop.

Similar to the case for the t-test, for the ANOVA test there are two possible formulas that we can use: one formula when the group variances are the same, and one formula for when the group variances are not the same. So, before we conduct an ANOVA test we have to conduct a pre-test to establish whether the group variances are the same or not. The test we use for this in the Bartlett test. If we reject the null for the ANOVA test we then go on to look for the specific differences between groups. This means that when we conduct an ANOVA test we usually have three tests to conduct:

1. A Bartlett test to check whether the group variances are the same
2. The actual ANOVA test (with either equal or unequal variance)
3. Pair-wise t-tests, with an adjustment for multiple comparisons (if we reject the null for the ANOVA test).

We will work through these steps below. As with all previous statistical tests, we will

set the alpha level at 0.05.

16.1 ANOVA - One Factor

For this example we will use the base **R** data set **chickwts**, which has the weights of chickens fed six different types of feed. The data is in long format. This means we have one continuous variable (weight) and one factor variable (feed) that has six levels. When working with multiple groups it is easiest to work with data in long format.

Note: This is the same data set used in the Equal Variance Testing in R reference guide. What we are trying to do is help a farmer identify which feed(s) promote the greatest weight gain in chickens.

Let's start by looking at our data and creating a box plot:

```
# get summary & structure and create a boxplot of our differences
str(chickwts)

## 'data.frame': 71 obs. of  2 variables:
##  $ weight: num  179 160 136 227 217 168 108 124 143 140 ...
##  $ feed  : Factor w/ 6 levels "casein","horsebean",..: 2 2 2 2 2 2 2 2 2 2 ...

summary(chickwts)

##      weight             feed
##  Min.   :108.0   casein   :12
##  1st Qu.:204.5   horsebean:10
##  Median :258.0   linseed  :12
##  Mean   :261.3   meatmeal :11
##  3rd Qu.:323.5   soybean  :14
##  Max.   :423.0   sunflower:12

with(chickwts,boxplot(weight~feed,
     col= "lightgray",
     main= "",
     xlab= "Feed type", ylab= "Weight (g)", ylim= c(100,450), las= 1))
```

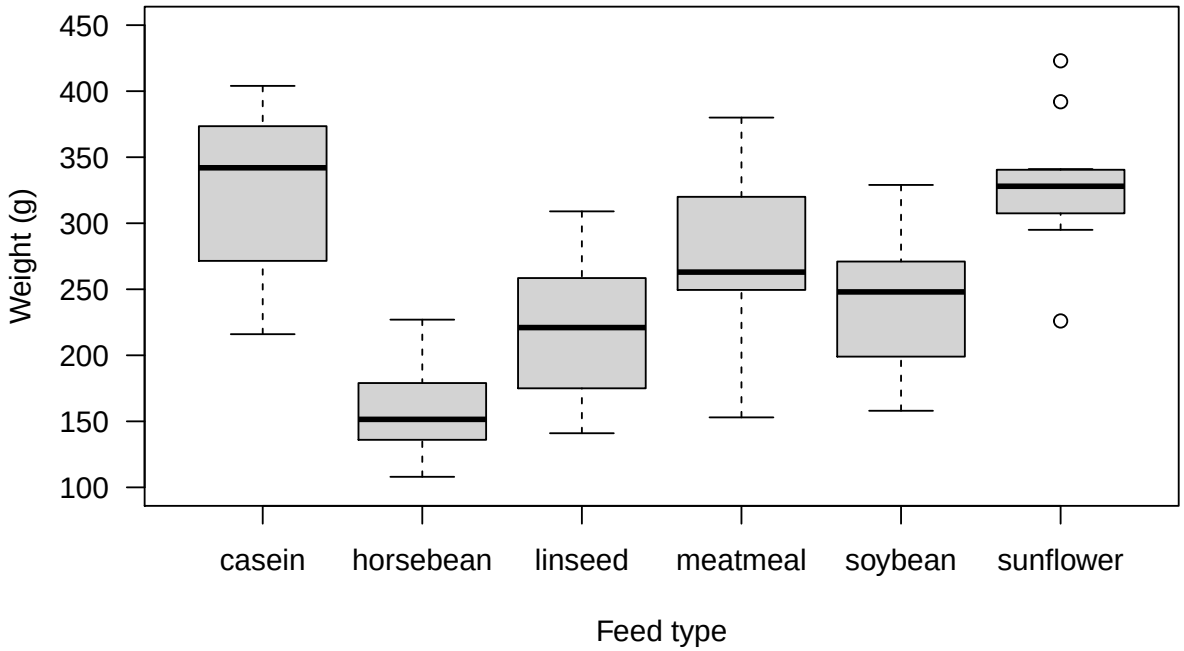


Figure 16.1: Chicken weight distributions for six different feed types

Based on the boxplot there look to be clear differences between at least some of the means, but it is not easy to see what the actual values are. Let's use a new function – *tapply()* – to calculate the mean and standard deviation for each of our groups. Here you are applying the function *mean* and *sd* to the vector `weight`, grouped by `feed`. The *tapply()* function works when we have data in long format. The function generates results similar to the way results are generated using the pivot table function in MS Excel. Personally, I prefer to use MS Excel to work through data management issues, but you can do similar things with **R**.

```
with(chickwts, tapply(weight, feed, mean))

##    casein horsebean   linseed  meatmeal   soybean sunflower
##  323.5833  160.2000  218.7500  276.9091  246.4286  328.9167

with(chickwts, tapply(weight, feed, sd))

##    casein horsebean   linseed  meatmeal   soybean sunflower
##  64.43384  38.62584  52.23570  64.90062  54.12907  48.83638
```

As we saw in our boxplot, there is quite a difference between some of our group means, particularly horse bean on the lower end, and sunflower and casein on the higher end. What about our standard deviations? Remember, the standard deviation is the square root of the variance so the group standard deviations gives us an idea of whether the group variances

are equal or not. The units of measurement for the standard deviation are just a bit easier to interpret than the units of measurement for the variance. Look carefully at the reported standard deviation values and also at the boxplot. Can you see the link between the visual display of the data and the group standard deviation values?

Now let's formally test for equal variance across the groups. Because we have multiple groups the test we use is the Bartlett Test of Homogeneity of Variances. The **Null hypothesis** for the test is that the group variances are all equal; the **Alternate hypothesis** is that the group variances not are equal. The implementation of the test follows the standard format we use for most tests. First we use the *with()* command to direct **R** to the dataset we want to use. Next we use the *bartlett.test()* command to tell **R** what function to apply. Finally we specify the column names that contain the data we are working with. Although this explanation is quite brief, there is a separate reference guide on the Bartlett test that explains the steps in greater detail.

```
with(chickwts, bartlett.test(weight ~ feed))  # long data format, use ~ not ,

##
##  Bartlett test of homogeneity of variances
##
## data:  weight by feed
## Bartlett's K-squared = 3.2597, df = 5, p-value = 0.66
```

As the test `p-value = 0.66`, which is greater than 0.05, we do not reject the null hypothesis that the group variances are all equal: we have insufficient evidence to reject the null.

As such we set the parameter `var.equal` to TRUE in our ANOVA test of whether the groups means are the same. Note: If we reject the null for the Bartlett test (i.e. we have a test p-value that is less than 0.05) when we conduct the ANOVA test we set the parameter `var.equal` to FALSE.

To implement the ANOVA test we use the function *oneway.test()* , and so for our formal test we have:

Step 1: Set the Null and Alternate Hypotheses

- Null hypothesis: The group means are all equal
- Alternate hypothesis: At least one group mean is different from the other groups

Step 2: Implement Analysis of Variance Test

```
with(chickwts, oneway.test(weight ~ feed, var.equal = TRUE))

##
##  One-way analysis of means
##
```

```
## data:  weight and feed
## F = 15.365, num df = 5, denom df = 65, p-value = 5.936e-10
```

Step 3: Interpret the Results
From the output we see the `p-value = 5.936e-10`; which we report as < 0.001. As the test p-value is smaller than 0.05 we:

- **Reject** the null hypothesis that all the means are equal.

In our output we can also see the F-statistic (`15.36`). This is the statistic that our p-value is based on. For large numbers we reject the null.

So what now? If you were a farmer wanting to know which feed to give your chickens, which feed would you choose? You could decide from your boxplot, likely choosing casein or sunflower, but the figure doesn't tell you if there is a feed which is statistically the highest, and the ANOVA test only tells you that the feeds are not ALL the same. This is where pair-wise comparisons can be helpful.

Pair-wise Comparisons

Let's come back to our research question - which feed(s) should the farmer consider based on the weights of chickens eating that feed. Here we will use the function *pairwise.t.test()* to compare all our individual group means. With this function the parameter `pool.sd` is the equivalent of `var.equal` for the *t.test()*. We use the Bartlett test result to decide whether we set this parameter to TRUE or FALSE.

If we do not reject the null for the Bartlett test we set the parameter to TRUE. If we reject the null for the Bartlett test we set the parameter to FALSE. So, for this example we will set it to TRUE.

The *pairwise.t.test()* function does two things that make our life easy. First, it calculates all the different possible pair-wise combinations for us at once. Rather than having to type the *t.test()* command 15 times (one for each pair-wise combination) we just need to write one command. The second thing that the *pairwise.t.test()* command does for us is to automatically adjust the reported p-values to control the Type I error rate. The automatic adjustment makes life easy for us as we can still apply our standard p-value decision rule to the results, where we know that the values have been adjusted to address the multiple comparison issue. This is something that only specialist software such as **R** is able to do, and it really saves a lot of time compared to the alternative of working through the adjustments manually.

It is possible to use a variety of different adjustments for multiple comparisons. For example, by changing the parameter `p.adjust.method` and providing "none" for no adjustment (lower p-values), or "bon" for a more restrictive adjustment (higher p-values). The default adjustment used by R is the holm adjustment, and this is a good choice. As such we will not be changing this parameter value. So conducting all the required comparisons is quite easy.

```r
with(chickwts, pairwise.t.test(weight, feed, pool.sd = TRUE))
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  weight and feed
##
##           casein  horsebean linseed meatmeal soybean
## horsebean 2.9e-08 -         -       -        -
## linseed   0.00016 0.09435   -       -        -
## meatmeal  0.18227 9.0e-05   0.09435 -        -
## soybean   0.00532 0.00298   0.51766 0.51766  -
## sunflower 0.81249 1.2e-08   8.1e-05 0.13218  0.00298
##
## P value adjustment method: holm
```

As you probably suspected based on the boxplot, we have some significant differences between groups (e.g. horse bean and casein) and some non-significant differences (e.g meat meal and linseed). The feeds with the two highest means, casein and sunflower, are not significantly different. We also see that the feed with the third highest mean, meat meal, is also not significantly different from either casein or sunflower. So, it looks like the farmer has some options in terms of the feed they select.

To make a final decision the farmer might now look to other factors such as cost or availability to decide what feed to provide her chickens. It is this final step that is important. The statistical test is just one part of the process of working out what to do. You then typically need to look at other information to workout what you actually need to do.

Reporting Test Results

How you present numerical results in a table is just as important as how you present figures. You must include information in a clear and concise way, without causing distraction with unnecessary information. Table 1 presents the results of our pair-wise comparisons. This is a good format for you to follow.

Table 16.1: Pair-wise t-test results (p-values):
effect of feed on chicken weights

	Casein	Horsebean	Linseed	Meatmeal	Soybean
Horsebean	< 0.001	-	-	-	-
Linseed	< 0.001	0.094	-	-	-
Meatmeal	0.182	< 0.001	0.094	-	-
Soybean	0.005	0.003	0.518	0.518	-
Sunflower	0.812	< 0.001	< 0.001	0.132	0.003

Note: Holm adjusted p-values

16.2 The notation of the ANOVA model

The ANOVA test can be motivated by saying that when looking at grouped data the observed variation can be separated into two parts: (i) the variation due to differences in the group means; and (ii) the variation of individual observations around each individual group mean. The variatition between group means is called the between group variation. The variation of individual observations around each individual group mean is called the within group variation.

The ANOVA test can be summarised with two propositions:

1. If the variation due to differences in the group means is relatively large, while the variation of observations around each group mean is relatively small, it is likely there are true differences in the group means. For this scenario the test will generate a small p-value.

2. If the variation due to differences in the group means is relatively small, while the variation of observations around the group mean is relatively large, it is unlikely there are true differences in the group means. For this scenario the test will generate a relatively large p-value.

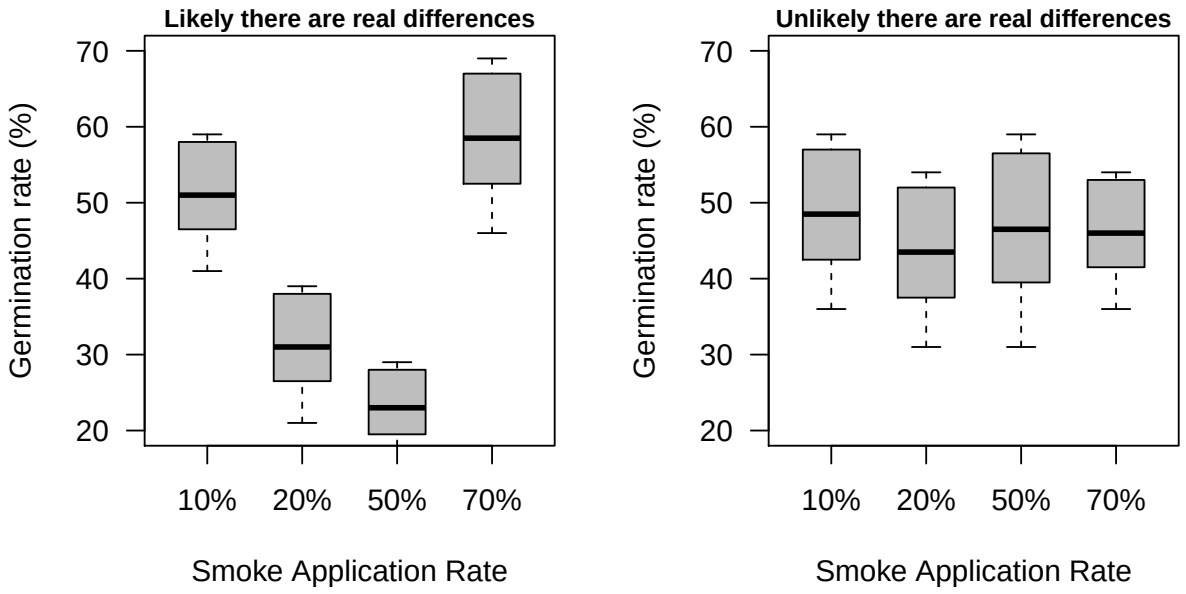


Figure 16.2: A comparison of where differences are likely and unlikely

If you understand the interpretation of the ANOVA test presented above you understand the fundamental element of ANOVA test. Unfortunately, the formal notation used to set out the ANOVA model is relatively intimidating. The complexity of the notation is one of the reasons a great many people refuse to come to terms with ANOVA. For the vast majority of people formal notation serves to obscure rather than highlight what is going on. The ANOVA model is especially bad in this respect. The type of notation used for ANOVA analysis will appear again and again throughout your studies. Because of this, a relatively formal introduction to ANOVA notation is presented below.

Introduction to notation

Consider the case where we have three groups. In group 1 we have 4 observations, in group 2 we have 5 observations, and in Group 3 we have 4 observations. That means we have something like the following:

For group 1 we have four observations: X_1, X_2, X_3, and X_4 which we summarise as X_j, $j = 1, ..., 4$. For group 2 we have five observations: X_1, X_2, X_3, X_4,and X_5, which we summarise as X_j, $j = 1, ..., 5$. For group 3 we have four observations: X_1, X_2, X_3, and X_4 which we summarise as X_j, $j = 1, ..., 4$. What you can see is that when we have groups of observations things get confusing if we use a single sub-script to index observations.

In the above expressions j was used to keep track of the number of observations in each group. To keep track of which group an individual observation belongs to we need to introduce a second subscript. This looks complex, but it we take things slowly it makes sense. When we have groups we use the notation X_{ij} to keep track of all the observations. The

i tells us which group the observation belongs to and the j keeps track of the individual observations within each group.

So for the specific case we have:

1. For group 1 we have: X_{11}, X_{12}, X_{13}, and X_{14}
2. For group 2 we have: X_{21}, X_{22}, X_{23}, X_{24}, and X_{25};
3. For group 3 we have: X_{31}, X_{32}, X_{33}, and X_{34}.

The number of observations in each group can be different. In our example there are 4 observations in group 1 and group 3, but 5 observations in group 2. The indexation for j (the number of observations within each group) must reflect this, and so we write $j = 1, ..., n_i$. This says that the number of observations in group i runs from 1 through to the maximum number for group i , and so reflects the fact that the number of observations in each group can vary. In our example we have $n_1 = 4$, $n_2 = 5$, and $n_3 = 4$.

We can denote the individual group means as $\overline{X}_1$, $\overline{X}_2$, and $\overline{X}_3$, and the average across all observations as $\overline{X}_G$, which we call the grand mean. We know how to calculate these values. Formally, we write the group mean for group i as:

$$\overline{X}_i = \frac{1}{n_i} \sum_j X_{ij}.$$

This expression says, to find the mean for any given group, take all the observations for that group, add them up, and divide through by the total number of observations for that group. The expression $\sum_j$ is read *sum over all posssible values of* j. So, if we take group 1 to illustrate we would have:

$$\begin{aligned} \overline{X}_1 &= \frac{1}{n_1} \sum_j X_{1j} \\ &= \frac{1}{4} (X_{11} + X_{12} + X_{13} + X_{14}) \\ &= \text{average for group 1.} \end{aligned}$$

To find the overall grand mean for the data we know that we simply have to add up all the observations and divide through by the total number of observations. While in practice this is a simple operation in Excel or R, the formal notation is a little involved. First we have to implement a rule that will give us the total number of observations, which we will denote with N . The formal notation we use to denote the total number of observations is

as follows:

$$\begin{aligned} N &= \sum_i n_i \\ &= n_1 + n_2 + n_3 \\ &= 4 + 5 + 4 \\ &= 13. \end{aligned}$$

Now, we need a way to say add up in a sequence all the observations in group 1, group 2, and group 3. If we had a single group we could use the expression $\sum_j X_j$, which say add up all the observations indexed by j. However, we need to say add up all the observations in group 1, group 2, and group 3. To see how this works let us start by writing out in full what we want to do. We want to write something that will say add up all 13 observations:

$$X_{11} + X_{12} + X_{13} + X_{14} + X_{21} + X_{22} + X_{23} + X_{24} + X_{25} + X_{31} + X_{32} + X_{33} + X_{44}$$

We know how to say add up all the values within a group. For group 1 we write this as $\sum_j X_{1j}$, for group 2 we write $\sum_j X_{2j}$, and for group 3 we write $\sum_j X_{3j}$. Now we want to add all these values together so we have something that looks like:

$$\sum_j X_{1j} + \sum_j X_{2j} + \sum_j X_{3j}.$$

What we need is a second summation sign that says add up across the groups. Formally we write $\sum_i \sum_j X_{ij}$ if we want to say add up all the observations from all groups. This expression says add up all the observations in all groups; where the groups are indexed by i and the individual observations within a group are indexed by j. For the three groups we have considered the term $\sum_i \sum_j X_{ij}$ translates as follows:

$$\sum_i \sum_j X_{ij} = X_{11} + X_{12} + X_{13} + X_{14} + X_{21} + X_{22} + X_{23} + X_{24} + X_{25} + X_{31} + X_{32} + X_{33} + X_{44}.$$

To find the average of the observations so we need to divide through by the total number of observation, which is $N = \sum_i n_i$. So, formally we denote the grand mean as:

$$\begin{aligned} \overline{X}_G &= \frac{1}{N} \sum_i \sum_j X_{ij} \\ &= \frac{1}{13} \left(X_{11} + X_{12} + X_{13} + X_{14} + X_{21} + X_{22} + X_{23} + X_{24} + X_{25} + X_{31} + X_{32} + X_{33} + X_{44} \right). \end{aligned}$$

In terms of notation this is as complicated as things get, and we now have all the elements we need to consider a formal ANOVA model.

The ANOVA model

The proposition presented regarding the ANOVA test is that if the variation due to differences in the group means is relatively large, while the variation of observations around each group mean is relatively small it is likely there are true differences in the group means. The alternative contrasting scenario is one where the variation due to differences in the group means is relatively small, while the variation of observations around each group mean is relatively large it is unlikely there are true differences in the group means.

The ANOVA test statistic is:

$$\frac{\text{standardisied measure of between group variation}}{\text{standardisied measure of within group variation}}.$$

This test statistic follows the F-distribution, and large values for F (p-value < 0.05) lead us to reject the null hypothesis of no differences between the groups. When adding up the differences between individual observations and the group mean, or differences between the group mean and the grand mean, positive and negative values will cancel out. To ensure that we have a positive value for the variation we therefore square each difference calculation.

Between group variation

We find the total between group variation for each group as the difference between the group mean and the grand mean multiplied by the number of observations in each group. This measure can be written as:

$$n_1\left(\overline{X}_1 - \overline{X}_G\right)^2 + n_2\left(\overline{X}_2 - \overline{X}_G\right)^2 + n_3\left(\overline{X}_3 - \overline{X}_G\right)^2.$$

In our example we have 4 observations in group 1, 5 observations in group 2, and 4 observations in group 3, so we have $n_1 = 4$, $n_2 = 5$, $n_3 = 4$ or

$$4 \times \left(\overline{X}_1 - \overline{X}_G\right)^2 + 5 \times \left(\overline{X}_2 - \overline{X}_G\right)^2 + 4 \times \left(\overline{X}_3 - \overline{X}_G\right)^2.$$

We can however also use the summation operator to write this in a formal way as:
$\sum_i n_i\left(\overline{X}_i - \overline{X}_G\right)^2 =$ the total between group variation,

and we call this measure the between group sum of squares. Note that the indexation subscript uses i as we are comparing group means to the grand mean. The standardisation we use is to divide through by the degrees of freedom. In a statistics context degrees of freedom works as follows. When we consider the sum of squares calculations there is a constraint that means if we have all but the last piece of information we are able to retrieve the final missing value. The final value is not free to take any value but rather is constrained to the value required by the formula we are using.

For our between group variation, if we know any $k-1$ values we can retrieve the final value. The standardisation we use for the between group sum of squares is therefore $k-1$.

So, to summarise we have:

$$\sum\nolimits_i n_i \left(\overline{X}_i - \overline{X}_G\right)^2 = \text{the between group sum of squares.}$$

We then have as our standardised measure of the between group variation:

$$\begin{aligned} \frac{\sum_i n_i \left(\overline{X}_i - \overline{X}_G\right)^2}{k-1} &= \text{the standardised between group sum of squares} \\ &= \text{the between group mean squares.} \end{aligned}$$

Within group variation

For the within group variation we have the same problem with positive and negative values that would cancel out, so again we square each deviation around the individual group means. The variation we want to add up is the variation of the individual observations around their respective group means. Specifically, the values that we want to add up are the following values:

$$\begin{aligned} \text{Sum of squared deviations} & \\ \text{from group means} &= \left(X_{11} - \overline{X}_1\right)^2 + \cdots + \left(X_{14} - \overline{X}_1\right)^2 + \\ &\quad \left(X_{21} - \overline{X}_2\right)^2 + \cdots + \left(X_{25} - \overline{X}_2\right)^2 + \\ &\quad \left(X_{31} - \overline{X}_3\right)^2 + \cdots + \left(X_{34} - \overline{X}_3\right)^2 \end{aligned}$$

We know that we can write the sum of the squared deviations from the mean for the groups individually as:

- within group variation for Group 1 $= \sum_j \left(X_{1j} - \overline{X}_1\right)^2$
- within group variation for Group 2 $= \sum_j \left(X_{2j} - \overline{X}_2\right)^2$
- within group variation for Group 3 $= \sum_j \left(X_{3j} - \overline{X}_3\right)^2$

This means that we can, in the first instance pull together the information that describes the within group variation as:

$$\begin{aligned} \sum\nolimits_i \sum\nolimits_j \left(X_{ij} - \overline{X}_i\right)^2 &= \left(X_{11} - \overline{X}_1\right)^2 + \cdots + \left(X_{14} - \overline{X}_1\right)^2 + \\ &\quad \left(X_{21} - \overline{X}_2\right)^2 + \cdots + \left(X_{25} - \overline{X}_2\right)^2 + \\ &\quad \left(X_{31} - \overline{X}_3\right)^2 + \cdots + \left(X_{34} - \overline{X}_3\right)^2 \end{aligned}$$

We call this the within group sum of squares. Within each group there is one value that is not free to vary, and so the degrees of freedom is equal to $N - k$; the total number of observations minus the number of groups. In the example we have been working with there are 19 observations and 3 groups so we have 16 degrees of freedom. We then have as our

standardised measure of the within group variation:

$$\frac{\sum_i \sum_j \left(X_{ij} - \overline{X}_i\right)^2}{N-k} = \text{the standardised within group sum of squares} = \text{within group mean squares.}$$

Again, technically this value is a variance estimate which follows a chi-squared distribution.

The ANOVA test is then conducted as:

$$F = \frac{\text{Between Group MS}}{\text{Within Group MS}},$$

and we reject the null hypothesis of no significant difference in the group means if this value is large, where large is defined by the F-distribution (We use the fact that F-value, as the ratio of two chi-squared variables, follows the F-distribution).

In practice we need to know how to implement the test, and we can implement the test in R using the `oneway.test()` function.

16.3 Advanced: Technical Details on ANOVA and F-test

The method of analysis of variance in this context focuses on answering one question: is the variability in the sample means so large that it seems unlikely to be from chance alone? This question is different from earlier testing procedures since we will *simultaneously* consider many groups, and evaluate whether their sample means differ more than we would expect from natural variation. We call this variability the **mean square between groups (MSG)**, and it has an associated degrees of freedom, $df_G = k-1$ when there are k groups. The MSG can be thought of as a scaled variance formula for means. If the null hypothesis is true, any variation in the sample means is due to chance and shouldn't be too large. Details of MSG calculations are provided in the footnote,[133] however, we typically use software for these computations.

The mean square between the groups is, on its own, quite useless in a hypothesis test. We need a benchmark value for how much variability should be expected among the sample means if the null hypothesis is true. To this end, we compute a pooled variance estimate, often abbreviated as the **mean square error (MSE)**, which has an associated degrees of freedom value $df_E = n - k$. It is helpful to think of MSE as a measure of the variability within the groups. Details of the computations of the MSE are provided in the footnote[134]

[133] Let $\bar{x}$ represent the mean of outcomes across all groups. Then the mean square between groups is computed as

$$MSG = \frac{1}{df_G} SSG = \frac{1}{k-1} \sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{x}\right)^2$$

where SSG is called the **sum of squares between groups** and n_i is the sample size of group i.

[134] Let $\bar{x}$ represent the mean of outcomes across all groups. Then the **sum of squares total (SST)** is

for interested readers.

When the null hypothesis is true, any differences among the sample means are only due to chance, and the MSG and MSE should be about equal. As a test statistic for ANOVA, we examine the fraction of MSG and MSE:

$$F = \frac{MSG}{MSE} \tag{1}$$

The MSG represents a measure of the between-group variability, and MSE measures the variability within each of the groups.

We can use the F statistic to evaluate the hypotheses in what is called an **F test**. 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 3 and 323 degrees of freedom, corresponding to the F statistic for the baseball hypothesis test, is shown in Figure 16.3.

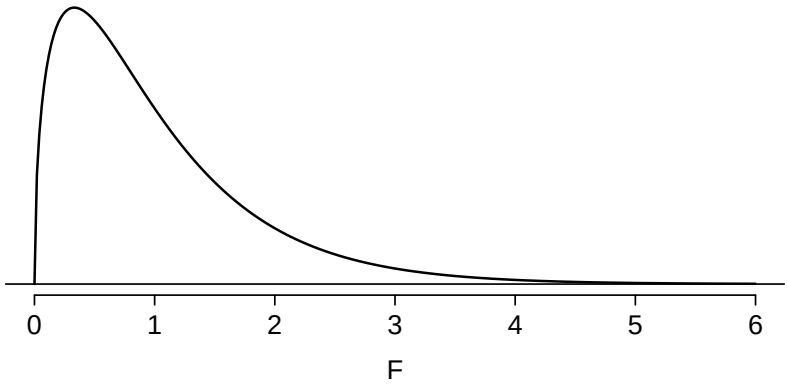


Figure 16.3: An F distribution with $df_1 = 3$ and $df_2 = 323$.

The larger the observed variability in the sample means (MSG) relative to the within-group observations (MSE), the larger F will be and the stronger the evidence against the null hypothesis. Because larger values of F represent stronger evidence against the null hypothesis, we use the upper tail of the distribution to compute a p-value.

---

computed as

$$SST = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

where the sum is over all observations in the data set. Then we compute the **sum of squared errors** (SSE) in one of two equivalent ways:

$$\begin{aligned} SSE &= SST - SSG \\ &= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2 \end{aligned}$$

where s_i^2 is the sample variance (square of the standard deviation) of the residuals in group i. Then the MSE is the standardized form of SSE: $MSE = \frac{1}{df_E} SSE$.

The F statistic and the F test

Analysis of variance (ANOVA) is used to test whether the mean outcome differs across 2 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 represent the p-value.

⊙ **Guided Practice 16.2** The test statistic for the baseball example is $F = 1.994$. Shade the area corresponding to the p-value in Figure 16.3. [135]

● **Example 16.3** The p-value corresponding to the shaded area in the solution of Guided Practice 2 is equal to about 0.115. Does this provide strong evidence against the null hypothesis?

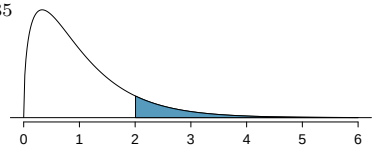
The p-value is larger than 0.05, indicating the evidence is not strong enough to reject the null hypothesis at a significance level of 0.05. That is, the data do not provide strong evidence that the average on-base percentage varies by player's primary field position.

Multiple comparisons and controlling Type 1 Error rate

When we reject the null hypothesis in an ANOVA analysis, we might wonder, which of these groups have different means? To answer this question, we compare the means of each possible pair of groups. For instance, if there are three groups and there is strong evidence that there are some differences in the group means, there are three comparisons to make: group 1 to group 2, group 1 to group 3, and group 2 to group 3.

When making these comparisons it is necessary to implement an adjustment to control the type 1 error rate. We use a default adjustment implemented by R.

135



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.

17 Regression

Our focus is on fitting a trend line to observations. There are many different decision rules that could be used to decide how to fit the 'best' trend line to a data set. The decision rule we will use to fit a trend line is to minimise the squared vertical distances between the trend line and the actual data points. We call the difference between where we fit the trend line and the actual data points the error term. In practical terms our decision rule means when we fit a trend line we try different intercept and different slope values until we get a line that fits the data as best as possible, where as best as possible means we have 'minimised the sum of the squared error terms.'

The reason we minimise the squared error terms rather than just minimise the sum of the error terms is that sometimes the errors are positive and sometimes they are negative, so we need a decision rule that takes this into account. The two easiest solutions to the problem are to either take the absolute value of the error term or square the error term values. With either decision rule the error terms are all positive. For reasons of computational ease, the convention is to use squared error terms rather than the absolute value. This approach gives rise to the estimation method name: least squares estimation. If you fit a trend line using the sum of the absolute values of the error terms the approach is called least absolute deviation, but this second approach is but much less common than the method of least squares.

The general term used for the modelling approach of fitting a trend line to data is linear regression. Linear regression does not mean that we are always going to fit a straight line to the data, but fitting a straight line to the data will be our initial focus. Once the basic approach is mastered it is easy to extend the modelling approach to more complicated examples.

In the language of linear regression we say that the decision rule we use to fit the trend line is an 'estimator'. So, we use the least squares estimator (decision rule) to obtain estimates of the slope and the intercept of the trend line. In linear regression the variable on the y-axis (vertical axis) is called the dependent variable or the response variable. The variable on the x-axis (horizontal axis) is called the explanatory variable or independent variable. Generally we think of the variable on the x-axis as being able to 'explain' the variable on the y-axis.

17.1 Estimating a Trend Line

In this example our data are the carapace sizes of lobsters at different distances to a no-take marine sanctuary. The research question is whether there is a relationship between lobster size and distance to the sanctuary. Specifically, we are interested in whether lobsters are smaller the further they are found from the no-take area.

We are not comparing lobster sizes at two difference distances, but rather we have data collected at a whole range of distances.

Let's read in our data and have a look at it. Note that with grouped data we generally look at a boxplot, and with ungrouped data (a single numerical variable) we create a histogram or a horizontal boxplot. Here we are looking at the 'relationship' rather than difference between two continuous (numerical) variables, so will create a scatter plot.

Note: There is a separate reference guide for scatter plots, so the descriptions given here are relatively brief.

Our `y` or dependent/response variable is lobster size, our `x` or independent/explanatory variable is distance to the no-take area (sanctuary). When plotting and creating linear regression models we give **R** the formula: `y ~ x`. Deciding which variable goes on the y-axis and which variable goes on the x-axis is tricky. The convention is to use the variable that we think is doing the explaining on the horizontal (x-axis). Because we think that distance from the no-take zone might 'explain' carapace width, we place distance on the horizontal axis and carapace width on the vertical.

```
# read in our data
lobster.data<- data.frame(
  distance= c(0, 2, 3, 4, 4, 5, 6, 7, 9, 12, 18, 20, 20, 23, 24, 26, 27, 28, 30),
   size= c(116.89, 96.90, 120.97, 116.40, 89.86, 96.83, 116.18, 117.02, 79.03,
          69.86, 105.60, 68.12, 78.12, 69.86, 66.64, 62.11, 59.11, 59.94, 44.43))

# have a look at the str, summary to mkae sure we understand the data
str(lobster.data) # we have 2 numerical or continuous variables

## 'data.frame': 19 obs. of  2 variables:
##  $ distance: num  0 2 3 4 4 5 6 7 9 12 ...
##  $ size    : num  116.9 96.9 121 116.4 89.9 ...

summary(lobster.data) # 6 number summary of both variables

##     distance          size
##  Min.   : 0.00   Min.   : 44.43
##  1st Qu.: 4.50   1st Qu.: 67.38
##  Median :12.00   Median : 79.03
##  Mean   :14.11   Mean   : 85.99
##  3rd Qu.:23.50   3rd Qu.:110.89
##  Max.   :30.00   Max.   :120.97
```

Now, rather than a boxplot we create a scatter plot. Note the form of the scatter plot formula is `y ~ x` so we have `size ~ distance` in the plot formula. Also note that in your own plots you will just have the figure caption, not a figure caption and a figure title.

```
with(lobster.data, plot(size~distance,
     pch= 4, # controls the dot format used
```

```
    # write title on two lines ( use \n or return for a new line)
    main= "Lobster size in relation to distance to a
           no-take marine sanctuary",
    ylab= "Lobster carapace size (mm)",
    xlab= "Distance from santuary (m)",
    ylim= c(40,130), xlim= c(0,30), las= 1))
```

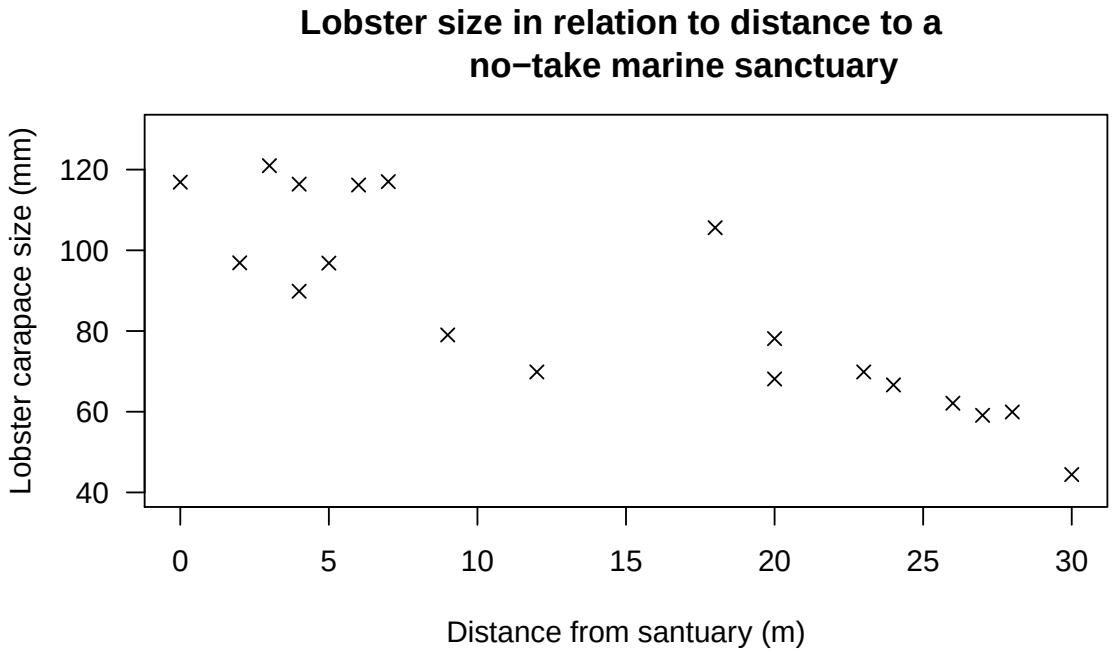


Figure 17.1: Scatter plot showing the relationship between distance to a no-take marine sanctuary and lobster carapace sizes.

What do you see here? Does there appear to be a linear relationship between the two variables, i.e. could you draw a straight line through the majority of the points? At the greatest distances from the sanctuary (25 - 30 m) we do see the smallest lobsters, and at the shortest distances (0 - 5 m) we have the largest lobsters, but in the middle values appear to be somewhat more random. The relationship is not perfect, but fitting a trend line to the data might help us understand what is going on.

Let's add a trend line to the data by estimating a 'linear' model with the function *lm()*, where lm stands for linear model. We will also save the model output as an object, giving it an intuitive name (`lm.lobster`) and calling on the data for the model using the parameter `data`.

Using the parameter `data` to indicate the data set we want **R** to use is a slight change in notation from the appraoch we have used for other models. For the ANOVA model, t-tests,

Kruskal-Wallis test, and Wilcoxon test we used the `with` command to tell **R** which data set to use.

To get the values for the trend line – the coefficients of the equation of the line - we use the function *coef()* calling on the model we created.

```
lm.lobster <- lm(size ~ distance, data = lobster.data)  # create our model
coef(lm.lobster)  # print our coefficients

## (Intercept)    distance
##  114.388201   -2.013081
```

Here we are given the values for our equation of the line: the intercept is 114.39 and slope for our explanatory variable (distance) is -2.01. Our slope is negative indicating we have a downward sloping trend. The reported intercept and slope values are the values that give a trend line that minimises the squared error terms between the actual data and the trend line.

As an equation this looks like: `fitted y-value = intercept + slope(x-data)`
Which we can write as: `y-hat = (114.388) + (-2.013)x`
Or in a clean general form: $\hat{y} = 114.4 - 2.0x$
Or, in a form that makes clear what is being measured: $\widehat{size} = 114.4 - 2.0 \times Distance$.

The equation tells us that for every 1 m increase in distance from the no-take marine sanctuary, lobster carapace sizes decrease, on average, by about 2 mm. When we have a regression equation we always add the comment, on average. The intercept says that at 0 m away from the sanctuary lobster carapaces are estimated to be, on average, 114 mm. Because the data used to derive the model includes 0 m from the sanctuary, this can be considered a value that has some meaning. In most cases the intercept coefficient is not informative and so we will generally not look at the intercept term in our models.

Note: You have MUCH less confidence using a model to extrapolate beyond the range of your data. Formally we only know what is happening for the data range we observe. So, this model may be useful for predicting lobster size estimates from `0` m to `121` m from this sanctuary. Outside this range the model has little to say, and it would be dangerous to extrapolate.

The last step is to graphically display our model. We can do this by adding a trend line to our scatter plot with the *abline()* function and identifying our line in a legend with the *legend()* function. Note we can add the line either by specifying the numerical values directly (`a` = intercept and `b` = slope), or by telling R to get the slope and intercept from our model (lm.lobster).

```
with(lobster.data, plot(size~distance,
     pch= 4,
     # write title on two lines ( use \n or return for a new line)
     main= "Lobster size in relation to distance to a
```

```
     no-take marine sanctuary",
     ylab= "Lobster carapace size (mm)",
     xlab= "Distance from santuary (m)",
     ylim= c(40,130), xlim= c(0,30), las= 1))

# add our trend line
abline(lm.lobster, lty= 2, lwd= 2, col= "blue")

# add our legend
legend("topright", legend= "trend line", lty= 2, lwd= 2, col= "blue", bty= "n")
```

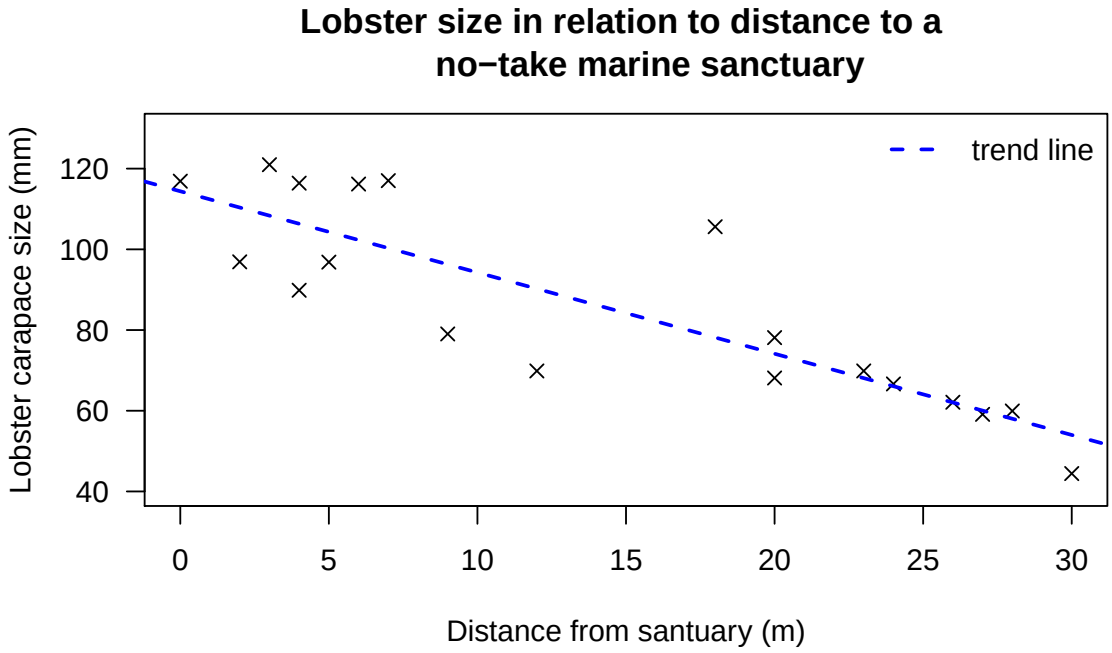


Figure 17.2: Scatter plot showing the relationship between distance to a no-take marine sanctuary and lobster carapace sizes, including linear model trend line

17.2 Statistical Significance of the Slope Estimate

The decision rule we use to generate an estimate of the slope and intercept (the least squares estimator) will always generate estimates of the slope and intercept. However, we are interested in knowing whether or not the slope estimate we obtain is statistically different from zero. The question we are asking is could the pattern we observe in the data be due to chance, or is it a real trend?

To determine whether or not the slope estimate is statistically different from zero we conduct a t-test. The t-test for the slope (and intercept) works the same way as our earlier t-tests, and we will use the same decision rule. Specifically, if the t-test p-value is less than 0.05 we will reject the null hypothesis that the slope estimate is equal to zero. In practice this decision rule means that we have set the test alpha level at 0.05.

To conduct the t-test we use the *summary()* function. The *summary()* function will conduct a t-test on both the slope and the intercept, and will also report some additional information on our linear model. This automated routine is one of the very useful things about **R**. While MS Excel will fit a trend line, the default output does not tell you whether or not the estimates are statistically different from zero. Let's run a formal test on the model we created previously, and saved as: `lm.lobster`.

Step 1: Set the Null and Alternate Hypotheses
Here we will be testing only the slope as we are not interested in the intercept.

- Null hypothesis: The slope is equal to zero
- Alternate hypothesis: The slope is not equal to zero

Step 2: Print the test output
We obtain all the output with a simple command *summary*(`lm.lobster`). The output we get has many elements but we will work through the detail in stages.

```
summary(lm.lobster)  # summary for our linear model

##
## Call:
## lm(formula = size ~ distance, data = lobster.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.3712  -8.5293   0.5658   7.0288  27.4473
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 114.3882     5.1334  22.283 5.08e-14 ***
## distance     -2.0131     0.2964  -6.791 3.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.98 on 17 degrees of freedom
## Multiple R-squared:  0.7307,Adjusted R-squared:  0.7148
## F-statistic: 46.12 on 1 and 17 DF,  p-value: 3.147e-06
```

Step 3: Interpret the Results

The output here is quiet a bit more complex than for basic t-tests. Let's focus, for now, on the 'Coefficients' section of the output. This section of the output provides the estimates for our trend line and the results for the test of statistical significance. Under the 'Estimate' column we see the same values as we obtained when using the *coef* function - the intercept and slope terms (distance is the slope estimate).

The next column is the 'Std. Error' column and the values in this column are the standard error for each coefficient. The standard error in a regression model has the same interpretation as the standard error of the mean in our earlier t-test examples. These values can be thought of as a measure of uncertainty for our slope and intercept values.

Next, we can see t-values. The t-values have been calculated using the approach we used for earlier a t-tests. Specifically, these values have been calculated as: Estimate value minus the null hypothesis test value divided by the standard error. In this instance our null hypothesis test value is zero; so the t-value is calculated as: (-2.0131 - 0 / 0.2964) = -6.79. Lastly we can see the p-values. The p-values are what we use to make a decision about whether or not the slope estimate is statistically different from zero. Since the p-value for the slope is: `3.15e-06` (< 0.001), which is smaller than 0.05, we:

- **Reject** the null hypothesis that the slope is equal to zero.

Why do we care if the slope estimate is different to zero? Well, if the slope is zero this means that there is no real trend in the data. The variation we have observed is just due to sampling variation. Fundamentally, we are interested in knowing whether the trend line we fit to the data represents a real trend line or not.

Another way to think about the null hypothesis for our test is that is asks the question: does a trend line fit the data better than a horizontal line at the mean of the data? A visual representation of the situation is given below. If the null were true, a horizontal line at the mean lobster carapace size would be a good description of the data. If the null is not true we are saying that a trend line fits the data better than a horizontal line. In our case, as we rejected the null, we are saying that the trend line is a better fit to the data than the horizontal line.

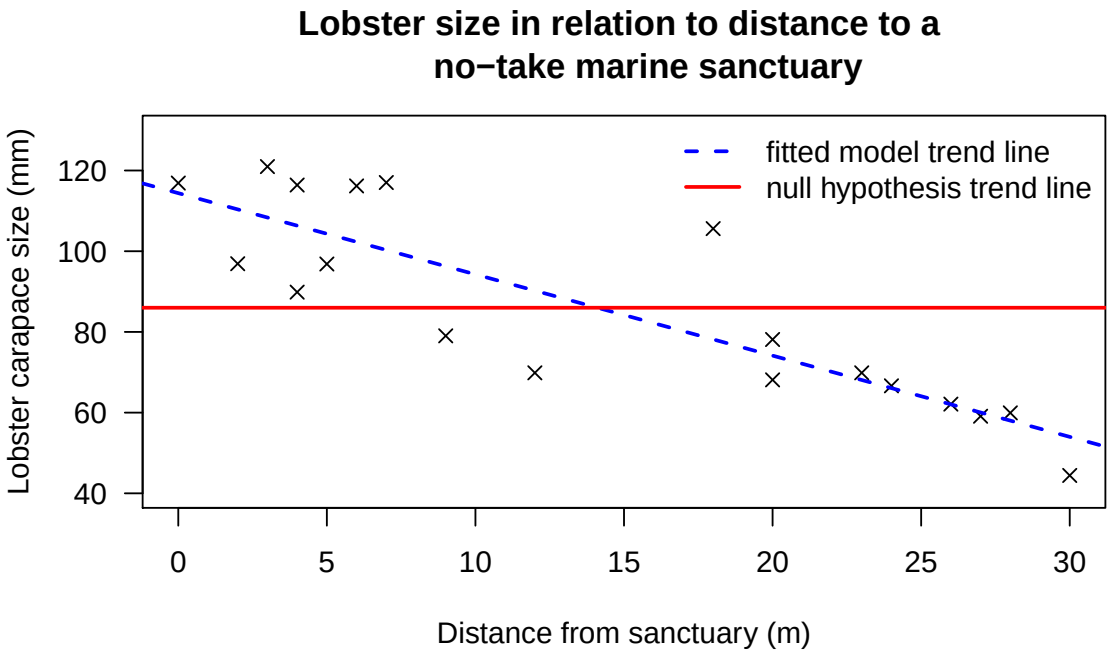


Figure 17.3: Scatter plot showing the relationship between distance to a no-take marine sanctuary and lobster carapace sizes.

17.3 Measure of Model Fit

Let's again look at the model output. The final thing we are interested in is the R^2 value. This value has an interpretation as the proportion of the variation in the data explained by the model. In **R**, the R^2 value is reported as the `Multiple R-squared value`, which is the second last line of the summary output. See if you can find it in the output below.

```
summary(lm.lobster)  # provide model created previously

##
## Call:
## lm(formula = size ~ distance, data = lobster.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.3712  -8.5293   0.5658   7.0288  27.4473
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 114.3882     5.1334  22.283 5.08e-14 ***
```

```
## distance     -2.0131     0.2964  -6.791 3.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.98 on 17 degrees of freedom
## Multiple R-squared:  0.7307,Adjusted R-squared:  0.7148
## F-statistic: 46.12 on 1 and 17 DF,  p-value: 3.147e-06

print(round(R_squared, 4))  # this is just to show you the value

## [1] 0.7307

# above is the R-squared value
```

As the R^2 value `= 0.731`, we say that the model explains 73.1% of the variation in the data. We don't really have a view regarding whether this is a good thing or a bad thing, but it is a commonly reported metric, and something that we will report.

There are other metrics of model fit, for example **R** also reports an `Adjusted R-squared value`. These other metrics do have advantages, but these alternative metrics do not have the same nice interpretation of percent/proportion of the variation in the data explained by the model. For this reason we will stick with reporting the R^2 value.

17.4 The Summary Table

Because there are multiple t-tests (one for the intercept and one for the slope), and because there is additional information to report, such as the R^2 value, the convention for reporting linear regression results in to use a summary table. The summary table contains: (i) the estimates of the slope and the intercept; (ii) the standard error of the slope and the intercept, where a '*' sign is used to denote statistical significance; (iii) the number of observations in the data set; and (iv) the R^2 value. The difference between the estimate and the standard error is usually indicated with parentheses. Because it is possible to report either standard error information or t-value information, the convention is to add a footnote to the table to say "Standard errors in parentheses." The thresholds chosen for statistical significance also vary.

Because we use p-value < 0.05 as the critical decision threshold for our t-tests, when we move to a more general system of critical thresholds we use alpha values around 0.05 that are both more and less stringent i.e. the threshold values we use for the '*' sign system of denoting statistical significance are '*' for p-value < 0.1; '**' for p-value < 0.05; '***' for p-value < 0.01.

Below is an example of an acceptable format for a regression summary table, but there are many other acceptable formats.

Table 17.1: Lobster linear regression

	Lobster size
Intercept	114.39*** (5.13)
Distance	−2.01*** (0.30)
Observations	19
R^2	0.73
Note:	*p<0.1; **p<0.05; ***p<0.01
	Standard errors in parentheses.

17.5 Advanced: Technical Details on Regression

Linear regression is a very powerful statistical technique. Many people have some familiarity with regression just from reading the news, where graphs with straight lines are overlaid on scatterplots. Linear models can be used for prediction or to evaluate whether there is a linear relationship between two numerical variables.

Figure 17.4 shows two variables whose relationship can be modelled perfectly with a straight line. The equation for the line is

$$y = 5 + 57.49x$$

Imagine what a perfect linear relationship would mean: you would know the exact value of y just by knowing the value of x. This is unrealistic in almost any natural process. For example, if we took family income x, this value would provide some useful information about how much financial support y a college may offer a prospective student. However, there would still be variability in financial support, even when comparing students whose families have similar financial backgrounds.

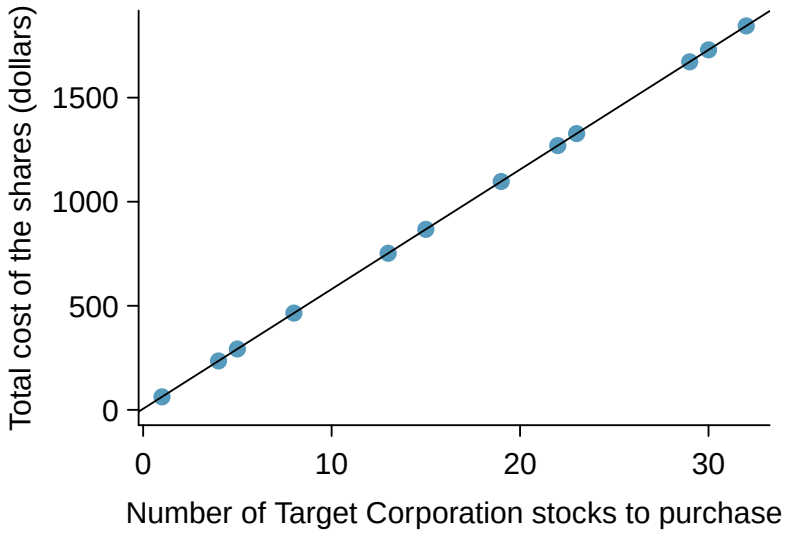


Figure 17.4: Requests from twelve separate buyers were simultaneously placed with a trading company to purchase Target Corporation stock (ticker TGT, April 26th, 2012), and the total cost of the shares were reported. Because the cost is computed using a linear formula, the linear fit is perfect.

Linear regression assumes that the relationship between two variables, x and y, can be modelled by a straight line:

$$y = \beta_0 + \beta_1 x \tag{1}$$

β_0, β_1 Linear model parameters

where β_0 and β_1 represent two model parameters (β is the Greek letter *beta*). These parameters are estimated using data, and we write their point estimates as b_0 and b_1. When we use x to predict y, we usually call x the explanatory or **predictor** variable, and we call y the response.

It is rare for all of the data to fall on a straight line, as seen in the three scatterplots in Figure 17.5. In each case, the data fall around a straight line, even if none of the observations fall exactly on the line. The first plot shows a relatively strong downward linear trend, where the remaining variability in the data around the line is minor relative to the strength of the relationship between x and y. The second plot shows an upward trend that, while evident, is not as strong as the first. The last plot shows a very weak downward trend in the data, so slight we can hardly notice it. In each of these examples, we will have some uncertainty regarding our estimates of the model parameters, β_0 and β_1. For instance, we might wonder, should we move the line up or down a little, or should we tilt it more or less? As we move forward in this chapter, we will learn different criteria for line-fitting, and we will also learn about the uncertainty associated with estimates of model parameters.

We will also see examples in this chapter where fitting a straight line to the data, even if there is a clear relationship between the variables, is not helpful. One such case is shown

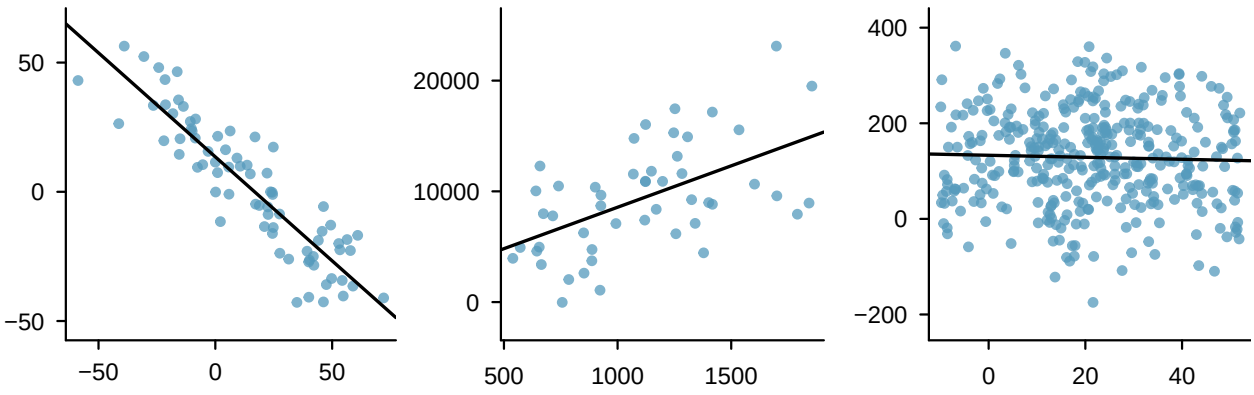


Figure 17.5: Three data sets where a linear model may be useful even though the data do not all fall exactly on the line.

in Figure 17.6 where there is a very strong relationship between the variables even though the trend is not linear. We will discuss some aspects of nonlinear trends in Chpater 8 but the details of fitting nonlinear models are saved for a later course.

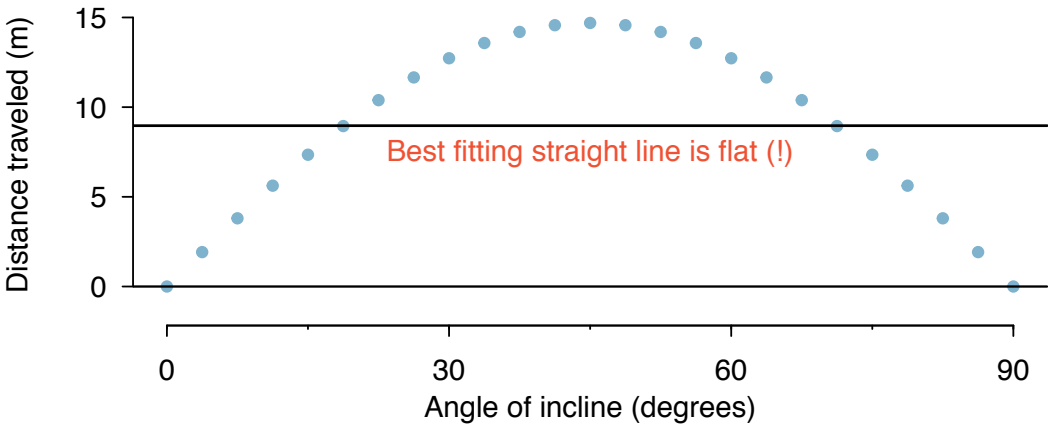


Figure 17.6: A linear model is not useful in this nonlinear case. These data are from an introductory physics experiment.

Line fitting, residuals, and correlation

It is helpful to think deeply about the line fitting process. In this section, we examine criteria for identifying a linear model and introduce a new statistic, *correlation*.

Beginning with straight lines

Scatterplots were introduced in Chapter 5 as a graphical technique to present two numerical variables simultaneously. Such plots permit the relationship between the variables to be examined with ease. Figure 17.7 shows a scatterplot for the head length and total length of 104 brushtail possums from Australia. Each point represents a single possum from the data.

The head and total length variables are associated. Possums with an above average total length also tend to have above average head lengths. While the relationship is not perfectly linear, it could be helpful to partially explain the connection between these variables with a straight line.

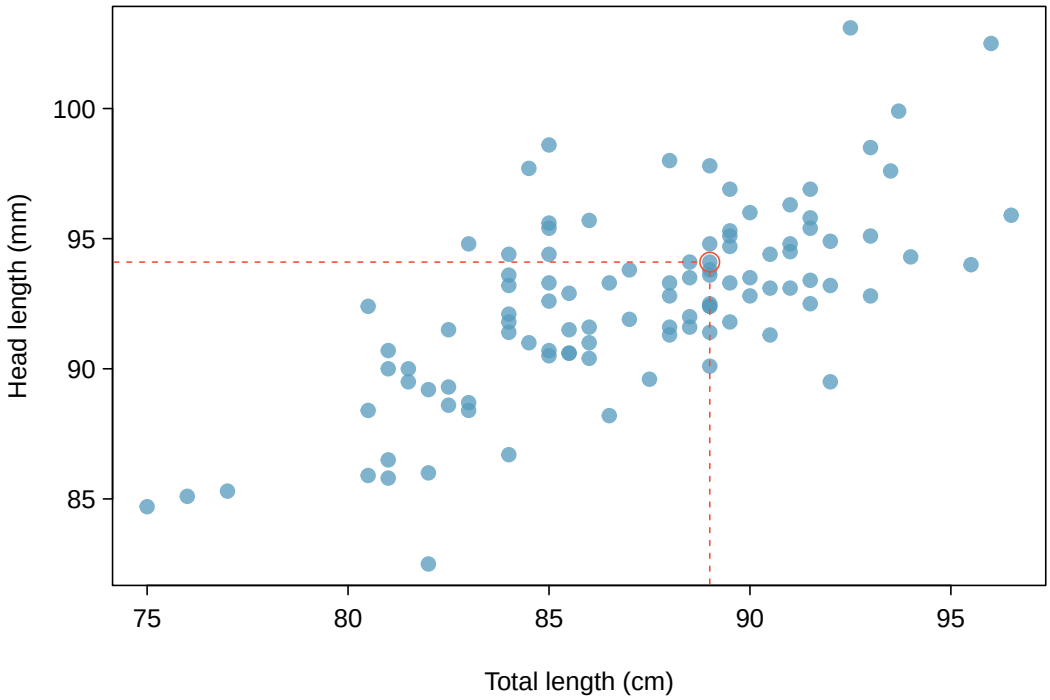


Figure 17.7: A scatterplot showing head length against total length for 104 brushtail possums. A point representing a possum with head length 94.1mm and total length 89cm is highlighted.

Figure 17.8: The common brushtail possum of Australia.Photo by Greg Schechter (https://flic.kr/p/9BAFbR).CC BY 2.0 license.

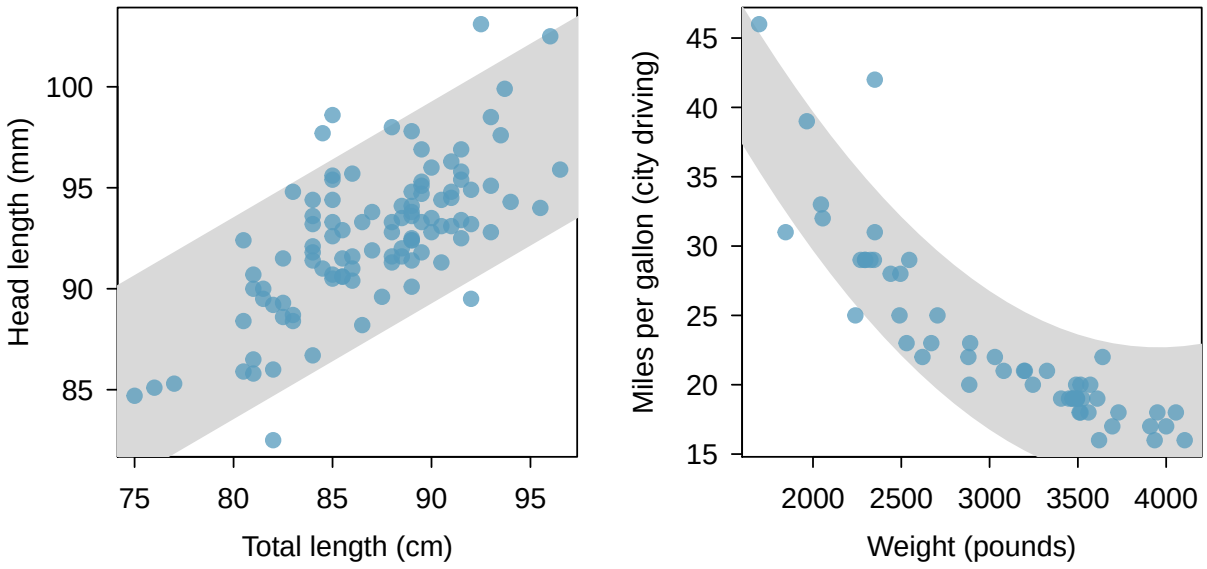


Figure 17.9: The figure on the left shows head length versus total length, and reveals that many of the points could be captured by a straight band. On the right, we see that a curved band is more appropriate in the scatterplot for `weight` and `mpgCity` from the `cars` data set.

Straight lines should only be used when the data appear to have a linear relationship, such as the case shown in the left panel of Figure 17.9. The right panel of Figure 17.9 shows a case where a curved line would be more useful in understanding the relationship between the two variables.

Caution: Watch out for curved trends

We only consider models based on straight lines in this chapter. If data show a nonlinear trend, like that in the right panel of Figure 17.9, more advanced techniques should be used.

Fitting a line by eye

We want to describe the relationship between the head length and total length variables in the possum data set using a line. In this example, we will use the total length as the predictor variable, x, to predict a possum's head length, y. We could fit the linear relationship by eye, as in Figure 17.10. The equation for this line is

$$\hat{y} = 41 + 0.59x \tag{2}$$

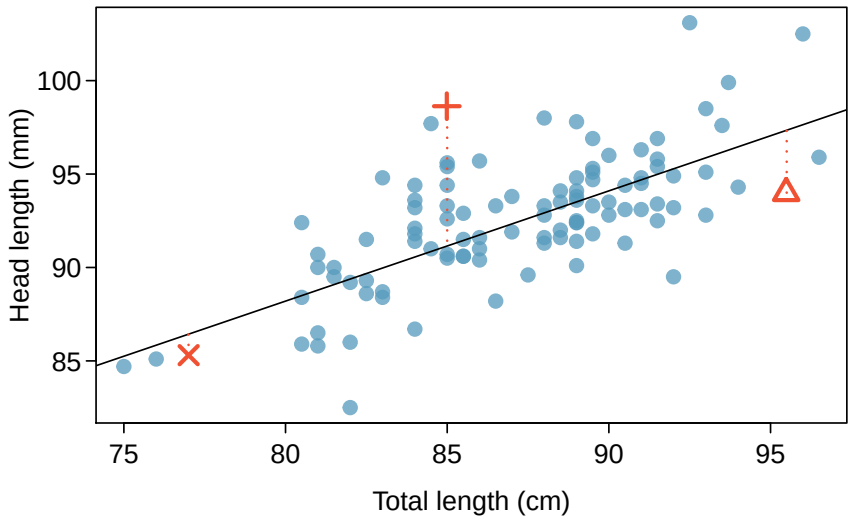


Figure 17.10: A reasonable linear model was fit to represent the relationship between head length and total length.

We can use this line to discuss properties of possums. For instance, the equation predicts a possum with a total length of 80 cm will have a head length of

$$\begin{aligned}\hat{y} &= 41 + 0.59 \times 80 \\ &= 88.2\end{aligned}$$

A "hat" on y is used to signify that this is an estimate. This estimate may be viewed as an average: the equation predicts that possums with a total length of 80 cm will have an average head length of 88.2 mm. Absent further information about an 80 cm possum, the prediction for head length that uses the average is a reasonable estimate.

Residuals

Residuals are the leftover variation in the data after accounting for the model fit:

$$\text{Data} = \text{Fit} + \text{Residual}$$

Each observation will have a residual. If an observation is above the regression line, then its residual, the vertical distance from the observation to the line, is positive. Observations below the line have negative residuals. One goal in picking the right linear model is for these residuals to be as small as possible.

Three observations are noted specially in Figure 17.10. The observation marked by an "×" has a small, negative residual of about -1; the observation marked by "+" has a large residual of about +7; and the observation marked by "△" has a moderate residual of about -4. The size of a residual is usually discussed in terms of its absolute value. For example, the residual for "△" is larger than that of "×" because $|-4|$ is larger than $|-1|$.

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 we would predict based on the model fit $(\hat{y}_i)$:

$$e_i = y_i - \hat{y}_i$$

We typically identify $\hat{y}_i$ by plugging x_i into the model.

Example 17.3 The linear fit shown in Figure 17.10 is given as $\hat{y} = 41 + 0.59x$. Based on this line, formally compute the residual of the observation $(77.0, 85.3)$. This observation is denoted by "×" on the plot. Check it against the earlier visual estimate, -1.

We first compute the predicted value of point "×" based on the model:

$$\hat{y}_\times = 41 + 0.59x_\times = 41 + 0.59 \times 77.0 = 86.4$$

Next we compute the difference of the actual head length and the predicted head length:

$$e_\times = y_\times - \hat{y}_\times = 85.3 - 86.4 = -1.1$$

This is very close to the visual estimate of -1.

Guided Practice 17.4 If a model underestimates an observation, will the residual be positive or negative? What about if it overestimates the observation?[136]

Guided Practice 17.5 Compute the residuals for the observations $(85.0, 98.6)$ ("+" in the figure) and $(95.5, 94.0)$ ("△") using the linear relationship $\hat{y} = 41 + 0.59x$. [137]

Residuals are helpful in evaluating how well a linear model fits a data set. We often display them in a **residual plot** such as the one shown in Figure 17.11 for the regression line in Figure 17.10. The residuals are plotted at their original horizontal locations but with the vertical coordinate as the residual. For instance, the point $(85.0, 98.6)_+$ had a residual

[136] If a model underestimates an observation, then the model estimate is below the actual. The residual, which is the actual observation value minus the model estimate, must then be positive. The opposite is true when the model overestimates the observation: the residual is negative.

[137] (+) First compute the predicted value based on the model:

$$\hat{y}_+ = 41 + 0.59x_+ = 41 + 0.59 \times 85.0 = 91.15$$

Then the residual is given by

$$e_+ = y_+ - \hat{y}_+ = 98.6 - 91.15 = 7.45$$

This was close to the earlier estimate of 7.

(△) $\hat{y}_\triangle = 41 + 0.59x_\triangle = 97.3$. $e_\triangle = y_\triangle - \hat{y}_\triangle = -3.3$, close to the estimate of -4.

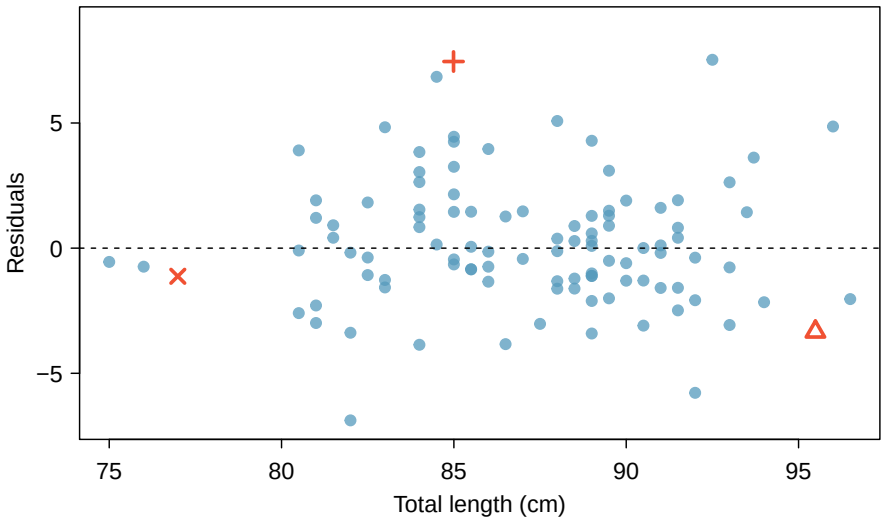


Figure 17.11: Residual plot for the model in Figure 17.10.

of 7.45, so in the residual plot it is placed at $(85.0, 7.45)$. Creating a residual plot is sort of like tipping the scatterplot over so the regression line is horizontal.

Example 17.6 One purpose of residual plots is to identify characteristics or patterns still apparent in data after fitting a model. Figure 17.12 shows three scatterplots with linear models in the first row and residual plots in the second row. Can you identify any patterns remaining in the residuals?

In the first data set (first column), the residuals show no obvious patterns. The residuals appear to be scattered randomly around the dashed line that represents 0.

The second data set shows a pattern in the residuals. There is some curvature in the scatterplot, which is more obvious in the residual plot. We should not use a straight line to model these data. Instead, a more advanced technique should be used.

The last plot shows very little upwards trend, and the residuals also show no obvious patterns. It is reasonable to try to fit a linear model to the data. However, it is unclear whether there is statistically significant evidence that the slope parameter is different from zero. The point estimate of the slope parameter, labelled b_1, is not zero, but we might wonder if this could just be due to chance. We will address this sort of scenario in Section 17.5.

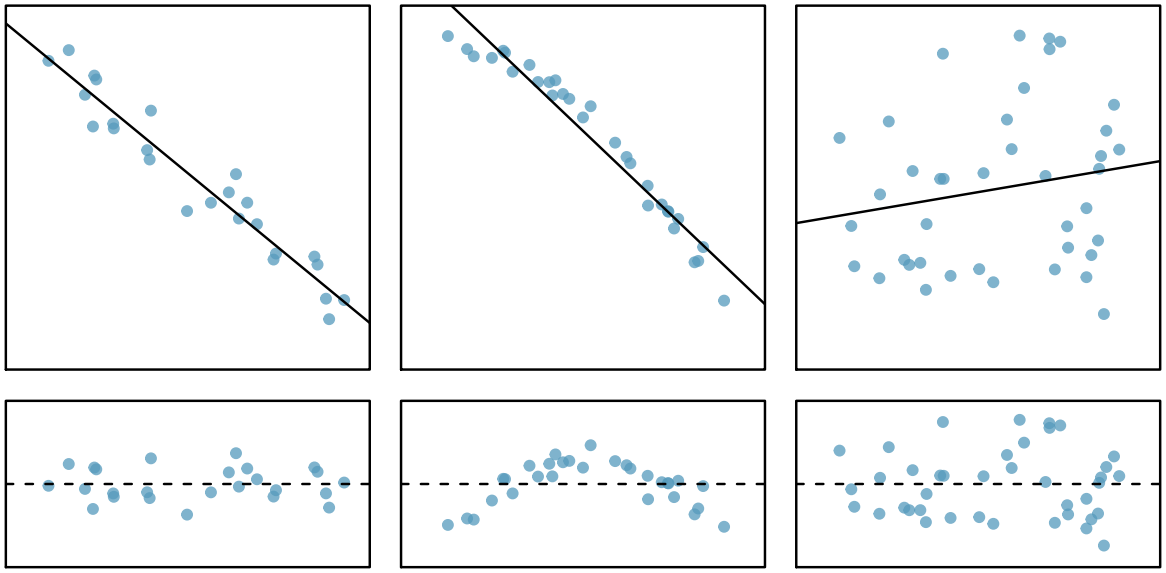


Figure 17.12: Sample data with their best fitting lines (top row) and their corresponding residual plots (bottom row).

Describing linear relationships with correlation

Correlation: strength of a linear relationship

Correlation, which always takes values between -1 and 1, describes the strength of the linear relationship between two variables. We denote the correlation by R.

We can compute the correlation using a formula, just as we did with the sample mean and standard deviation. However, this formula is rather complex,[138] so we generally perform the calculations on a computer or calculator. Figure 17.13 shows eight plots and their corresponding correlations. Only when the relationship is perfectly linear is the correlation either -1 or 1. If the relationship is strong and positive, the correlation will be near +1. If it is strong and negative, it will be near -1. If there is no apparent linear relationship between the variables, then the correlation will be near zero.

The correlation is intended to quantify the strength of a linear trend. Nonlinear trends, even when strong, sometimes produce correlations that do not reflect the strength of the relationship; see three such examples in Figure 17.14.

[138] Formally, we can compute the correlation for observations (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) using the formula

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$

where $\bar{x}$, $\bar{y}$, s_x, and s_y are the sample means and standard deviations for each variable.

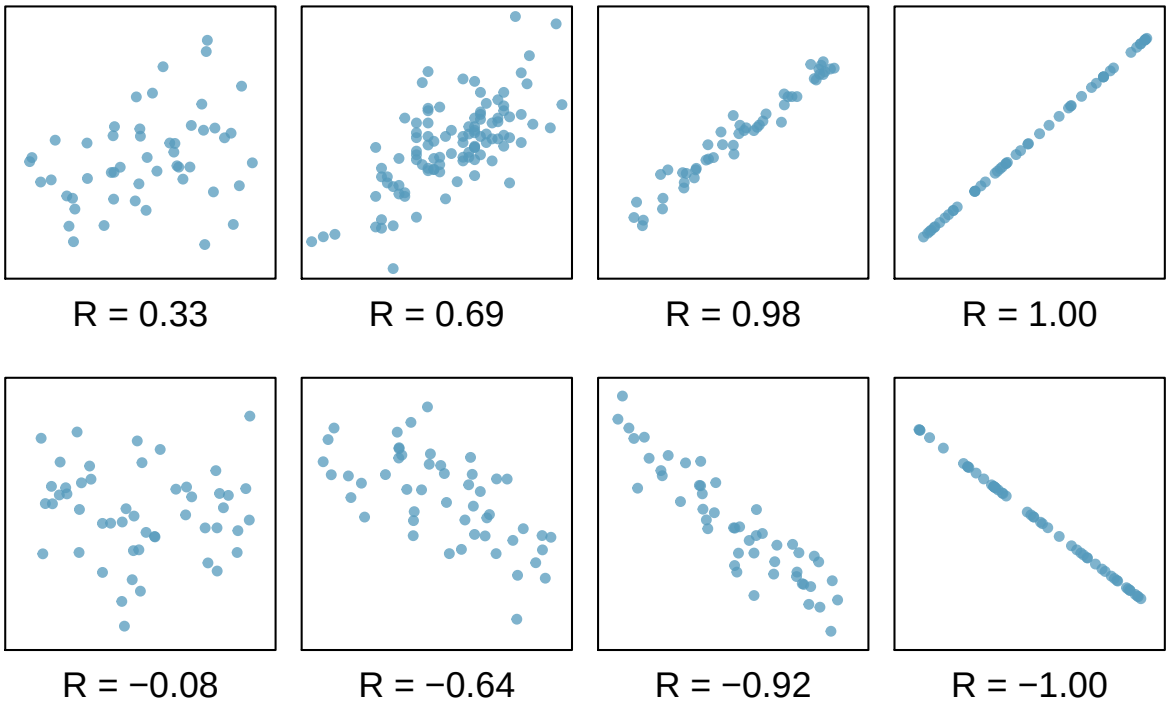


Figure 17.13: Sample scatterplots and their correlations. The first row shows variables with a positive relationship, represented by the trend up and to the right. The second row shows variables with a negative trend, where a large value in one variable is associated with a low value in the other.

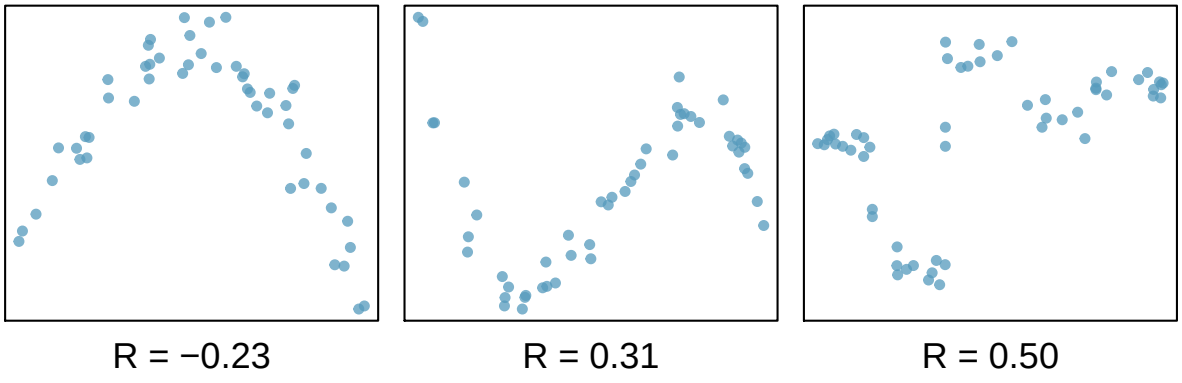


Figure 17.14: Sample scatterplots and their correlations. In each case, there is a strong relationship between the variables. However, the correlation is not very strong, and the relationship is not linear.

⊙ **Guided Practice 17.7** It appears no straight line would fit any of the datasets represented in Figure 17.14. Try drawing nonlinear curves on each plot. Once you create a curve for each, describe what is important in your fit.[139]

Fitting a line by least squares regression

Fitting linear models by eye is open to criticism since it is based on an individual preference. In this section, we use *least squares regression* as a more rigorous approach.

This section considers Western Rock Lobster data from section 17.1. A scatterplot of the data is shown in Figure 17.15 along with two linear fits. The lines follow a negative trend in the data; lobsters caught further from the sactuary tend to be larger in size.

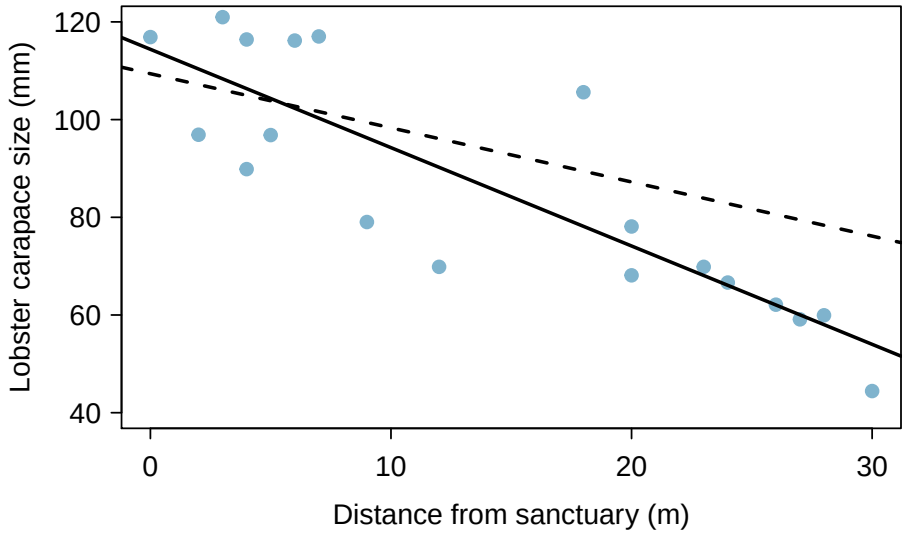


Figure 17.15: Carapace size and distance from sanctuary for a sample of 19 lobsters. Two lines are fit to the data, the solid line being the *least squares line*.

⊙ **Guided Practice 17.8** Is the correlation positive or negative in Figure 17.15?[140]

An objective measure for finding the best line

We begin by thinking about what we mean by "best". Mathematically, we want a line that has small residuals. Perhaps our criterion could minimize the sum of the residual magnitudes:

$$|e_1| + |e_2| + \cdots + |e_n| \tag{9}$$

[139] We'll leave it to you to draw the lines. In general, the lines you draw should be close to most points and reflect overall trends in the data.

[140] Further distance from teh sanctuary is associated with smaller carapace size, so the correlation will be negative. Using a computer, the correlation can be computed: -0.855.

which we could accomplish with a computer program. The resulting dashed line shown in Figure 17.15 demonstrates this fit can be quite reasonable. However, a more common practice is to choose the line that minimizes the sum of the squared residuals:

$$e_1^2 + e_2^2 + \cdots + e_n^2 \tag{10}$$

The line that minimizes this **least squares criterion** is represented as the solid line in Figure 17.15. This is commonly called the **least squares line**. The following are three possible reasons to choose Criterion (10) over Criterion (9):

1. It is the most commonly used method.

2. Computing the line based on Criterion (10) is much easier by hand and in most statistical software.

3. In many applications, a residual twice as large as another residual is more than twice as bad. For example, being off by 4 is usually more than twice as bad as being off by 2. Squaring the residuals accounts for this discrepancy.

The first two reasons are largely for tradition and convenience; the last reason explains why Criterion (10) is typically most helpful.[141]

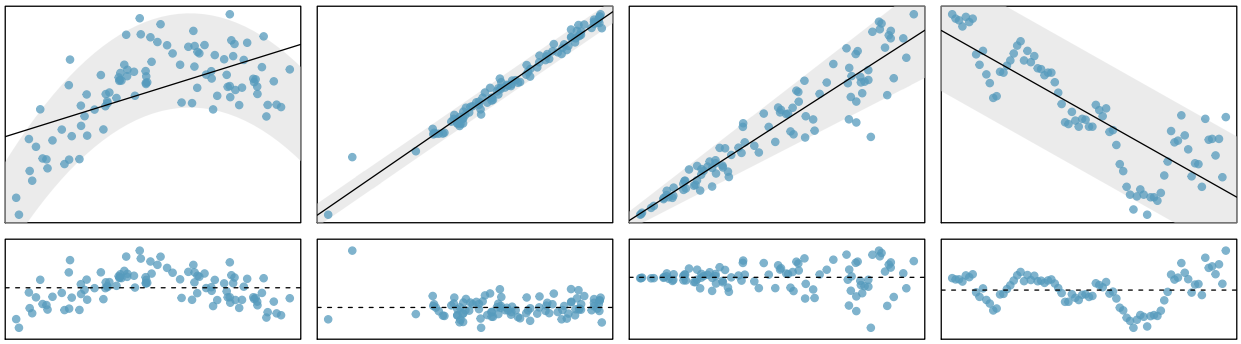


Figure 17.16: Four examples showing when the methods in this chapter are insufficient to apply to the data. In the left panel, a straight line does not fit the data. In the second panel, there are outliers; two points on the left are relatively distant from the rest of the data, and one of these points is very far away from the line. In the third panel, the variability of the data around the line increases with larger values of x. In the last panel, a time series data set is shown, where successive observations are highly correlated.

[141] There are applications where Criterion (9) may be more useful, and there are plenty of other criteria we might consider. However, this book only applies the least squares criterion.

⊙ **Guided Practice 17.11** Should we have concerns about applying least squares regression to the lobster data in Figure 17.15?[142]

Finding the least squares line

For the lobster data, we could write the equation of the least squares regression line as

$$\widehat{\text{length}} = \beta_0 + \beta_1 \times \text{distance from sanctuary}$$

Here the equation is set up to predict size based on distance from the sanctuary, which would be useful when deciding where to fish. These two values, β_0 and β_1, are the *parameters* of the regression line.

The parameters are estimated using observed data. In practice, this estimation is done using a computer in the same way that other estimates, like a sample mean, can be estimated using a computer or calculator. However, we can also find the parameter estimates by applying two properties of the least squares line:

- The slope of the least squares line can be estimated by

$$b_1 = \frac{s_y}{s_x} R \tag{12}$$

  where R is the correlation between the two variables, and s_x and s_y are the sample standard deviations of the explanatory variable and response, respectively.

- If $\bar{x}$ is the mean of the horizontal variable (from the data) and $\bar{y}$ is the mean of the vertical variable, then the point $(\bar{x}, \bar{y})$ is on the least squares line.

We use b_0 and b_1 to represent the point estimates of the parameters β_0 and β_1.

b_0, b_1
Sample estimates of β_0, β_1

⊙ **Guided Practice 17.13** Table 17.2 shows that the sample means for the distance and size are 14.1m and 86mm respectively. Plot the point $(14.1, 86)$ on Figure 17.15 on page 226 to verify it falls on the least squares line (the solid line).

	distance from sanctuary, in metres ("x")	carapace size, in millimetres ("y")
mean	$\bar{x} = 14.1$	$\bar{y} = 86.0$
sd	$s_x = 10.3$	$s_y = 24.3$
		$R = -0.855$

Table 17.2: Summary statistics for distance from sanctuary and carapace size.

[142] The trend appears to be linear, the data fall around the line with no obvious outliers, the variance is roughly constant. These are also not time series observations. Least squares regression can be applied to these data.

Guided Practice 17.14 Using the summary statistics in Table 17.2, compute the slope for the regression line of carpace length against distance from sanctuary.[143]

You might recall the **point-slope** form of a line from ealier maths classes (another common form is *slope-intercept*). Given the slope of a line and a point on the line, (x_0, y_0), the equation for the line can be written as

$$y - y_0 = slope \times (x - x_0) \tag{15}$$

A common exercise to become more familiar with foundations of least squares regression is to use basic summary statistics and point-slope form to produce the least squares line.

TIP: Identifying the least squares line from summary statistics
To identify the least squares line from summary statistics:

- Estimate the slope parameter, b_1, using Equation (12).
- Noting that the point $(\bar{x}, \bar{y})$ is on the least squares line, use $x_0 = \bar{x}$ and $y_0 = \bar{y}$ along with the slope b_1 in the point-slope equation:

$$y - \bar{y} = b_1(x - \bar{x})$$

- Simplify the equation.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24.3193	1.2915	18.83	0.0000
family_income	-0.0431	0.0108	-3.98	0.0002

Table 17.3: Summary of least squares fit for the Elmhurst data.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	114.39	5.13	22.28	0.00
distance	-2.01	0.30	-6.79	0.00

Table 17.4: Summary of least-squares fit to lobster catch data

[143] Apply Equation (12) with the summary statistics from Table 17.2 to compute the slope:

$$b_1 = \frac{s_y}{s_x} R = \frac{24.3}{10.3}(-0.855) = -2.02$$

Example 17.16 Examine the second, third, and fourth columns in Table 17.4. Can you guess what they represent?

We'll describe the meaning of the columns using the second row, which corresponds to β_1. The first column provides the point estimate for β_1, as we calculated in an earlier example: -2.01. The second column is a standard error for this point estimate: 0.2964. The third column is a t-test statistic for the null hypothesis that $\beta_1 = 0$: $T = -6.79$. The last column is the p-value for the t-test statistic for the null hypothesis $\beta_1 = 0$ and a two-sided alternative hypothesis: 0.0000.

Example 17.17 Suppose a cray fisherwoman is considering fishing for lobster in the area outside the sanctuary. Can she simply use the linear equation that we have estimated to calculate the size of the lobster she will catch?

She may use it as an estimate, though some qualifiers on this approach are important. Firstly, the data all may have come from one season, and the realtionship between the distance from the sanctuary and size may change from season to season. Secondly, the equation will provide an imperfect estimate. While the linear equation is good at capturing the trend in the data, no individual lobster's size will be perfectly predicted.

Interpreting regression line parameter estimates

Interpreting parameters in a regression model is often one of the most important steps in the analysis.

Example 17.18 The slope and intercept estimates for the Elmhurst data are 114.39 and 114.39. What do these numbers really mean?

Interpreting the slope parameter is helpful in almost any application. For each additional metre away from the sanctuary, we would expect the size of the lobster to be -114.39cm smaller. We must be cautious in this interpretation: while there is a real association, we cannot interpret a causal connection between the variables because these data are observational.

The estimated intercept $b_0 = 114.39$cm describes the average size of teh lobster at the border of the sanctuary.

Interpreting parameters estimated by least squares

The slope describes the estimated difference in the y variable if the explanatory variable x for a case happened to be one unit larger. The intercept describes the average outcome of y if $x = 0$ *and* the linear model is valid all the way to $x = 0$, which in many applications is not the case.

Extrapolation is treacherous

When those blizzards hit the East Coast this winter, it proved to my satisfaction that global warming was a fraud. That snow was freezing cold. But in an alarming trend, temperatures this spring have risen. Consider this: On February 6^{th} it was 10 degrees. Today it hit almost 80. At this rate, by August it will be 220 degrees. So clearly folks the climate debate rages on.

Stephen Colbert
April 6th, 2010[144]

Linear models can be used to approximate the relationship between two variables. However, these models have real limitations. Linear regression is simply a modelling framework. The truth is almost always much more complex than our simple line. For example, we do not know how the data outside of our limited window will behave.

[144] www.cc.com/video-clips/l4nkoq

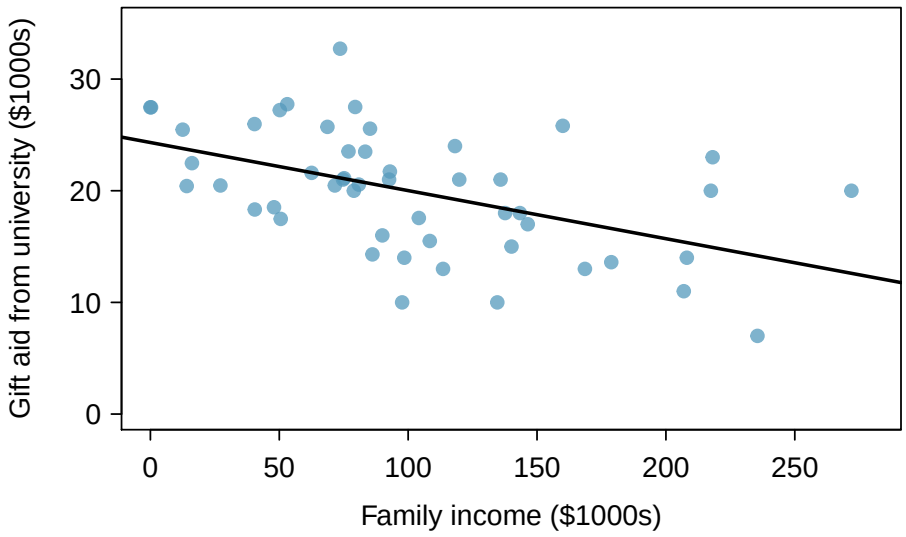


Figure 17.17: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College, shown with the least squares regression line.

Example 17.19 Use the model $\widehat{\text{size}} = 114.39 - 2.01 \times \text{distance (m)}$ to estimate the size of lobsters caught 0.1 kilometres from the sanctuary

---

Recall that the units of distance are metres, so we want to calculate the size for a distance of 100 metres.

$$114.39 - 2.01 \times 100 = -86.61$$

The model predicts that lobsters caught in this area will have a carapace length of -86.61 (!).

Applying a model estimate to values outside of the realm of the original data is called **extrapolation**. Generally, a linear model is only an approximation of the real relationship between two variables. If we extrapolate, we are making an unreliable bet that the approximate linear relationship will be valid in places where it has not been analysed.

Using R^2 to describe the strength of a fit

We evaluated the strength of the linear relationship between two variables earlier using the correlation, R. However, it is more common to explain the strength of a linear fit using R^2, called **R-squared**. If provided with a linear model, we might like to describe how closely the data cluster around the linear fit.

The R^2 of a linear model describes the amount of variation in the response that is explained by the least squares line. For example, consider the lobster data. The variance of the response variable, carapace length, is $s^2_{aid} = 590.93$. However, if we apply our least squares line, then this model reduces our uncertainty in predicting aid using a student's

family income. The variability in the residuals describes how much variation remains after using the model: $s^2_{RES} = 168.52$. In short, there was a reduction of

$$\frac{s^2_{aid} - s^2_{RES}}{s^2_{aid}} = \frac{590.93 - 168.52}{590.93} = 0.71$$

or about 29% in the data's variation by using information about distance from sanctuary for predicting size using a linear model. This corresponds to the adjusted-R-squared value from the R output: 0.71

⊙ **Guided Practice 17.20** 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?[145]

Calculator videos
Videos covering how to find regression coefficients using TI and Casio graphing calculators are available at openintro.org/videos.

Categorical predictors with two levels

Categorical variables are also useful in predicting outcomes. Here we consider a categorical predictor with two levels (recall that a *level* is the same as a *category*). We'll consider Ebay auctions for a video game, *Mario Kart* for the Nintendo Wii, where both the total price of the auction and the condition of the game were recorded.[146] Here we want to predict total price based on game condition, which takes values `used` and `new`. A plot of the auction data is shown in Figure 17.18.

To incorporate the game condition variable into a regression equation, we must convert the categories into a numerical form. We will do so using an **indicator variable** called `cond_new`, which takes value 1 when the game is new and 0 when the game is used. Using this indicator variable, the linear model may be written as

$$\widehat{price} = \beta_0 + \beta_1 \times \texttt{cond\_new}$$

The fitted model is summarized in Table 17.5, and the model with its parameter estimates is given as

$$\widehat{price} = 42.87 + 10.90 \times \texttt{cond\_new}$$

For categorical predictors with just two levels, the linearity assumption will always be satisfied. However, we must evaluate whether the residuals in each group are approximately normal and have approximately equal variance. As can be seen in Figure 17.18, both of these conditions are reasonably satisfied by the auction data.

[145] About $R^2 = (-0.97)^2 = 0.94$ or 94% of the variation is explained by the linear model.
[146] These data were collected in Fall 2009 and may be found at openintro.org.

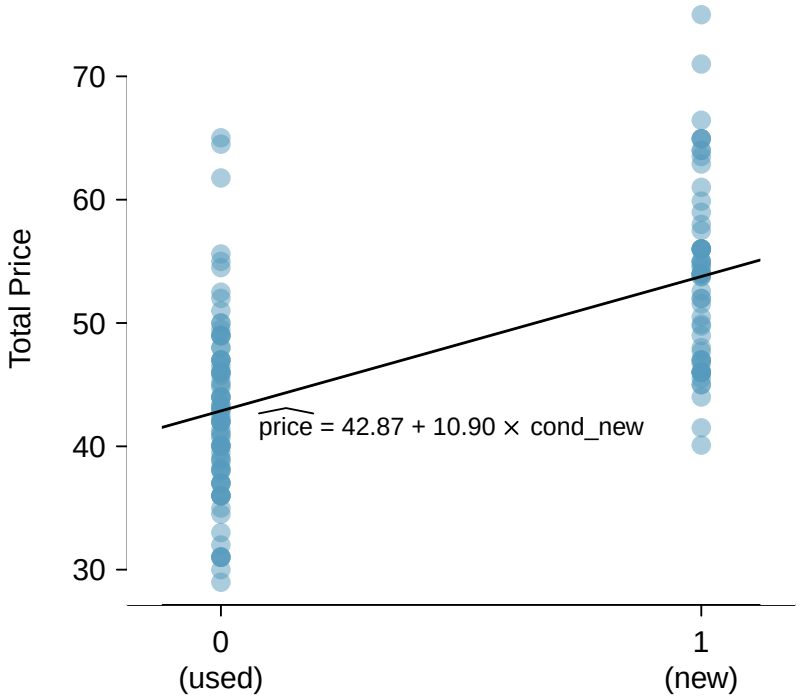


Figure 17.18: Total auction prices for the video game *Mario Kart*, divided into used ($x = 0$) and new ($x = 1$) condition games. The least squares regression line is also shown.

Example 17.21 Interpret the two parameters estimated in the model for the price of *Mario Kart* in eBay auctions.

The intercept is the estimated price when `cond_new` takes value 0, i.e. when the game is in used condition. That is, the average selling price of a used version of the game is $42.87.

The slope indicates that, on average, new games sell for about $10.90 more than used games.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42.87	0.81	52.67	0.0000
cond_new	10.90	1.26	8.66	0.0000

Table 17.5: Least squares regression summary for the final auction price against the condition of the game.

TIP: Interpreting model estimates for categorical predictors.
The estimated intercept is the value of the response variable for the first category (i.e. the category corresponding to an indicator value of 0). The estimated slope is the average change in the response variable between the two categories.

Inference for linear regression

In this section we discuss uncertainty in the estimates of the slope and y-intercept for a regression line. Just as we identified standard errors for point estimates in previous chapters, we first discuss standard errors for these new estimates. However, in the case of regression, we will identify standard errors using statistical software.

US elections and unemployment

Elections for members of the United States House of Representatives occur every two years, coinciding every four years with the U.S. Presidential election. The set of House elections occurring during the middle of a Presidential term are called midterm elections. In America's two-party system, one political theory suggests the higher the unemployment rate, the worse the President's party will do in the midterm elections.

To assess the validity of this claim, we can compile historical data and look for a connection. We consider every midterm election from 1898 to 2010, with the exception of those elections during the Great Depression. Figure 17.19 shows these data and the least-squares regression line:

$$\begin{aligned}&\%\text{ change in House seats for President's party}\\&\qquad = -6.71 - 1.00 \times (\text{unemployment rate})\end{aligned}$$

We consider the percent change in the number of seats of the President's party (e.g. percent change in the number of seats for Democrats in 2010) against the unemployment rate.

Examining the data, there are no clear deviations from linearity, the constant variance condition, or in the normality of residuals (though we don't examine a normal probability plot here). While the data are collected sequentially, a separate analysis was used to check for any apparent correlation between successive observations; no such correlation was found.

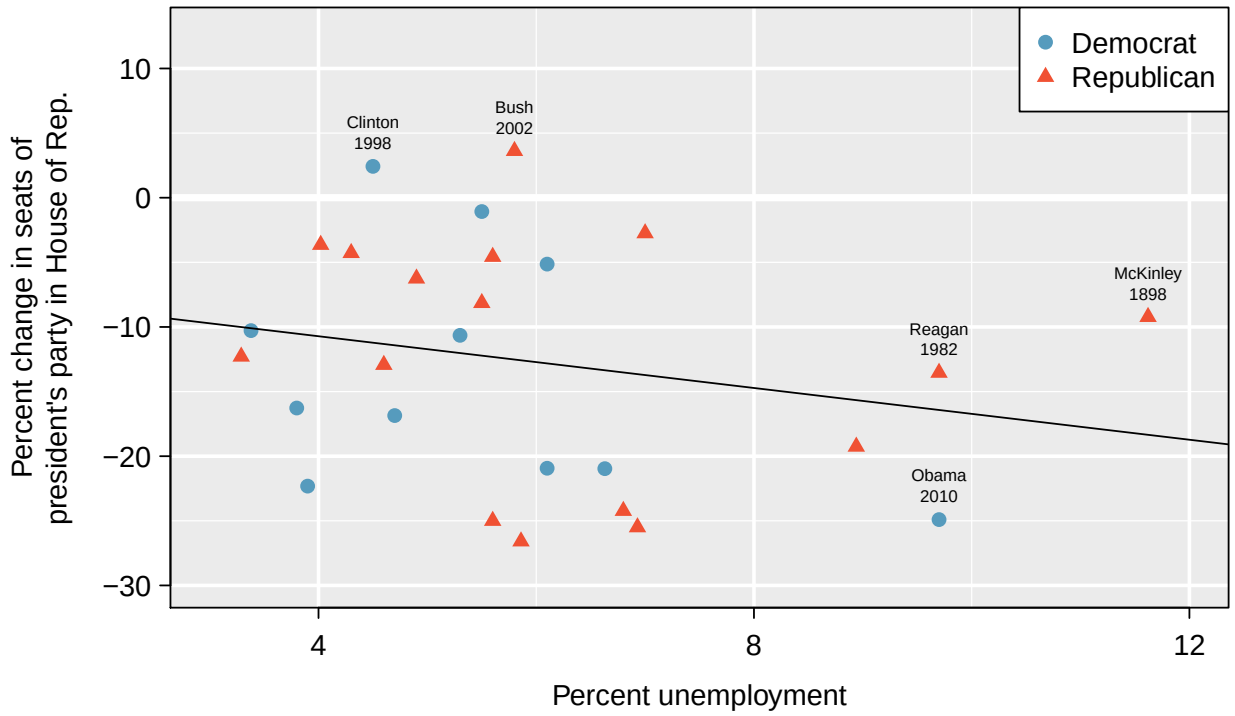


Figure 17.19: The percent change in House seats for the US President's party in each election from 1898 to 2010 plotted against the unemployment rate. The two points for the Great Depression have been removed, and a least squares regression line has been fit to the data.

There is a negative slope in the line shown in Figure 17.19. However, this slope (and the y-intercept) are only estimates of the parameter values. We might wonder, is this convincing evidence that the "true" linear model has a negative slope? That is, do the data provide strong evidence that the political theory is accurate? We can frame this investigation into a one-sided statistical hypothesis test:

H_0: $\beta_1 = 0$. The true linear model has slope zero.

H_A: $\beta_1 < 0$. The true linear model has a slope less than zero. The higher the unemployment, the greater the loss for the President's party in the House of Representatives.

We would reject H_0 in favour of H_A if the data provide strong evidence that the true slope parameter is less than zero. To assess the hypotheses, we identify a standard error for the estimate, compute an appropriate test statistic, and identify the p-value.

Understanding regression output from software

Just like other point estimates we have seen before, we can compute a standard error and test statistic for b_1. We will generally label the test statistic using a T, since it follows the t-distribution.

We will rely on statistical software to compute the standard error and leave the explanation of how this standard error is determined to a second or third statistics course. Table 17.6 shows software output for the least squares regression line in Figure 17.19. The row labelled *unemp* represents the information for the slope, which is the coefficient of the unemployment variable.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6.7142	5.4567	-1.23	0.2300
unemp	-1.0010	0.8717	-1.15	0.2617
				$df = 25$

Table 17.6: Output from statistical software for the regression line modelling the midterm election losses for the President's party as a response to unemployment.

Example 17.22 What do the first and second columns of Table 17.6 represent?

The entries in the first column represent the least squares estimates, b_0 and b_1, and the values in the second column correspond to the standard errors of each estimate.

We previously used a t-test statistic for hypothesis testing in the context of numerical data. Regression is very similar. In the hypotheses we consider, the null value for the slope is 0, so we can compute the test statistic using the T (or Z) score formula:

$$T = \frac{\text{estimate} - \text{null value}}{\text{SE}} = \frac{-1.0010 - 0}{0.8717} = -1.15$$

We can look for the one-sided p-value – shown in Figure 17.20 – using the probability table for the t-distribution or statistical software.

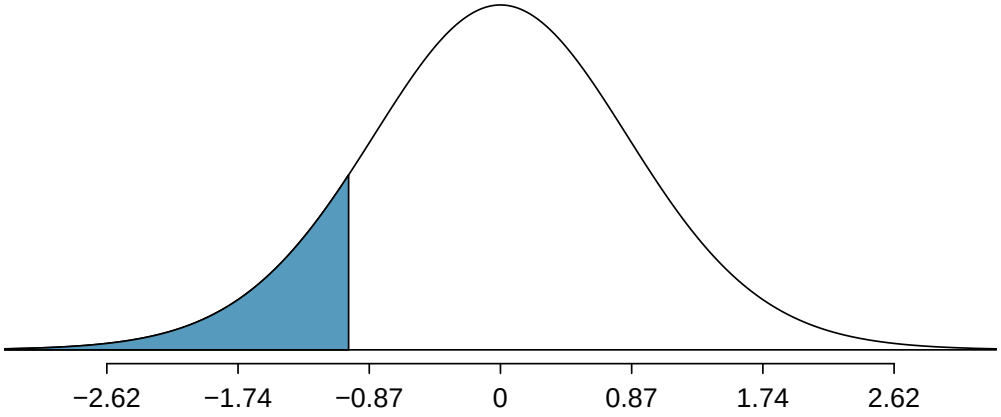


Figure 17.20: The distribution shown here is the sampling distribution for b_1, if the null hypothesis was true. The shaded tail represents the p-value for the hypothesis test evaluating whether there is convincing evidence that higher unemployment corresponds to a greater loss of House seats for the President's party during a midterm election.

Inference for regression

We usually rely on statistical software to identify point estimates and standard errors for parameters of a regression line. After verifying conditions hold for fitting a line, we can use the methods learned in Section 10.5 for the t-distribution to create confidence intervals for regression parameters or to evaluate hypothesis tests.

TIP: Always check assumptions

If conditions for fitting the regression line do not hold, then the methods presented here should not be applied. The standard error or distribution assumption of the point estimate – assumed to be normal when applying the t-test statistic – may not be valid.

18 Regression with Transformations

Once we add the log transformation as a possibility – for either the x-variable, the y-variable, or both – we can describe many possible data trends. The only issue is that we need to make sure we know how to interpret the slope estimate in our model after the transformation. The four possible scenarios are described below.

Model 1. linear-linear model:

$$Y_i = \alpha + \beta X_i + e_i \tag{1}$$

How to interpret the slope for model 1: We say a one unit change in X, on average, leads to a β unit change in Y.

Model 2. linear-log model:

$$Y_i = \alpha + \beta log X_i + e_i \tag{2}$$

How to interpret the slope for model 2: We say a one percent change in X, on average, leads to a $\beta \div$ 100 unit change in Y.

Model 3. log-linear model:

$$log Y_i = \alpha + \beta X_i + e_i \tag{3}$$

How to interpret the slope for model 3: We say a one unit change in X, on average, leads to a $\beta \times$ 100 percent change in Y.

Model 4. log-linear model:

$$log Y_i = \alpha + \beta log X_i + e_i \tag{4}$$

How to interpret the slope for model 4: We say a one percent change in X, on average, leads to a β percent change in Y.

18.1 Log Transformations

For this example we are interested in modelling the relationship between a biodiversity index measure and altitude. Specifically, we are interested in understanding if biodiversity varies with altitude.

Our `y` or dependent/response variable is the diversity index measure, and our `x` or independent/explanatory variable is altitude above sea level. When creating linear regression models and working with scatterplots we give **R** the formula: `y ~ x`. Deciding which variable goes on the y-axis and which variable goes on the x-axis is tricky. The convention is to use the variable that we think is doing the explaining on the horizontal (x-axis). Because we

think that altitude above sea level might 'explain' the observed level of biodiversity, we place altitude on the horizontal axis and biodiversity on the vertical.

Let's read in the data and have a look at what we have. Recall that normally you would have the data in an Excel file. Here I type the data in so there is a complete record of the values used.

```
# read in our data
diversity.data<- data.frame(
  altitude= c(56, 57.4 , 59.3, 60.6, 61.1, 62.6, 69.5, 80.6,
              93, 104.3, 109.4, 120.3, 132.9, 138.3     ),
  d.index= c(0.1846, 0.2760, 0.4635, 0.0898, 0.1082, 0.2593, 0.06408, 0.0419,
             0.0670, 0.0405, 0.0320, 0.0174, 0.0156, 0.0399 ))

# Use both str, summary to look at data
str(diversity.data)

## 'data.frame': 14 obs. of  2 variables:
##  $ altitude: num  56 57.4 59.3 60.6 61.1 ...
##  $ d.index : num  0.1846 0.276 0.4635 0.0898 0.1082 ...

summary(diversity.data)

##     altitude         d.index
##  Min.   : 56.00   Min.   :0.01560
##  1st Qu.: 60.73   1st Qu.:0.04005
##  Median : 75.05   Median :0.06554
##  Mean   : 86.09   Mean   :0.12141
##  3rd Qu.:108.12   3rd Qu.:0.16550
##  Max.   :138.30   Max.   :0.46350
```

As we have two continuous variables, rather than a boxplot we create a scatter plot. Note the form of the scatter plot formula is `y ~ x` so we have `d.index ~ altitude` in the plot formula. Also note that in your own plots you will just have the figure caption, not a figure caption and a figure title.

```
with(diversity.data, plot(d.index~altitude, # create our scatter plot
     pch= 4,
     main= "Biodiversity and altitude relationship",
     ylab= "Biodiversity (Index)",
     xlab= "Height above sea level (m)",
     ylim= c(0,0.5), xlim= c(20,150), las= 1))
```

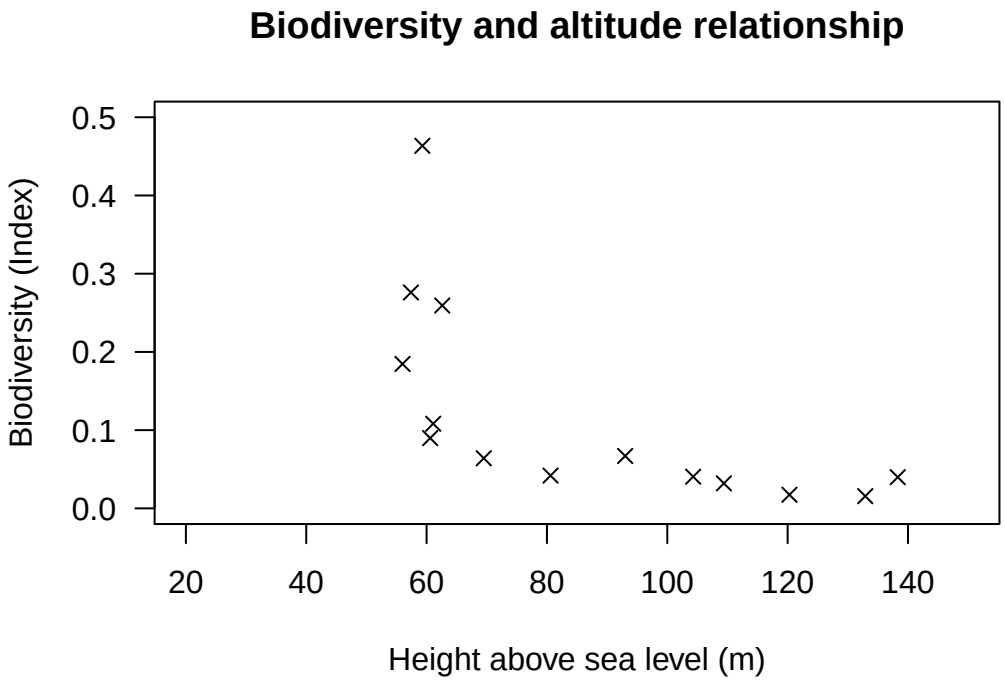


Figure 18.1: Scatter plot showing the relationship between biodiversity and height above sea level.

The plot looks quite different from the previous lobster example. The plot shows a clear non-linear relationship. Let's create three scatter plots showing the possible log transformations of our data (log-linear, linear-log and log-log), and our original plot (linear-linear) to see if any of these transformation generate an approximately linear relationship.

To get the log of a value or set of values we use the function *log()*. We will also use a new function which allows us to set different graphical parameters for our figures, *par()*. We will use *par()* to set the number of plots per plotting window using parameter `mfrow`. Here we would like 2 rows × 2 columns so we can see all 4 plots together to more easily compare them. i.e. we are creating a grid to display each individual plot at the same time.

```
# set the number of plots (row, column) per window
par(mfrow= c(2,2)) # 2 rows and 2 columns = four plots

# plot 4 scatter plots
with(diversity.data, plot(d.index~altitude,
                  main= "Raw Data: linear - linear")) # untransformed
with(diversity.data, plot(log(d.index)~altitude,
                  main= "Transformed: log - linear")) # log(y)
with(diversity.data, plot(d.index~log(altitude),
```

```
                    main= "Transformed: linear - log")) # log(x)
with(diversity.data, plot(log(d.index)~log(altitude),
                    main= "Transformed: log - log"))          # log(x) & (y)
```

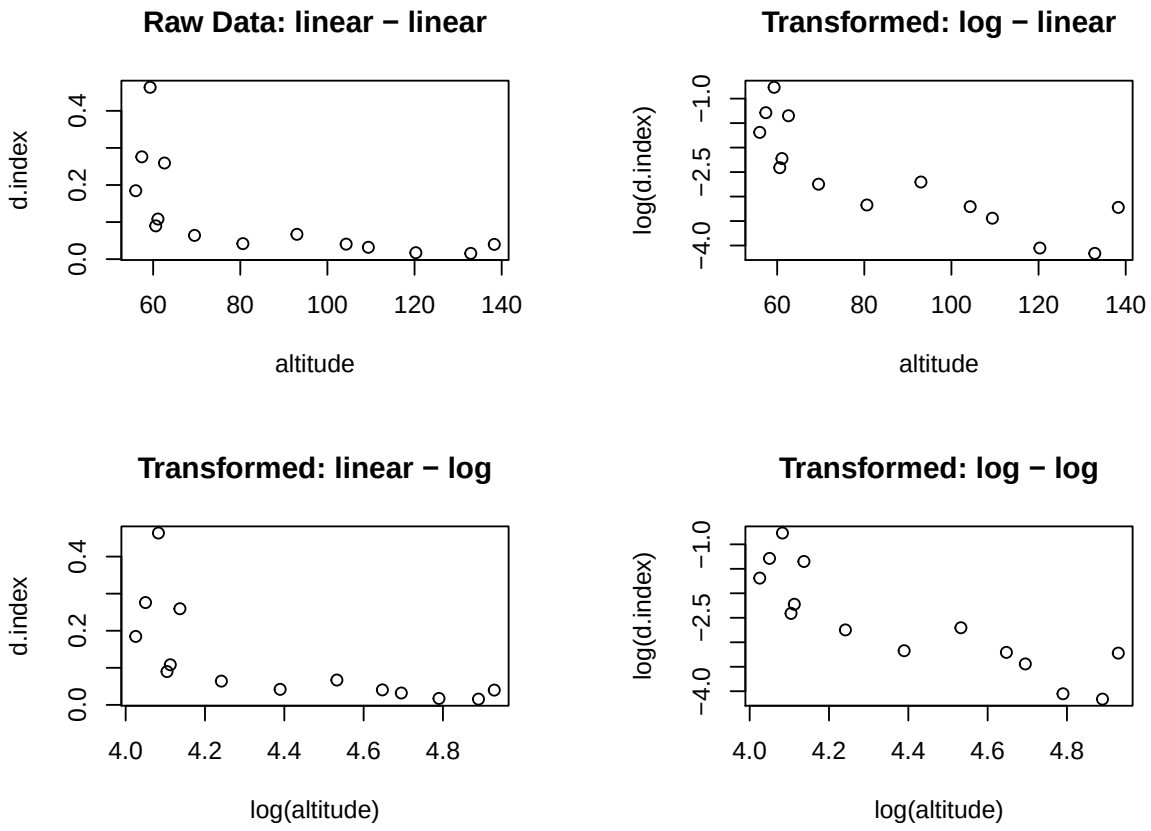


Figure 18.2: Scatter plots showing the relationship between biodiversity and altitude using untransformed data and three different log transformations.

Which, if any, of the transformations achieved an approximately linear relationship? It is not 100 percent clear, but i am going to say that the log-log model seems to be the best fit across the available options. Let's create a linear trend line for the data using the log-log model and add the trend line to a scatter plot on the log-log scale.

```
# create our model Note the way we tell R the data file name
lm.Diversity <- lm(log(d.index) ~ log(altitude), data = diversity.data)

# print the intercept and slope coefficients
coef(lm.Diversity)
```

```
##   (Intercept) log(altitude)
##      9.157182     -2.671504
```

From the output have the intercept estimate of 9.157 and the slope estimate -2.672. As we have fitted a log-log model, the equation tells us that for every 1 percent increase in altitude, biodiversity decreases, on average, by 2.67 percent. When we have a regression equation we always add the comment, on average. The intercept does not have an intuitive meaning in this model, so we restrict our focus to the slope.

Note: You have MUCH less confidence using a model to extrapolate beyond the range of your data. Formally we only know what is happening for the data range we observe. So, this model may be useful for predicting biodiversity only for altitudes between 56 m and 138 m above sea level. Outside this range the model has little to say, and it would be dangerous to extrapolate to other altitudes.

The last step is to graphically display our model. We can do this by adding a trend line to our scatter plot with the *abline()* function and identifying our line in a legend with the *legend()* function. Note we can add the line either by specifying the numerical values directly (a = intercept and b = slope), or by telling R to get the slope and intercept from our regression model (lm.diversity).

```
#  Base plot
with(diversity.data, plot(log(d.index)~log(altitude),
     pch= 4,
     ylab= "Biodiversity Index (log scale)",
     xlab= "Height above sea level (log scale)",
     las= 1 ))

# add our trend line
abline(lm.Diversity, lty= 2, lwd= 2, col= "blue")

# add a pretty fancy legend
legend("topright", legend=c("Trend line","Observation"), lty=c(2,NA),
       pch =c(NA,4),lwd=c(2,NA), col =c("blue","black"), bty="n")
```

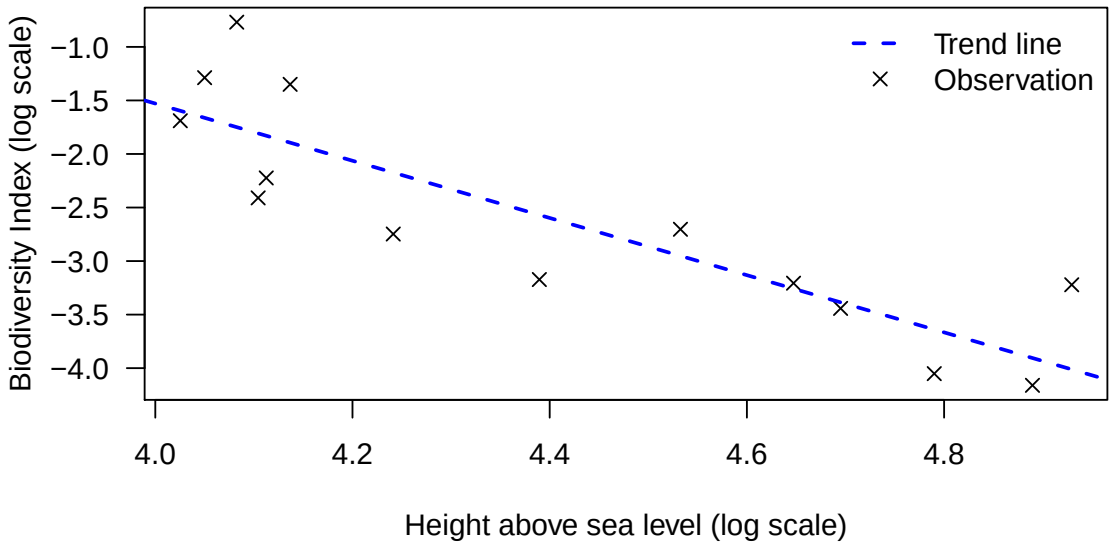


Figure 18.3: Relationship between biodiversity and altitude (log-log scale).

18.2 Statistical Significance of the Slope Estimate

The decision rule we use to generate an estimate of the slope and intercept (the least squares estimator) will always generate estimates of the slope and intercept. However, we are interested in knowing whether or not the slope estimate we obtain is statistically different from zero. The question we are asking is could the pattern we observe in the data be due to chance, or is it a real trend?

To determine whether or not the slope estimate is statistically different from zero we conduct a t-test. The t-test for the slope (and intercept) works just the same way as our earlier t-tests, and we will use the same decision rule. Specifically, if the t-test p-value is less than 0.05 we will reject the null hypothesis that the slope estimate is equal to zero. In practice this decision rule means that we have set the test alpha level at 0.05. With regression models it is often the case that people use multiple standards for testing rather than just 0.05. This added feature is usually captured with an additional footnote in the summary table.

To conduct the t-test we use the *summary()* function. The *summary()* function will conduct a t-test on both the slope and the intercept, and will also report some additional information on our linear model. This automated routine is one of the very useful things about **R**. While MS Excel will fit a trend line, the default output does not tell you whether

or not the estimates are statistically different from zero. Let's run a formal test on the model we created previously, and saved as: `lm.Diversity`.

Step 1: Set the Null and Alternate Hypotheses
Here we will be testing only the slope as we are not interested in the intercept.

- Null hypothesis: The slope is equal to zero
- Alternate hypothesis: The slope is not equal to zero

Step 2: Print the test output
We obtain all the output with a simple command *summary*(`lm.Diversity`). The output we get has many elements but we will work through the detail in stages.

```
summary(lm.Diversity)

##
## Call:
## lm(formula = log(d.index) ~ log(altitude), data = diversity.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60309 -0.40759 -0.07473  0.34353  0.98058
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.157      1.982   4.621 0.000589 ***
## log(altitude)   -2.671      0.449  -5.950 6.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5436 on 12 degrees of freedom
## Multiple R-squared:  0.7468,Adjusted R-squared:  0.7257
## F-statistic:  35.4 on 1 and 12 DF,  p-value: 6.715e-05
```

Step 3: Interpret the Results
The output here is quiet a bit more complex than for basic t-tests. Let's focus, for now, on the 'Coefficients' section of the output. This section of the output provides the estimates for our trend line and the results for the test of statistical significance. Under the 'Estimate' column we see the same values as we obtained when using the *coef()* function - the intercept and slope terms (distance is the slope estimate).

The next column is the 'Std. Error' column and the values in this column are the standard error for each coefficient. The standard error in a regression model has the same interpretation as the standard error of the mean in our earlier t-test examples. These values can be thought of as a measure of uncertainty for our slope and intercept intercept values.

Next, we can see t-values. The t-values have been calculated using the standard approach. Specifically, these values have been calculated as: Estimate value minus the null hypothesis test value divided by the standard error. In this instance our null hypothesis test value is zero, so the t-value is calculated as: (-2.671 - 0 / 0.449) = -5.95. Lastly we can see the p-values. The p-values are what we use to make a decision about whether or not the slope estimate is statistically different from zero. Since the p-value for the slope is: `6.71e-05` (< 0.001), which is smaller than 0.05, we:

- **Reject** the null hypothesis that the slope is equal to zero.

Why do we care if the slope estimate is different to zero? Well, if the slope is zero this means that there is no real trend in the data. The variation we have observed is just due to sampling variation. Fundamentally, we are interested in knowing whether the trend line we fit to the data represents a real trend line or not. We want to know whether or not there really is a relationship between altitude and diversity.

18.3 Measure of Model Fit

The final thing we are interested in is the R^2 value. This value has an interpretation as the proportion of the variation in the data explained by the model. In **R**, the R^2 value is reported as the `Multiple R-squared value`, which is the second last line of the summary output. See if you can find it in the output below.

```
summary(lm.Diversity)

##
## Call:
## lm(formula = log(d.index) ~ log(altitude), data = diversity.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60309 -0.40759 -0.07473  0.34353  0.98058
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.157      1.982   4.621 0.000589 ***
## log(altitude)   -2.671      0.449  -5.950 6.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5436 on 12 degrees of freedom
## Multiple R-squared:  0.7468,Adjusted R-squared:  0.7257
## F-statistic:  35.4 on 1 and 12 DF,  p-value: 6.715e-05
```

As the R^2 value = `0.747`, we say that the model explains 74.7% of the variation in the data. We don't really have a view regarding whether this is a good thing or a bad thing, but it is a commonly reported metric, and something that we will report. There are other metrics of model fit, for example **R** also reports an `Adjusted R-squared value`. These other metrics do have advantages, but these alternative metrics do not have the same nice interpretation of percent/proportion of the variation in the data explained by the model. For this reason we will stick with reporting the R^2 value.

18.4 The Summary Table

Because there are multiple t-tests (one for the intercept and one for the slope), and because there is additional information to report, such as the R^2 value, the convention for reporting linear regression results in to use a summary table. The summary table contains: (i) the estimates of the slope and the intercept; (ii) the standard error of the slope and the intercept, where a '*' sign is used to denote statistical significance; (iii) the number of observations in the data set; and (iv) the R^2 value. The difference between the estimate and the standard error is usually indicated with parentheses. Because it is possible to report either standard error information, or t-value information, the convention is to add a footnote to the table to say "Standard errors in parentheses." The thresholds chosen for statistical significane also vary.

Because we use p-value < 0.05 as the critical decision threshold for our t-tests, when we move to a more general system we use threshold values around 0.05 that are both more and less stringent i.e. the threshold values we use for the '*' sign system of denoting statistical significance are '*' for p-value < 0.1; '**' for p-value < 0.05; '***' for p-value < 0.01. Below is an example of an acceptable format for a regression summary table, but there are many other acceptable formats. Because we use p-value < 0.05 as the critical decision threshold for our t-tests, when we move to a more general system we use threshold values around 0.05 that are both more and less stringent i.e. the threshold values we use for the '*' sign system of denoting statistical significance are '*' for p-value < 0.1; '**' for p-value < 0.05; '***' for p-value < 0.01. Below is an example of an acceptable format for a regression summary table, but there are many other acceptable formats.

Table 18.1: Altitude and Diversity linear regression

	Log Diversity Index
Intercept	9.16^{***} (1.98)
Log Altitude	-2.67^{***} (0.45)
Observations	14
R^2	0.75
Note:	*p<0.1; **p<0.05; ***p<0.01 Standard errors in parentheses.

Reference

Marek Hlavac (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2. http://CRAN.R-project.org/package=stargazer

Yihui Xie (2016). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.14.http://CRAN.R-project.org/package=knitr

19 Nonparametric Methods

The statistical inference methods that we have seen so far can be classed as *parametric* methods. We have an understanding of the the theoretical properties of different distributions of random variables. These distributions are defined by their parametric form: the Normal distribution has a mean and a variance; the t-distribution is defined by its mean and degrees of freedom. We then make assumptions about the underlying structure of the data we observe or about the population from which the data are drawn in order to make use of these distributions, and we make inference about the parameters of the distribution.

If we are not so sure that these parametric assumptions hold, we may prefer methods that are valid without such assumptions: *non-parametric* methods. These methods are robust to mis-specification of the underlying distribution of the data because they do not assume a particular distribution in the first place.

Non-parametric methods have long been used to establish whether or not two or more groups are the same. The tests work by first ranking the data and then using the rank values, not the actual data values. For this reason the tests are sometimes referred to as robust tests: the tests are robust to the influence of outliers. The situation is analogous to the comparison of the mean and the median. To find the median we rank the data from highest to lowest and look at the middle value. The median is therefore referred to as a measure that is robust to the influence of outliers. This is in contract to the mean, which can be influenced by outliers.

The difference between using the actual data for a test, and using the rank data can be understood with a simple example. Consider the set: `10, 5, 15, 4, 7, 14, 79`. This set of numbers has one relatively extreme value, and this value has a strong effect on the mean but not the median. Specifically the mean is `17.1`, which is greater than six of the seven numbers, while the median is `10`.

Rather than use the raw data values, non-parametric methods convert the raw data into ranks first and then use the rank values in the test. The mapping of the raw data to rank values is shown in Table 1. The smallest value in the data set is `4` and so the value `4` is given a rank value of `1`. The second smallest value in the data set is `5` and so the value `5` is given a rank value of `2`. The process continues until we get to the largest value in the data set, which is `79`, and `79` is given a rank value of `7`.

Table 19.1: Mapping raw data to rank values

Raw data values	10	5	15	4	7	14	79
Data rank values	4	2	6	1	3	5	7

If the assumptions of the parametric model are valid, then parametric methods have a greater ability to detect a difference, if a difference exists. For this reason we generally only use non-parametric methods when we can not appeal to the central limit theorem. Recall that the central limit theorem tells us that for large sample sizes, the distribution of the

sample mean will tend towards a Mornal distribution. So we prefer non-parametric methods generally when we have small samples. Deciding when it is necessary to use a non-parametric approach is tricky, so we will leave that for now and just focus on making sure we understand how to implement the key non-parametric approaches.

As with all previous statistical tests, we will set the alpha level at 0.05.

19.1 Two Sample Non-Parametric Test: Mann-Whitney Test

Rather than testing whether two samples come from populations with different means (the mean being a parameter), a non-parametric approach tests whether two samples came from two populations with the same distribution, or different distributions. The key here is that we don't make any assumption about what type of distributions we are dealing with, we only care about whether the distributions are the same or not.

Under the null hypothesis that the two samples are independent samples from the same distribution, we know what the probabilities are of a unit in one sample being higher or lower than the units in the other sample. The Mann-Whitney test compares these theoretical probabilities under the null hypothesis to the observed realisations of which units are higher or lower than others in the two samples. In this way we get a probability of having observed two samples at least as different as the two samples in front of us, if the null hypothesis were true. Thus, we have a p-value for the test of whether the two samples come from the same distribution or not.

For this example we use the same phytoremediation data used for the two sample t-test. Recall, phytoremediation is the use of plants for remediating contaminated soil and water, and plant species are selected based on their ability to uptake or stabilize specific contaminants at a site.

Here we will look at the efficiency of two crop plants (redbeet and barley) at removing cadmium from the top 20 cm of soil at a contaminated site. The data we have is the percent reduction of cadmium after one harvest. Because we only have 15 observations for each plant we decide to use a non-parametric test. Because we have two independent samples we use a Mann-Whitney two sample test. So that the example can be reproduced, below I type in the data. In general your first step in to read in the data from a .csv file not type the data directly into **R** .

```
# read in our data (wide format)
Cd.BeetBarley<- data.frame(
  redbeet= c(18, 5, 10, 8, 16, 12, 8, 8, 11, 5, 6, 8, 9, 21, 9),
  barley= c(8, 5, 10, 19, 15, 18, 11, 8, 9, 4, 5, 13, 7, 5, 7))

# check the data structure
str(Cd.BeetBarley)
```

```
## 'data.frame': 15 obs. of  2 variables:
##  $ redbeet: num  18 5 10 8 16 12 8 8 11 5 ...
##  $ barley : num  8 5 10 19 15 18 11 8 9 4 ...

#  the structure information tells un we have two columns of data, and
#  that there are 15 observations in each group
```

Once we understand the data structure, and are confident the data has been read into **R** correctly, we then create a boxplot to look at the data. Remember, the boxplot is an essentially part of the process whether we are using a parametric test or a non-parametric test.

```
with(Cd.BeetBarley,boxplot(redbeet,barley,
     col= "lightgrey",
     main= "Phytoremediation Efficiency of Crop Plants",
     xlab= "Crop type", ylab= "Cadmium reduction (%)",
     names= c("Redbeet","Barley"),
     ylim=  c(0,25), las= 1,
     boxwex=0.6))
```

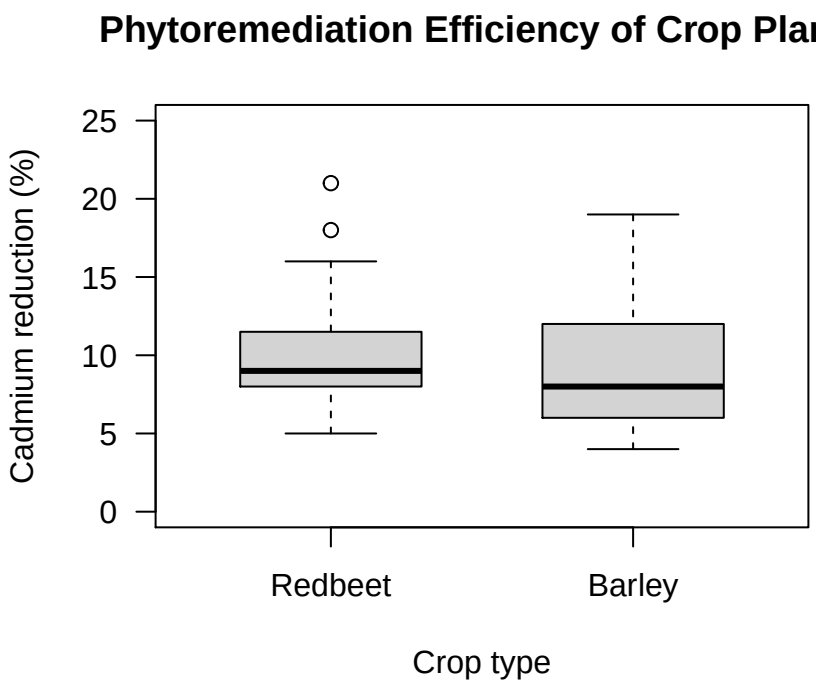


Figure 19.1: Box plot comparing the phytoremediation efficiency of redbeet and barley crop plants for removing cadmium from contaminated soil at depths of 0 to 20 cm.

Unlike the two sample t-test, for non-parametric tests there is no variance assumption to think about. We just jump straight to our main test for a difference between the groups. For non-parametric tests the statement about the null and the alternate hypothesis do not refer to the mean, but are just general statements.

Step 1: Set the Null and Alternate Hypotheses

- Null hypothesis: The effectiveness of Redbeet and Barley in removing cadmium from the top 20cm of soil is the same
- Alternate hypothesis: The effectiveness of Redbeet and Barley in removing cadmium from the top 20cm of soil is not the same

Step 2: Implement the test

We use the *wilcox.test()* function to implement the test, and the test notation is essentially the same as for the *t.test()* function. Although there is no equal/unequal variance parameter to set for the test, to suppress the warnings – which can be really annoying – when using the test we set the parameter `exact = FALSE`.

```
with(Cd.BeetBarley, wilcox.test(redbeet, barley, exact = FALSE))

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  redbeet and barley
## W = 126.5, p-value = 0.5728
## alternative hypothesis: true location shift is not equal to 0

# use exact=FALSE to supress the warnings
```

Step 3: Interpret the Results

From the output we see the `p-value = 0.5728`. Since 0.573 is greater than 0.05 (alpha), we:

- **Fail to reject** the null hypothesis that the effectiveness of Redbeet and Barley in removing cadmium from the top 20cm of soil is the same.

In other words: We do **not** have **sufficient evidence** to identify any kind of difference. If you were trying to determine which plant to use to remediate a particular site you would likely look to other factors, such as cost, environmental conditions or additional contaminants to be removed, to assist you in making your decision. Alternatively, you could think about collecting further data. With more data there is an increase in test power, and hence an increase in the ability to detect a difference if a difference exists.

19.2 Paired Non-Parametric Test: Wilcoxon Test

As with the Mann-Whitney test, the Wilcoxen test looks at pairs of observations and compares what is observed with what would be expected under the null hypothesis. The Wilcoxen test also makes use of the paired nature of the data, to test the null hypothesis that the the true distribution of the difference between observations is centered at zero and is symmetric. The difference we are talking about here is calculated for each individual in the sample, the difference of that individual's observation under one condition minus the observation under the other.

This example is similar to the paired t-test example, and we will again be looking at before and after data from a contaminated site. Groundwater at the site was remediated using a Pump & Treat method with the intention of removing unacceptable levels of Total Petroleum Hydrocarbons (TPC). TCP is a general term encompassing hundreds of hydrocarbon based compounds. The paired measurements, recorded in micrograms per litre, were taken at observation wells around the contaminated site before and after the treatment.

We would like to determine whether the treatment has been effective - i.e. we want to determine whether or not there are differences between the measurements taken before and after the treatment. However, as we only have nine before and after measurements we decide we will use a non-parametric test rather than a paired t-test.

Generally you will read the data in from an MS Excel .csv file or similar, but here we input the data directly to provide a complete record of all the data needed for the test.

```
#test
# read in our data (wide format)
TPH.remediation<- data.frame(
before= c(1475.7, 1292.2, 1575.9, 1440.8, 1606.1, 1425.1, 1502.3, 1327.4, 1526.4),
after= c(695.1, 706.1, 675.5, 706.6, 717.8, 729.4, 722.3, 668.2, 714.4))

# Look at the raw data
str(TPH.remediation)
```

```
## 'data.frame': 9 obs. of  2 variables:
##  $ before: num  1476 1292 1576 1441 1606 ...
##  $ after : num  695 706 676 707 718 ...
```

```
summary(TPH.remediation)
```

```
##      before         after
##  Min.   :1292   Min.   :668.2
##  1st Qu.:1425   1st Qu.:695.1
##  Median :1476   Median :706.6
##  Mean   :1464   Mean   :703.9
##  3rd Qu.:1526   3rd Qu.:717.8
##  Max.   :1606   Max.   :729.4
```

With our data in **R**, we then move on to our standard process of analysis. The first step is a boxplot. Because we have paired data we are then going to create a horizontal boxplot of the difference in the observation pairs.

```
with(TPH.remediation,boxplot(before - after,
     col= "lightgray", boxwex= 1.2,
     main= "Pump and Treat Groundwater Remediation of
     Total Petroleum Hydrocarbons (TPH)",
     xlab= "Differences in TPH (ug/L)", ylim= c(500,1000),
      horizontal= TRUE))
```

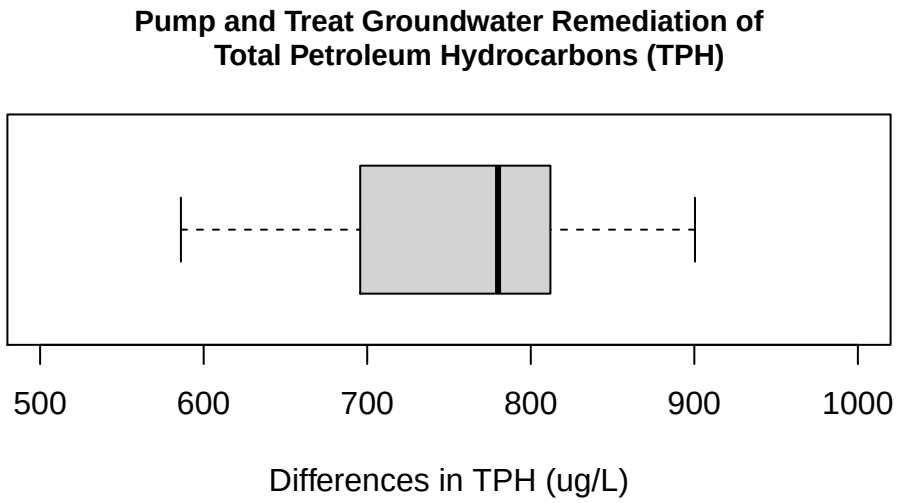


Figure 19.2: Box plot showing the differences in groundwater measurements of Total Petroleum Hydrocarbons (ug/L) at monitoring wells before and after pump and treat remediation.

From the horizontal boxplot it looks like our remediation was successful. All the differences are positive. Although, remember that for the test we do not use the actual data but the rank values.

Step 1: Set the Null and Alternate Hypotheses

- Null hypothesis: There is no difference in the before and after measurements
- Alternate hypothesis: There is a difference in the before and after measurements

Step 2: Implement the Wilcoxon test for paired data

Here we set the parameter: `paired=TRUE` to indicate that we have a paired test, and we also set the parameter `exact=FALSE` to suppress the warnings. As always the alpha level is set at 0.05.

```r
with(TPH.remediation, wilcox.test(before, after, paired = TRUE, exact = FALSE))

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  before and after
## V = 45, p-value = 0.009152
## alternative hypothesis: true location shift is not equal to 0
```

Step 3: Interpret the Results

From the output we see the `p-value = 0.009152`, which we would generally report as < 0.001, which is smaller than 0.05. So, we:

- **Reject** the null hypothesis that the means are equal

In words: We have **sufficient evidence** to say the concentration of Total Petroleum Hydrocarbons is different after remediation.. We know remediation had a positive effect by looking at the data values and the boxplot. That the remediation has had an impact is great news; but we still needs to be compare actual pollution levels to acceptable levels for the relevant land use, and understand things such as the cost effectiveness of the remediation. As always, the formal testing aspect is just a small part of the overall process.

19.3 Multiple Group: Kruskal-Wallis Test

The Kruskal-Wallis test is the non-parametric analogue of the familiar ANOVA for difference between means in multiple groups. The test examines the ranks of observations in the sample as a whole and measures how these ranks vary between groups in order to test whether or not all observations in the sample come from the same distribution.

For this example we will use the base **R** data set **chickwts**. The data set includes information on the weights of chickens fed six different types of feed, and the data is in long format. When data is in long format it means we have one continuous variable (weight) and one factor variable (feed) that has six levels. When working with multiple groups it is easiest to work with data in long format. What we are trying to do is help a farmer identify a feed which promotes the greatest weight gain in chickens.

Just as for the parametric approach, we start our investigation by looking at our data using the *str()* and *summary()* functions, and creating a boxplot. When the data are in long format if we want to get the quantile information that is plotted in the boxplot we use the *tapply()* function.

Note: This is the same data set used in the ANOVA R reference guide, so you can compare the different results.

```r
# get summary & structure and create a boxplot of our differences
str(chickwts)
```

```
## 'data.frame': 71 obs. of  2 variables:
##  $ weight: num  179 160 136 227 217 168 108 124 143 140 ...
##  $ feed  : Factor w/ 6 levels "casein","horsebean",..: 2 2 2 2 2 2 2 2 2 2 ...
```

```r
summary(chickwts)
```

```
##      weight             feed
##  Min.   :108.0   casein   :12
##  1st Qu.:204.5   horsebean:10
##  Median :258.0   linseed  :12
##  Mean   :261.3   meatmeal :11
##  3rd Qu.:323.5   soybean  :14
##  Max.   :423.0   sunflower:12
```

```r
with(chickwts,tapply(weight,feed,quantile))
```

```
## $casein
##     0%    25%    50%    75%   100%
## 216.00 277.25 342.00 370.75 404.00
##
## $horsebean
##     0%    25%    50%    75%   100%
## 108.00 137.00 151.50 176.25 227.00
##
## $linseed
##     0%    25%    50%    75%   100%
## 141.00 178.00 221.00 257.75 309.00
##
## $meatmeal
##    0%   25%   50%   75%  100%
## 153.0 249.5 263.0 320.0 380.0
##
## $soybean
##     0%    25%    50%    75%   100%
## 158.00 206.75 248.00 270.00 329.00
##
## $sunflower
##     0%    25%    50%    75%   100%
## 226.00 312.75 328.00 340.25 423.00
```

```r
with(chickwts,boxplot(weight~feed,
     col= "lightgray",
     main= "",
     xlab= "Feed type", ylab= "Weight (g)", ylim= c(100,450), las= 1))
```

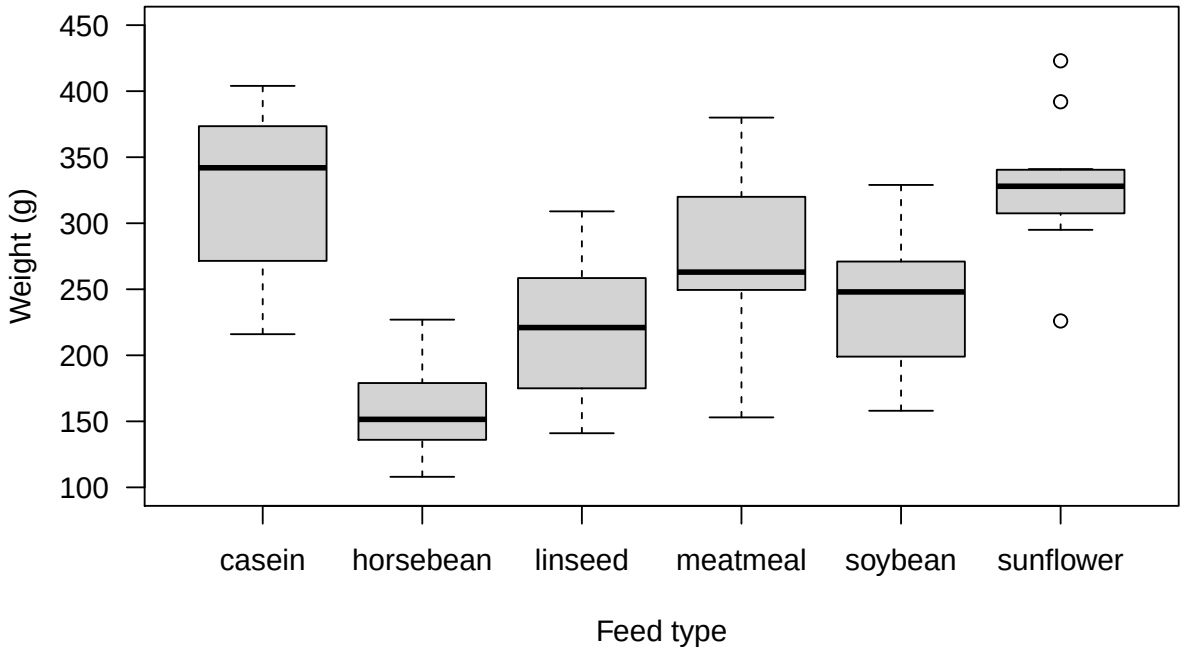


Figure 19.3: Chicken weight distributions for six different feed types.

Personally, I would be pretty comfortable with using an ANOVA test, but let's say that because we only have between ten and 14 observations per group we decide to use the non-parametric Kruskal-Wallis test rather than an ANOVA test.

There are no distributional assumptions with theKruskal-Wallis test, so we do not need the Bartlett test of constant group variance. Also, the test is not a test of differences in the group means, but rather is a test for some general kind of difference. The Null and Alternative hypothesis we use are therefore different to the Null and Alternative hypothesis used for the ANOVA test.

We use the function *kruskal.test()* to implement the Kruskal-Wallis test, and so for our formal test we have:

Step 1: Set the Null and Alternate Hypotheses

- Null hypothesis: The feeds all have the same effect on chicken weight gain
- Alternate hypothesis: The feeds do not all have the same effect on chicken weight gain

Step 2: Implement Kruskal-Wallis Test

When there is a tie in the rank values R will print a warning. This can be really annoying. The issue is trivial. In my experience the issue is at the 4th decimal place, and so I think it safe to ignore these warnings. For reporting purposes we truncate our p-value reporting at p-value < 0.001, so there is never anything misleading in the way we report results.

Unfortunately there is no easy way to turn off the warnings within the *kruskal.test()* function, so for this test we just have to live with the warnings when they apply.

```
with(chickwts, kruskal.test(weight ~ feed))

##
##  Kruskal-Wallis rank sum test
##
## data:  weight by feed
## Kruskal-Wallis chi-squared = 37.343, df = 5, p-value = 5.113e-07
```

Step 3: Interpret the Results
From the output we see the `p-value = 5.113e-07`; which we report as < 0.001. As the test p-value is smaller than 0.05 we:

- **Reject** the null hypothesis that the feeds all have the same effect on chicken weight gain

In our output we can also see that the χ^2-statistic is (`37.34`). Technically, this is our test statistic and it is this value that our test p-value is based on. For decision making it is much easier to just focus on the p-value decision rule rather than working with test statistics, and that is what we do here.

Similar to the way we proceeded with the ANOVA example, given we reject the null, we now move on to try to identify the specific differences using pairwise comparisons.

Pairwise Comparisons

Recall the research question: which feed(s) should the farmer consider using if she wants heavier chickens. Here we will use the function *pairwise.wilcox.test()* to compare all feed types. Remember, for these pair-wise tests we are not looking for differences in the group means but just some general difference in the groups.

To implement the tests we use the *pairwise.wilcox.test()* function. The function works in a similar way to the the *pairwise.t.test()* function. Specifically, it calculates all the different possible combinations for us at once. This means we don't need to type the *wilcox.test()* command 15 times (one for each pairwise combination). The *pairwise.wilcox.test()* also automatically adjusts the reported p-values to control the Type I error rate, which is important when conducting multiple comparisons.

The automatic adjustment to the p-values makes life easy for us as we can still apply our standard p-value decision rule to the results, where we know that the values have been adjusted to address the multiple comparison issue. This is something that only specialist software such as **R** is able to do, and it really saves a lot of time compared to the alternative of working through the adjustments manually. In all our testing we will rely on the default adjustment used by R, which is the holm adjustment.

Similar to when using the *wilcox.test()* function, when using the *pairwise.wilcox.test()* function it is possible to turn off the warnings by setting the parameter `exact = FALSE`. With pair-wise tests, it is possible to get a lot of warnings if you do not turn them off.

```
# Note i turn off the warnings using exact =FALSE

with(chickwts, pairwise.wilcox.test(weight, feed, exact = FALSE))

##
##  Pairwise comparisons using Wilcoxon rank sum test
##
## data:  weight and feed
##
##           casein horsebean linseed meatmeal soybean
## horsebean 0.0027 -         -       -        -
## linseed   0.0122 0.0644    -       -        -
## meatmeal  0.3622 0.0091    0.2181  -        -
## soybean   0.0474 0.0091    0.7100  0.7100   -
## sunflower 1.0000 0.0017    0.0032  0.3472   0.0140
##
## P value adjustment method: holm
```

As you probably suspected based on the boxplot, we have some significant differences between groups (e.g. horse bean and casein) and some non-significant differences (e.g meat meal and linseed). With reference to the boxplot, it can be seen that the two feeds that seem to have the highest distribution (i.e. they are the groups that feature at the top of the boxplot) are casein and sunflower. The test indicates no statistically significant difference for these groups. We also see meat meal, is also not significantly different from either casein or sunflower. So, it looks like the farmer has some options. She might now look to other factors such as cost or availability to decide what feed to provide her chickens. It is this final step that is important. The statistical test, whether it is a parametric test or a non-parametric test, is just one part of the process of working out what to do. You then typically need to look at other information – the boxplot and or the group means – to workout what you need to do.

Reporting Test Results

How you present numerical results in a table is just as important as how you present prepare figures. You must include information in a clear and concise way, without causing distraction with unnecessary information. Table 2 presents the results of our pair-wise comparisons, based on the non-parametric tests. For comparison purposes the results of the pair-wise t-tests are also shown in Table 3. Generally the p-values for the pair-wise t-tests (Table 3) are smaller.

Table 19.2: Pairwise Mann-Whitney test results: effect of feed on chicken weights

	Casein	Horsebean	Linseed	Meatmeal	Soybean
Horsebean	0.003	-	-	-	-
Linseed	0.012	0.064	-	-	-
Meatmeal	0.362	0.009	0.218	-	-
Soybean	0.047	0.009	0.910	0.710	-
Sunflower	1.000	0.002	0.003	0.347	0.014

Note: Holm adjusted p-values

Table 19.3: Pairwise t-test results: effect of feed on chicken weights

	Casein	Horsebean	Linseed	Meatmeal	Soybean
Horsebean	< 0.001	-	-	-	-
Linseed	< 0.001	0.094	-	-	-
Meatmeal	0.182	< 0.001	0.094	-	-
Soybean	0.005	0.003	0.518	0.518	-
Sunflower	0.812	< 0.001	< 0.001	0.132	0.003

Note: Holm adjusted p-values

19.4 Non-parametric Regression

Non-parametric methods are also used in regression. So far we have seen regression models defined parametrically: for example a linear regression model is defined by the slope and the intercept parameters. Rather than assume a linear relationship, or a log-linear relationship, or any other parametric relationship between two variables, we can relax this assumption and have the form of the relationship be defined by the observed data. Such methods are robust to model mis-specification (e.g. assuming a linear relationship when in fact the relationship is of some other form). These methods are covered in more advanced courses.