Contents

These course notes were written by Dr. Robin Donatello for MATH 315, Applied Statistics I at California State University, Chico, and updated on 2019-01-16 for the Spring 2019 semester.

I would like to thank Jack Foglassio (undergraduate student at CSU, Chico) for his tremedous help in revising these course notes over Winter break 2019.

Introduction

According to the 2006 Oxford Dictionary of Statistical Terms, statistics is the study of the collection, organization, analysis, interpretation and presentation of data. Some refer to it as "the science of learning from 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**. Training in Statistics is necessary to understand how to collect, visualize, describe, analyze, draw conclusions from, and understand the level of uncertainty and limitations in the findings.

An underlying concept is Data Literacy: how to understand, visualize, read and learn from data. For the critical role it plays in Science and Statistics, data literacy has been historically underemphasized. Right now data is being generated and collected on everything and everyone, and the amount and speed of this growth is faster than ever before. Data is everywhere, and if you don't know how to use it you are at a distinct disadvantage.

In the current world that we are living in, where we are being bombarded with "fake news" and "misleading reports", the ability to determine which content is valuable and meaningful is paramount to ensure the health and well-being of our everyday lives. From a scientific researcher's point of view, there's nothing worse than spending time and money on a study, only to find out that you can't use the data to answer your research question.
Statistics falls into two major domains: **Descriptive** and **Inferential** statistics.

Descriptive: A collection of methods to organize, describe, visualize and summarize the data. It answers the question, "What information does the data contain and what relationships exist?" These methods are the basis of an Exploratory Data Analysis (EDA).

Inferential: A collection of methods to answer questions and hypotheses about the data, to make inferences about a **population** based upon data obtained from a **sample** of that population.
We will spend the first half of the class understanding how to explore, describe and investigate data. Then we can move towards using the data to make decisions.

Structure of Notes: There are two types of text boxes and two different fonts used in these notes.

Reading Assignment

Text in red blocks tell you which Open Intro Book sections correspond to the current chapter. These sections are to be read prior to class. Quiz material can come from the readings.

Example 0.1: Example

Blue boxes contain questions or example problems where you write your answer directly into the notes. You can work these out individually or with a peer. These will be used as door tickets, and sometimes you will be asked to present your answers to the class to start a discussion.

Bold and underlined words are **vocabulary terms.**

`R` code is noted in two ways. When referring to code within a paragraph of text, `this font is reserved for R code and/or special terms in R`. Often, however, you will see syntax-highlighted `R` code such as the two lines below. This is the direct code used to perform a task. The output follows the command immediately, and is prefaced with two pound signs.

```
2+2

## [1] 4

mean(rnorm(100))

## [1] -0.04586245
```

Data used in this class: Where possible, all data and examples used in this class come from real situations. All data will be available for download and should be downloaded prior to class time. Syntax to read in the data into `R` is provided with the download link. Detailed information on the data being analyzed will be made available where possible.

Chapter 1

Introduction to Data

Reading Assignment

OpenIntro Section 1.3

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

1.1 Populations and samples

Consider the following three research questions:

1. What is the average mercury content in swordfish in the Atlantic Ocean?
2. Over the last five years, what is the average time taken to complete a degree for Chico State 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. Oftentimes, 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 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.

Example 1.1: Identifying the Population

For each question above, identify the target population and what represents an individual case.

Consider the following possible responses to the three research questions (RQ):

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

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

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

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

1.1.1 Sampling from a population

Reading Assignment

OpenIntro (OI) Section 1.3.3

We might try to estimate the time-to-graduation for Chico State undergraduates in the last five years by collecting a sample of students. All graduates in the last five 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.

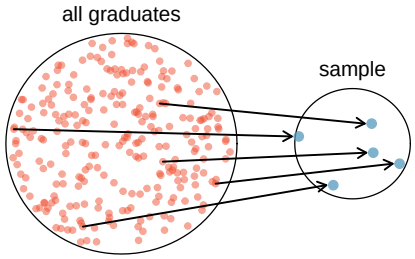


Figure 1.1: In this graphic, five graduates are randomly selected from the population to be included in the sample.

Example 1.2: Choosing a representative sample

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?

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

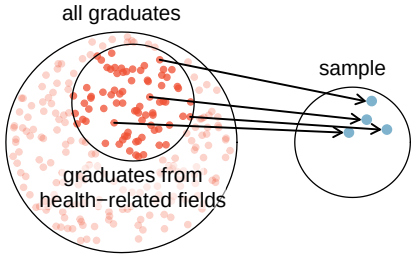


Figure 1.2: Instead of sampling from all graduates equally, a nutrition major might inadvertently pick graduates with health-related majors disproportionately often.

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

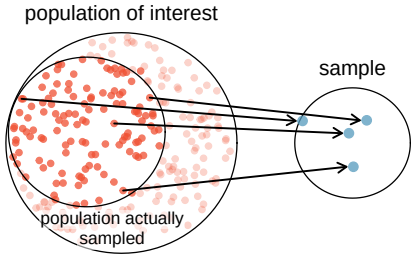


Figure 1.3: Due to the possibility of non-response, surveys studies may only reach a certain group within the population. It is difficult, and often impossible, to completely fix this problem.

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

Example 1.3: Online rating systems

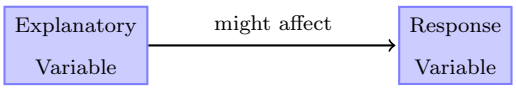
We can easily access ratings for products, sellers and companies through websites. These ratings are based only on those people who go out of their way to provide a rating. If 50% of online reviews for a product are negative, do you think this means that 50% of buyers are dissatisfied with the product?

1.2 Explanatory and Response variables

In studying the relationship between two variables, the variables are often viewed as either a **response variable** or an **explanatory variable**.

- The response variable is the variable or characteristic of the data that we want to learn something about.
- The explanatory variable is the variable whose effect on the response variable we are interested in establishing.

To identify the explanatory variable in a pair of variables, ask yourself which of the two is suspected of affecting the other.



Example 1.4: Identifying the parts of a relationship

Consider the following questions or pairs of variables and identify the response and explanatory variables.

1. Is federal spending, on average, higher or lower in counties with high rates of poverty?
2. College grade point average and high school grade point average.

Labeling variables as *explanatory* and *response* does not guarantee the relationship between the two is actually causal, even if there is an association identified between the two variables. We use these labels only to keep track of which variable we suspect affects the other.

1.3 Data Structure

Reading Assignment

OpenIntro (OI) Section 1.2

Effective presentation and description of data is a first step in most analyses. You can collect the best quality data in the world, but if you can't make sense of it then it's useless bits and bytes – and a waste of time and energy!

This section introduces the most common structure for organizing data, and some terminology that will be used throughout this book.

1.3.1 Observations, variables, and data matrices

	spam	num_char	line_breaks	format	number
1	0	11.3700	202	1	big
2	0	10.5040	202	1	small
3	0	7.7730	192	1	small
50	0	14.4310	296	1	small

Table 1.1: Four rows of data from the email data set

Table 1.1 displays rows 1, 2, 3, and 50 of the `email` data set. This data set contains information on emails received to one of the authors of the Open Intro textbook during early 2012, and was used to identify characteristics of spam mail.

The data in Table 1.1 represent a **data matrix**, which is a common way to organize data. It can also be referred to as **structured data**. 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.

Each row in the table represents a single email or **case** (**observation**). The columns represent characteristics, called **variables**, for each of the emails. For example, the first row represents email 1, which is 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 1.2. This is an example of a **codebook**, or "data dictionary".

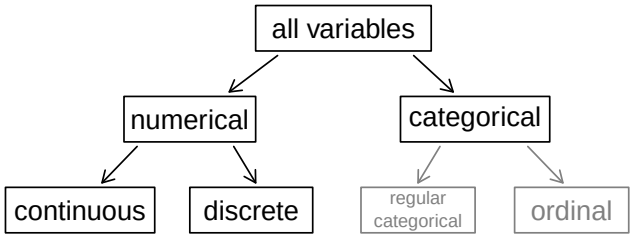
variable	description
`spam`	Specifies whether the message was spam
`num_char`	The number of characters in the email
`line_breaks`	The number of line breaks in the email (not including text wrapping)
`format`	Indicates if the email contained special formatting, such as bolding, tables, or links, which 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 1.2: Variables and their descriptions for the `email` data set.

1.3.2 Types of Data

When dealing with data there are two main types of data to consider: those that can be measured, and those that cannot.

- **Categorical** data is data that can be categorized, or grouped, by non-overlapping characteristics. **Cannot be measured (like with a ruler).**
- **Numerical** data is quantitative data that can take on a value on the number line, and where mathematical operations can be performed on it. **Can be measured (like, with a ruler).**



Example 1.5: County data

We consider a publicly available data set that summarizes information about the 3,143 counties in the United States. It is available for download on the course website as the `OScounty` data set. This data set includes information about each county: its name, the state where it resides, its population in 2000 and 2010, per capita federal spending, poverty rate and other characteristics.

state	name	fed_spend10	pop2010	smoking_ban
Alabama	Autauga County	6.0681	54571	none
Alabama	Baldwin County	6.1399	182265	none
Alabama	Barbour County	8.7522	27457	partial
Alabama	Bibb County	7.1220	22915	none
Alabama	Blount County	5.1309	57322	none
Alabama	Bullock County	9.9731	10914	none

Table 1.3: Sample of data from the county dataset

Q: Specify the data type (with units where appropriate) for each variable in the data matrix above.

- state:
- name:
- fed_spend10:
- pop2010:
- smoking_ban:

Each of these variables is inherently different from the other three, yet many of them share certain characteristics. First consider `fed_spend10`, which is said to be a **numerical** variable since it can take on 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 sums, differences and averages have no clear meaning.

The `pop2010` variable is also numerical, although it seems to be a little different than `fed_spend10`. This variable of the population count can only take whole numbers, i.e., non-negative integers $(0, 1, 2, \ldots)$. 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 federal spending variable is said to be **continuous**.

The variable `state` can take up to 51 values after accounting for Washington, DC. Because the responses themselves are categories, `state` is called a **categorical** variable (sometimes also called a **nominal** variable), and the possible values are called the variable's **levels**. Categorical variables that have only two levels (e.g., M/F) can also called **binary** variables.

Finally, consider the `smoking_ban` variable, which describes the type of county-wide smoking ban and takes values `none`, `partial` or `comprehensive` in each county. This variable seems to be a hybrid: it is a categorical variable but the levels have a natural ordering. A variable with these properties is called an **ordinal** variable. To simplify analyses, any ordinal variables in this class will be treated as categorical variables.

1.4 Basic Data Management

This section contains small excerpts from the Chapter on Data Preparation in the Applied Statistics Notebook `https://norcalbiostat.github.io/AppliedStatistics_notes/data-prep.html` (also written by your instructor). For more context or assistance follow the link provided here or through the course website to access this online notebook. The data in this section come from several sources including data on depression from the Afifi et.al. textbook and data on live births from North Carolina in 2014. These data sets can be found on the **Data** page of Dr. D's website.

1.4.1 Missing Data

R displays missing data as `NA` values. Use `table()` with the `useNA="always"` argument to identify missing categorical data and `summary()` for quantitative data.

```
summary(ncbirths$fage)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   14.00   25.00   30.00   30.26   35.00   55.00     171

table(ncbirths$habit, useNA="always")

##
## nonsmoker    smoker      <NA>
##       873       126         1
```

Change data values to missing codes Let's look at the religion variable in the depression data set.

```
table(depress$RELIG, useNA="always")

##
##    1    2    3    4    6 <NA>
##  155   51   30   56    2    0
```

The codebook for the depression data set shows that there is no category '6' for religion. Thus all 6's we see in the data set are errors. We need to change all values from 6 to `NA`

(missing). Then we recreate the table to visually confirm that our recode worked as intended.

```
depress$RELIG[depress$RELIG==6] <- NA
table(depress$RELIG, useNA="always")

##
##    1    2    3    4 <NA>
##  155   51   30   56    2
```

The results of the table now show 2 `NA` values and no values with a '6'. This code says take all rows where `RELIG` is equal to 6, and change them to `NA`. This technique works for all data types.

1.4.2 Creating factor variable from numeric

When variables have numerical levels it is necessary to ensure that the program knows it is a factor variable.

The following code uses the `factor()` function to take the marital status variable and convert it into a factor variable with specified labels that match the codebook. The ordering of the labels argument must be in the same order (left to right) as the factor levels themselves.

```
depress$marital_cat <- factor(depress$MARITAL,
  labels = c("Never Married", "Married", "Divorced", "Separated", "Widowed"))
table(depress$MARITAL, depress$marital_cat, useNA="always")

##
##        Never Married Married Divorced Separated Widowed <NA>
##   1               73       0        0         0       0    0
##   2                0     127        0         0       0    0
##   3                0       0       43         0       0    0
##   4                0       0        0        13       0    0
##   5                0       0        0         0      38    0
##   <NA>             0       0        0         0       0    0
```

It is important to confirm the recode worked by creating a frequency table of the old variable `MARITAL` against the new variable `marital_cat` to visually confirm that all 1's went to "Never Married" etc.

1.4.3 Collapsing factor levels

Lets combine the categories for Divorced and Separated. Here I create a new variable called `marital_4cat` (4 category marital variable). This example uses the function `fct_collapse` from the `forcats` package. This is not the only way to accomplish this, but it is the easiest. Notice that the new category name is **not** in quotes, but the list of levels that are being collapsed are in quotes.

```
depress$marital_4cat <- forcats::fct_collapse(depress$marital_cat,
                              Div_Sep = c("Divorced", "Separated"))
table(depress$marital_cat, depress$marital_4cat, useNA="always")

##
##                 Never Married Married Div_Sep Widowed <NA>
##   Never Married            73       0       0       0    0
##   Married                   0     127       0       0    0
##   Divorced                  0       0      43       0    0
##   Separated                 0       0      13       0    0
##   Widowed                   0       0       0      38    0
##   <NA>                      0       0       0       0    0
```

Again I create a two-way frequency table to confirm that all records that were recorded as either divorced or separated are now listed as "Div_Sep".

1.4.4 Creating binary variables

The `ifelse()` is hands down the easiest way to create a binary variable (dichotomizing, only 2 levels). Let's add a variable to identify if a mother in the North Carolina births data set was underage at the time of birth.

Make a new variable `underage` on the `NCbirths` data set. If `mage` is under 18, then the value of this new variable is `underage`, else it is labeled as `adult`.

```
ncbirths$underage <- ifelse(ncbirths$mage <= 18, "underage", "adult")
table(ncbirths$underage, useNA="always")

##
##    adult underage     <NA>
##      925       75        0
```

Since this is a quantitative variable, we can calculate summary statistics to make sure the age ranges are correctly coded. Here I use `dplyr` code to take the data set `ncbirths`, then group it by the categorical variable `underage`, then calculate the min and max of `mage` using `summarise`. For more information on how to use `dplyr` to calculate summary statistics refer to the `dplyr` cheat sheet found in R Studio or online via Google search.

```
ncbirths %>% group_by(underage) %>% summarise(min=min(mage),
                                              max=max(mage))

## # A tibble: 2 x 3
##   underage   min   max
##   <chr>    <dbl> <dbl>
## 1 adult       19    50
## 2 underage    13    18
```

Ages for individuals in the `underage` category range from 13 to 18, and ages for those in the `adult` category are all over 18. This recode was successfully conducted.

Chapter 2

Visualizing and Describing Distributions of Data

The first step to understanding relationships in your data is to look at each variable of interest. We could look at the **raw data** such as what appears in Table 1.1 and Table 1.3, but that is hard to ingest. Instead we explore the data by creating summary statistics, tables and charts/graphics/plots. Visualizing your data is a critical step in exploratory analysis and must not be ignored.

The order this section is presented is the order in which you should conduct any exploratory data analysis. Individual variables need to be looked at by themselves first (univariately), **before** relationships between two or more variables are examined.
Describing distributions of data consist of three things:

- A visualization (plot, graph, graphic)
- Summary statistics (mean, median, N, %)
- A sentence connecting the summary numbers and the plot, in a readable manner.

Plots should never stand alone. They should always be accompanied by a sentence that describes the features of the plot, along with the summary numbers to put the data in context (i.e. units).

2.1 Describing Univariate Numerical data

Reading Assignment

OpenIntro Section 1.6.2 - 1.6.6

In this section we will be introduced to techniques for exploring and summarizing numerical variables. Recall that outcomes of numerical variables are numbers on which it is reasonable to perform basic arithmetic operations.

There are three primary pieces of information that you want to be able to discuss when describing the distribution of a numerical variable.

- Location: What is the typical value? In what range does most of the data lie?
- Shape: What is the shape of the data? Is it symmetrical? skewed? Unimodal?
- Spread: How spread out is the data? What is the range of plausible values? Max and min?

We'll start out by exploring how to calculate **summary statistics**, or numbers that we can use to characterize the data in an attempt to answer the three questions above. Summary statistics include measures of center such as the mean, median and mode, and measures of spread such as the range, variance and standard deviation.

2.1.1 Summary statistics

Measures of center

- Mean: Sometimes called the **average**, is a common way to measure the center of a **distribution** of data. Let x_i be the value of a single data point for any $i = 1, \ldots, n$. The sample mean $\bar{x}$ is calculated as the sum of all x_i divided by the sample size.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.1}$$

- Median: The physical midpoint of the distribution. Half the observations are below this point, and half are above.
- Mode: This is the value of the data that occurs most frequently. Less commonly used than the mean or median.

Example 2.1: Calculating measures of central tendency on exam scores

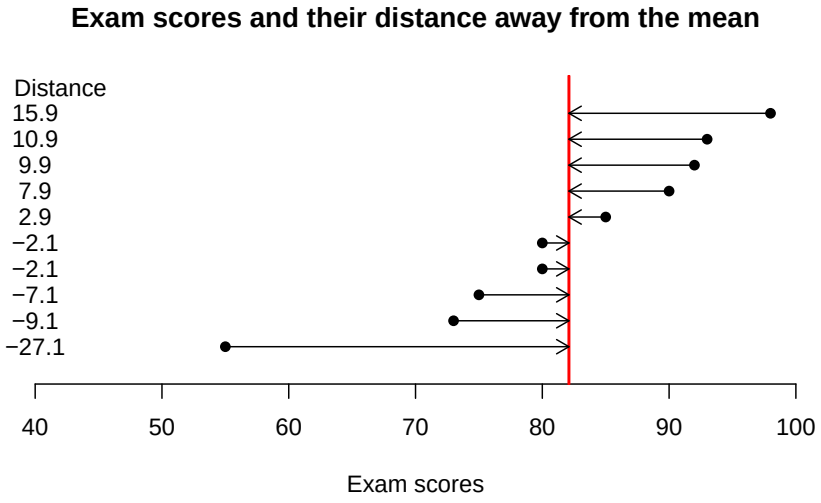
The scores from a recent exam are as follows: 55, 73, 75, 80, 80, 85, 90, 92, 93, 98
Calculate the Mean and Median for the exam scores.

- mean:

- median:

Measures of Spread

- Range: Simply the maximum value minus the minimum value.
- Variance: A measure of how far away the data points are from their mean.
- Standard Deviation: Calculated as $\sqrt{s^2} = s$, the standard deviation is used more in context than the variance, primarily because it is on the same measurement unit scale as the data.



The variance s^2 is almost the average of these distances $(x_i - \bar{x})$ squared.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.2}$$

Q: Why do we square this distance before we sum across data points? *Hint: Add up the distances without squaring. Does this number make sense?*

Example 2.2: Describing the distribution of exam scores

In a complete English sentence using the summary statistics above as supporting evidence, describe the distribution of exam scores.

Calculate summary statistics using R

When you have a small amount of data you can use the collection operator `c()` to create a vector of data and then perform calculations on that vector.

```
a <- c(55, 73, 75, 80, 80, 85, 90, 92, 93, 98)
mean(a)

## [1] 82.1

median(a)

## [1] 82.5

range(a)

## [1] 55 98

var(a) # variance

## [1] 157.4333

sqrt(var(a)) # square root of the variance == standard deviation

## [1] 12.54724

sd(a) # standard deviation

## [1] 12.54724
```

2.1.2 Dotplots, Histograms and Density plots

Here we consider the percent of people in each county living below the poverty level from the `county complete` data set. A sample of the raw data is in the table below.

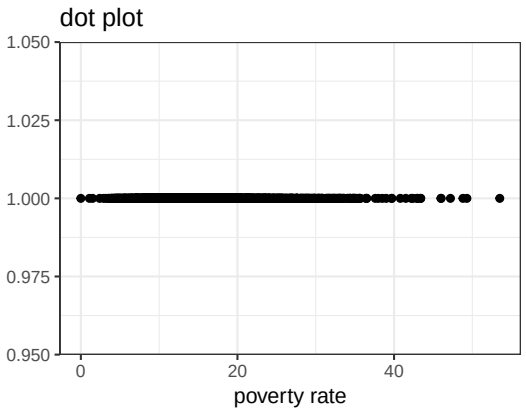
name	poverty
Autauga County	10.60
Monroe County	25.40
Pima County	16.40
Montrose County	10.90

Q: What data type is `poverty`? What are its units? Where did you find this information?

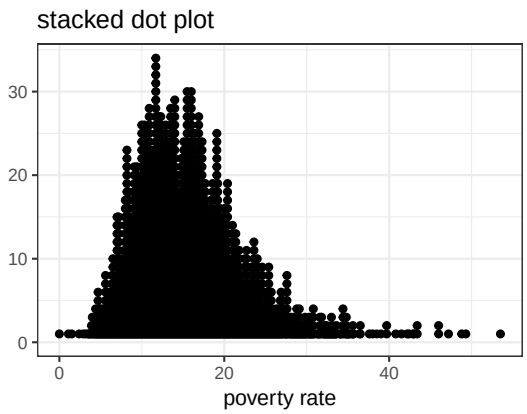
The plot on the left is a standard **dot plot**, where each dot is placed on the x-axis of the graph at the value of the data point and the same value of y, *the value of which is irrelevant here*. For many data sets, dot plots contain a lot of over-plotting and this makes them hard to read. How many counties have between 10% and 20% poverty? Hard to tell.

The plot on the right starts to make the picture clearer by plotting each dot with the same x value higher on the y scale. This **stacked dot plot** has a meaningful vertical value, namely the number of observations that have that exact value of x.

```
ggplot(county, aes(y=1, x=poverty)) +
  ylab("") + xlab("poverty rate") +
  geom_point() + ggtitle("dot plot")
```




```
ggplot(county, aes(y=1, x=poverty)) +
  ylab("") + xlab("poverty rate") +
  geom_point(position=position_stack()) +
  ggtitle("stacked dot plot")
```



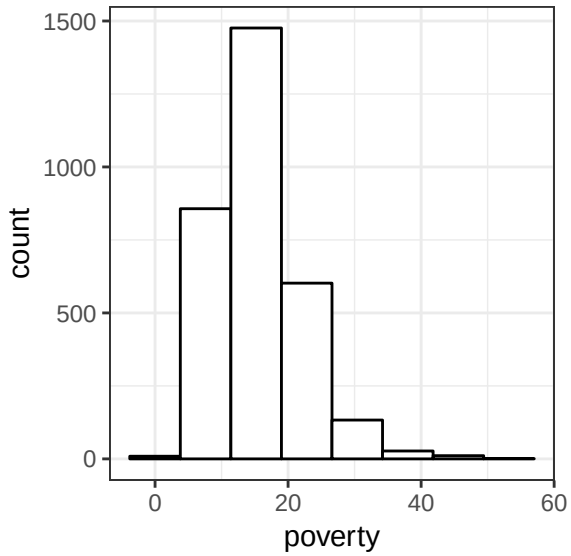
Histograms

Rather than showing the value of each observation, we prefer to think of the value as belonging to a *bin*. **Histograms** display the frequency of values that fall into those bins. For example if we cut the poverty rates into seven bins of equal width, the frequency table would look like this:

(-0.0535,7.64]	(7.64,15.3]	(15.3,22.9]	(22.9,30.6]	(30.6,38.2]	(38.2,45.9]	(45.9,53.6]
216	1442	1101	276	63	12	6

In a histogram, the binned counts are plotted as bars into a histogram. Note that the x-axis is continuous, so the bars touch. This is unlike the barchart, which has a categorical x-axis and vertical bars that are separated.

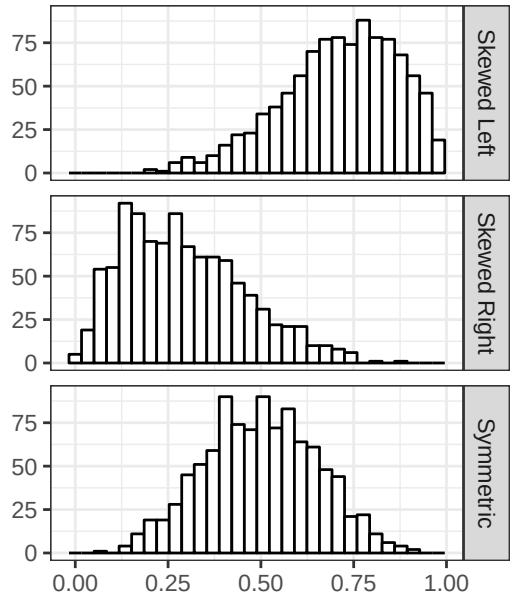
```
ggplot(county, aes(x=poverty)) +
  geom_histogram(binwidth=7.6, colour="black", fill="white")
```



Code Comments: The binwidth is set by looking at the cut points above that were used to create seven bins. The fill was set to white so that the outlines of the bars could be seen – darkgrey is the default.

The shape of the data distribution is considered **skewed right** if the histogram has a long right tail, and **skewed left** if the histogram has a long left tail. We generally hope to see a **symmetric** distribution to the data.

The size of the bins can reveal or hide attributes of the distribution (Figure 2.1).



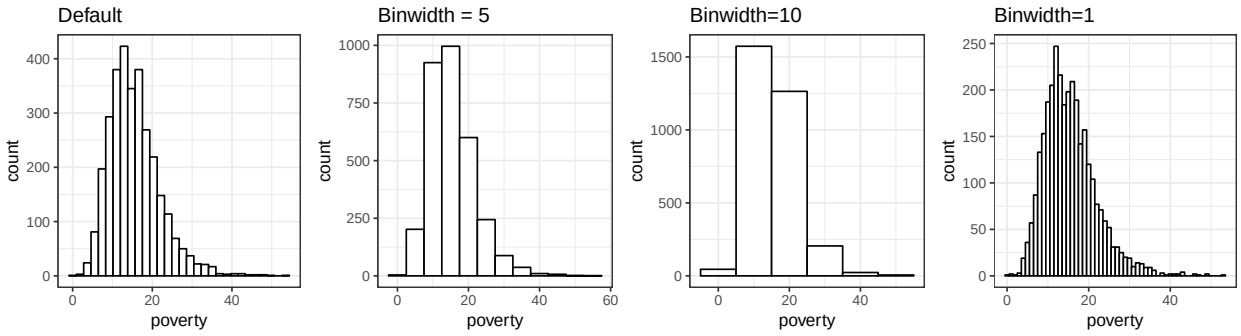


Figure 2.1: Histograms of poverty rate with varying number of bins.

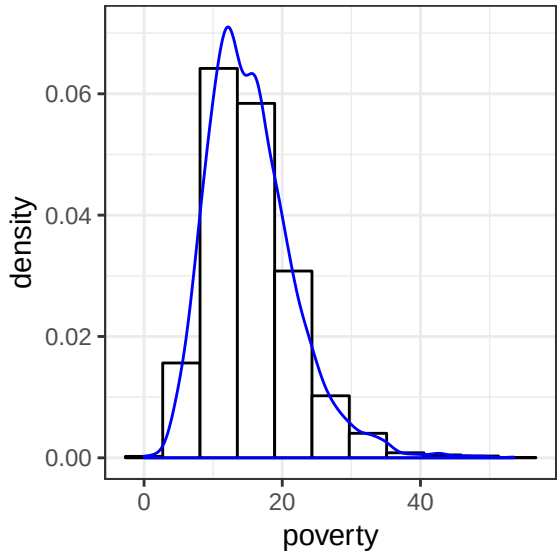
Example 2.3: Describing the distribution of poverty

In a complete English sentence using the following summary statistics as supporting evidence, describe the distribution of the county-level poverty rate. Don't forget to include the units.
mean: 15.5%, range: 0 - 53.5%, median: 14.7%, sd: 6.7%

Density plots

Notice that as the binwidth decreases, the shape of the histogram "smooths" out? If you let the width of the bins get infinitely small, you end up with a density curve.

```
ggplot(county, aes(x=poverty)) +
  geom_histogram(aes(y=..density..), binwidth=5.4, colour="black", fill="white") +
  geom_density(col="blue")
```



Notice that this density curve is a little "wigglier" than the histogram indicates. Most of the time density curves can give you a much better picture of the actual data distribution than histograms can. You always lose information when aggregating data like we did when we collapsed the real values of poverty rate into counts of rates within a specified bin. We will get more into probability densities in the next chapter, but this is a good spot to introduce it.

Code comment: Histograms plot frequencies, so the y-axis is a count of values. The area under a density curve must add up to 1, so the scale on the y-axis is different. In the `geom_histogram` *line you have to specify that the y value is* `..density..` *so that it knows to scale the y-axis correctly. Otherwise the blue line would not be visible.*

2.1.3 Boxplots and the five number summary

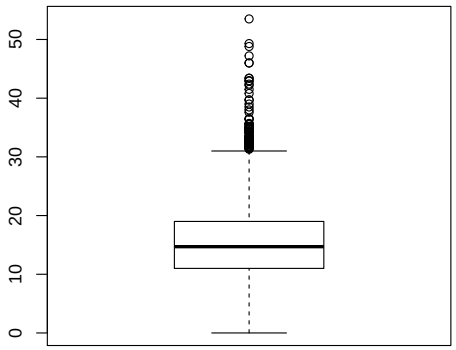
Another name for the median is the 50^{th} percentile (P_{50}), since 50% of the data is below this number. In general, the p^{th} percentile is the value where p% of the data points fall below that number. Other names: the 25^{th} percentile (P_{25}) is also called the first quartile (Q_1), and the 75^{th} percentile (P_{75}) is also known as the third quartile (Q_3). The **five number summary** is the minimum, Q_1, median, Q_3, and maximum values of a data set. The **interquartile range** (IQR) is the distance between Q_1 and Q_3.

```
summary(county$poverty)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   11.00   14.70   15.53   19.00   53.50

IQR(county$poverty)

## [1] 8
```

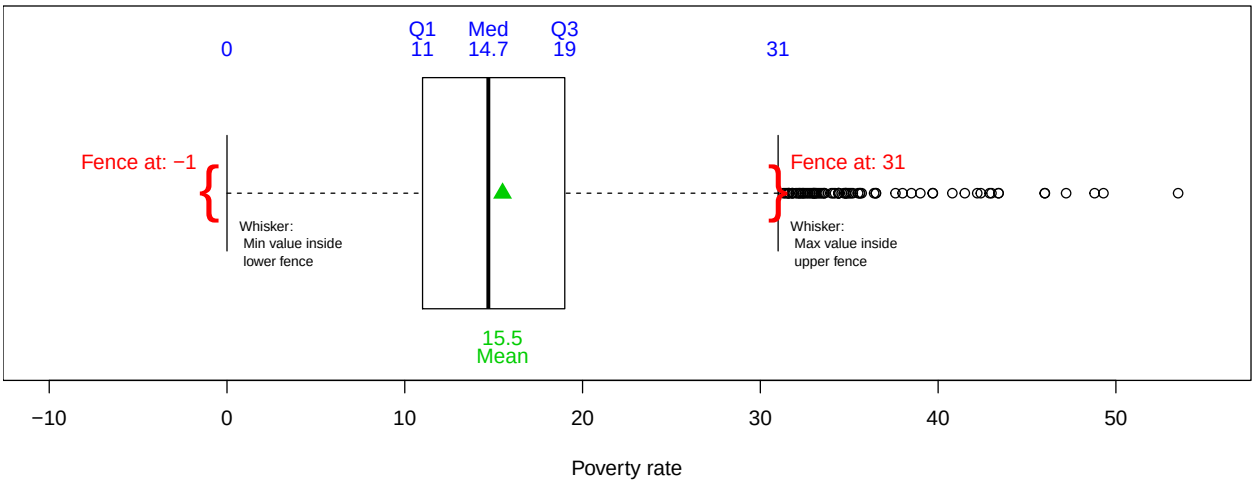
These five numbers are used to create a **boxplot**.

A basic univariate boxplot looks like this. The plot is split into quarters, with half of the data in the two sections that make up the box, and half the data in the two sections that make up the whiskers.

```
boxplot(county$poverty)
```



Most statistical programs draw a **modified boxplot**, where suspected outliers are drawn as open circles (or dots). Values are considered outliers if they are below $Q_1 - 1.5 * IQR$ or above $Q_3 + 1.5 * IQR$; these are called the **fences**. In the presence of outliers, the whisker is then drawn at the largest data point still within the $1.5 * IQR$ range.



The mean is **sensitive** to outliers, meaning a high-end outlier will pull the mean up and a low-end outlier will pull the mean down. The median is not sensitive to outliers since it is defined as the middle number, regardless of the value of the data.

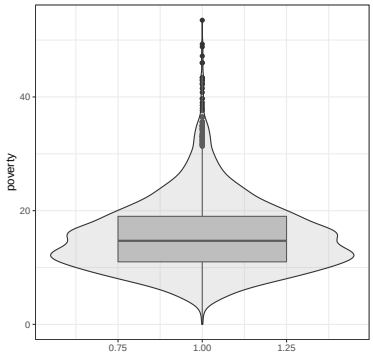
Q: Use the output from `summary()` to confirm the location fences on the plot above are correct.

2.1.4 Violin plots: where density meets boxplots

A fantastic addition to boxplots is an overlaid density curve. For aesthetic purposes, since a boxplot is symmetric about a midline down the center (connect the two whiskers), the density curve is reflected on both sides of this midline. The interpretation is the same as a density plot seen earlier: the wider the area under the curve, the more data lies in that area. Here I have adjusted the width of the boxplot, added a fill color and changed the transparency of the violin plot to be more readable.

.

```
ggplot(county, aes(x=1, y=poverty)) +
  geom_boxplot(width=.5, fill="grey") +
  geom_violin(alpha=.3, fill="grey") +
  xlab=("")
```



Example 2.4: Effect of outliers

Recall the exam grade data on page 22. What if the person recording those grades erroneously entered in the first score of 55 as 5? How does this affect the measures of center and spread?

- Mean:
- Median:
- Variance:
- Standard Deviation:

2.2 Describing Univariate Categorical data

Reading Assignment

OpenIntro Section 1.7

We will be using the full `email` data set that contains information on 3,921 emails. Two categorical variables of current interest are `spam` (0/1 binary indicator if an email is flagged as spam) and `number` which describes whether an email contains no numbers (`none`), only `small` numbers (values under 1 million), or at least one `big` number (a value of 1 million or more).

2.2.1 Tables

The distribution of a categorical variable is summarized using a **frequency table**. These are created using the `table()` function, where the argument is the categorical variable that you want to create the table for.

```
table(email$spam)
```

```
##
##    0    1
## 3554  367
```

The values displayed represent the number of records in the data set that fall into that particular category; e.g. 367 records in the `email` data set were flagged as spam.

2.2.2 Proportions

A summary statistic for categorical data is the **sample proportion**. This is calculated as the number of times a particular event (x) occurs divided by the total number of records (n).

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.3}$$

In a frequency table such as the one above, $\sum x_i$ is exactly what is displayed in the table entries, i.e., the total number of times an event occurs.

Example 2.5: Calculating proportions

Use the table below that contains the marginal total (Sum) to calculate the proportion of emails flagged as spam and not flagged as spam by dividing each frequency by the total.

```
addmargins(table(email$spam))

##
##    0    1  Sum
## 3554  367 3921
```

Q: What do these two proportions add up to? Will that always be the case?

A proportion is also called a **relative frequency**, that is, the frequency of an event occurring relative to the amount of non-events. The function `prop.table()` creates this relative frequency table when it is provided a table object as an argument.

```
prop.table(table(email$spam))

##
##          0          1
## 0.90640143 0.09359857
```

Proportion as the mean of a binary variable

You may have noticed that Equation (2.3) looks identical to Equation (4.1). This is what happens when the categorical variable of interest is coded as 0/1. This allows us to use the `mean()` function on a binary variable to calculate the proportion of events (when $x = 1$).

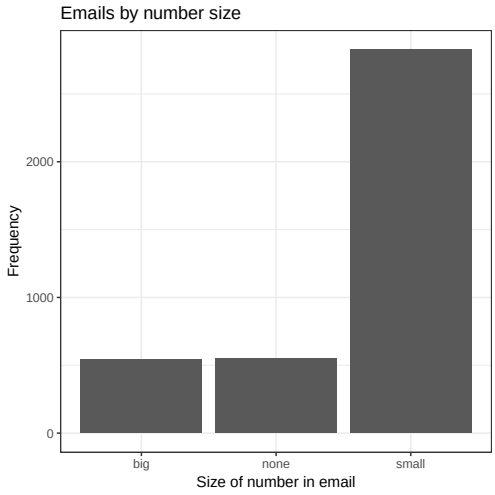
```
mean(email$spam)

## [1] 0.09359857
```

This is particularly advantageous because we will learn many methods for analyzing and comparing means of continuous variables. We can use many of those techniques on binary variables to compare proportions.

2.2.3 Bar Charts

The most common method to display frequencies in a graphical manner is with a **bar chart** (a.k.a. barplot or bar graph). It has one bar per distinct category with the height of the bar representing the frequency of the data that fall into that category.

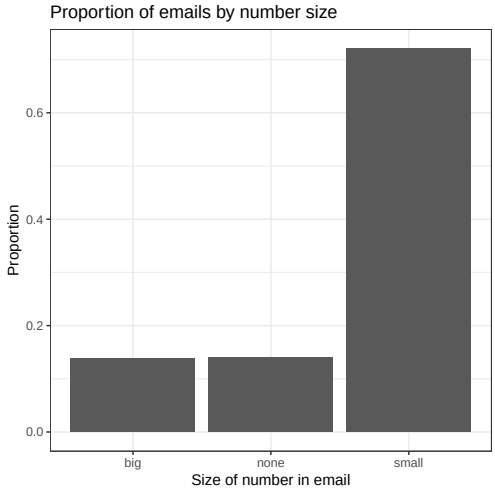
```
library(ggplot2)
ggplot(email, aes(x=number)) +
  geom_bar() +
  ylab("Frequency") +
  xlab("Size of number in email") +
  ggtitle("Emails by number size")
```



Q: What do the layers `ylab, xlab` and `ggtitle` do?

Now we just saw how comparing frequencies can be misleading, so let's redraw that bar chart where proportions are on the vertical y-axis. To accomplish this, we have to aggregate the data to calculate the proportions first, then plot the aggregated data using `geom_col` to create the columns.

```
a <- table(email$number) %>%
      prop.table() %>%  data.frame()

ggplot(a, aes(x=Var1, y=Freq)) +
  geom_col() +
  ylab("Proportion") +
  xlab("Size of number in email") +
  ggtitle("Proportion of emails by
          number size")
```



Q: What happens if you don't turn the `table` into a `data.frame`? *Hint: the first argument to* ***`ggplot()`*** *must be a data frame.*

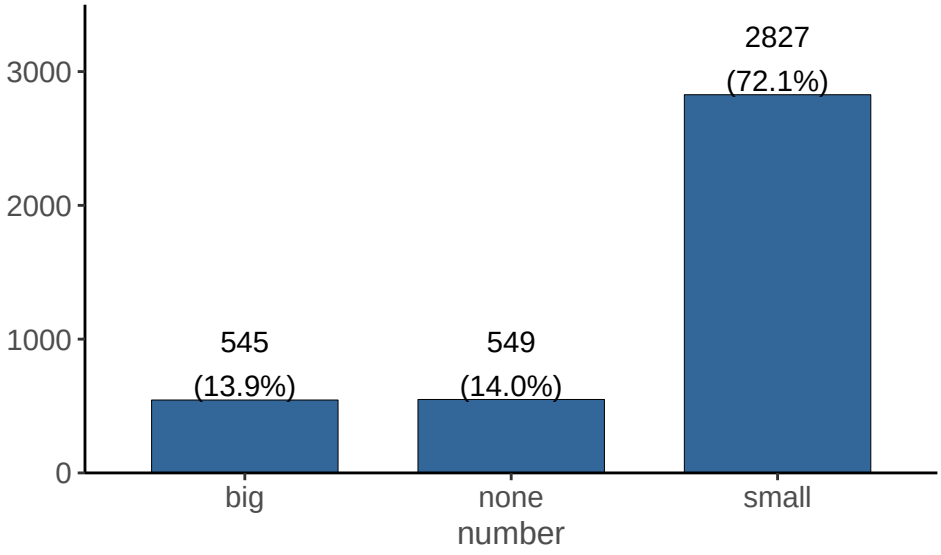
Q: Where did `Var1` and `Freq` come from? Why can't we just specify `number` on the x axis? *Hint: Look at the variable names of* ***`a`****.*

Q: What happens if you use `geom_bar` instead of `geom_col`? *Hint: Do this in R and look at the y-axis.*

Using the sjPlot package

The `sjPlot` package has some nice graphing and table alternatives to `ggplot2`. Here is one example. Don't forget you have to install the package once before you can call the `sjp.frq` function.

```
set_theme(base = theme_classic())
sjp.frq(email$number)
```

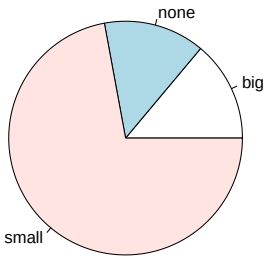


Nearly three-quarters (72%, 2,827) emails contain small numbers, with the rest split between having no numbers (n=549, 14%) or big numbers (n=545, 13.9%).

2.2.4 Pie Charts

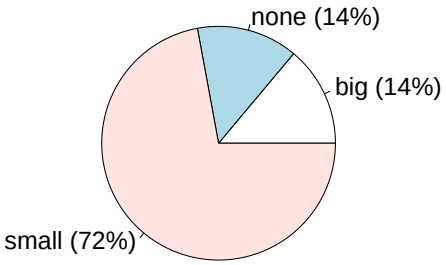
Pie charts are the most easily misused data visualization. I do not recommend them primarily because our eyes cannot compare angles very well. Which is larger, the `none` or `big` wedge? Nevertheless you can create them in R simply by using the `pie()` function.

```
pie(table(email$number))
```



If you want to use a pie chart then you *must* label the wedges with the proportions for each wedge. You can do this manually by using the `labels=` argument, but I like to do things programmatically (the hard, but reproducible way).

```
a <- table(email$number)
pie(a, labels=paste0(
        names(a),
        " (",
        round(a/sum(a),2)*100,
        "%)"))
```



2.3 Relationships between two variables

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

1. Is federal spending, on average, higher or lower in counties with high rates of poverty?
2. Which counties have a higher average income: those that enact one or more smoking bans or those that do not?

2.3.1 Categorical v. Categorical: Cross-tabs and Bar charts

Recall that you can examine the distribution of categorical variables in two ways: by looking at the frequency (count) of events, or the proportion of events.

Comparing frequencies

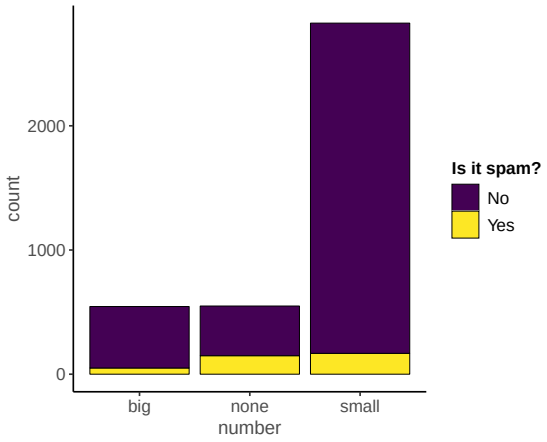
A table that summarizes data for two categorical variables is called a **contingency table** (or cross-tab). Each value in the table represents the number of times a particular combination of variable outcomes occurred. This is achieved by listing both variable names separated by a comma in the `table()` function.

```
table(email$spam, email$number)

##
##      big none small
##   0  495  400  2659
##   1   50  149   168
```

The value 149 corresponds to the number of emails in the data set that are spam *and* had no number listed in the email.

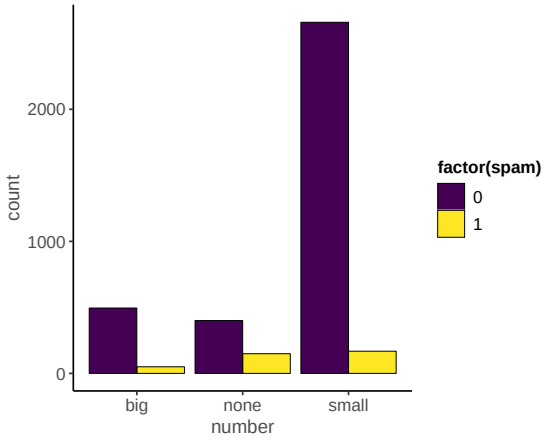
Bar charts are the standard method to compare frequencies across groups. The default method is to produce stacked bar charts. This is not typically advised as it is difficult to compare frequencies across categories.

```
ggplot(email, aes(x=number,
                  fill=factor(spam))) +
  geom_bar() +
  scale_fill_discrete(
        name="Is it spam?",
        breaks=c("0","1"),
        labels=c("No", "Yes"))
```

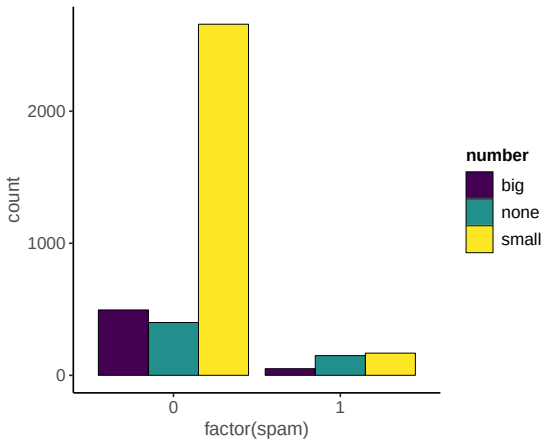


Side-by-side bar charts tend to be the best for comparing between categories. This is achieved by adding `position=position_dodge()` to the `geom_bar()` layer. You can easily change what is being compared by changing what variable is on the x-axis, and what variable is being used for the color fill.

```
ggplot(email, aes(x=number,
                  fill=factor(spam))) +
  geom_bar(position=position_dodge())
```




```
ggplot(email, aes(x=factor(spam),
                  fill=number) +
  geom_bar(position=position_dodge())
```



Q: Notice that all references to the `spam` variable in these plots use a `factor()` function around the variable. Why is this the case? I.e., why did I use `x=factor(spam)` in the last plot instead of simply `x=spam`?

Comparing proportions

Very often we are interested in comparing the proportion of one variable across levels of another variable. For instance, are emails with big numbers flagged as spam more often than emails containing no numbers? This can also be phrased using the term **rate**: How does the rate of spam differ across the number categories? Here we can use the `prop.table()` function again to get proportions, but with the `margin` option to specify *which* proportions are calculated.

No margin: When `margin` is not specified, the **cell percents** are shown.

```
table(email$spam, email$number) %>% prop.table(margin=1)

##
##           big      none     small
##   0 0.1392797 0.1125492 0.7481711
##   1 0.1362398 0.4059946 0.4577657
```

Row margin: When `margin=1`, the **row percents** are shown. The percentages across the rows add up to 1. The interpretation puts the rows as the denominator; e.g., 74.8% of non-spam emails contain small numbers.

```
table(email$spam, email$number) %>% prop.table(margin=1) %>% round(3)

##
##       big  none small
##   0 0.139 0.113 0.748
##   1 0.136 0.406 0.458
```

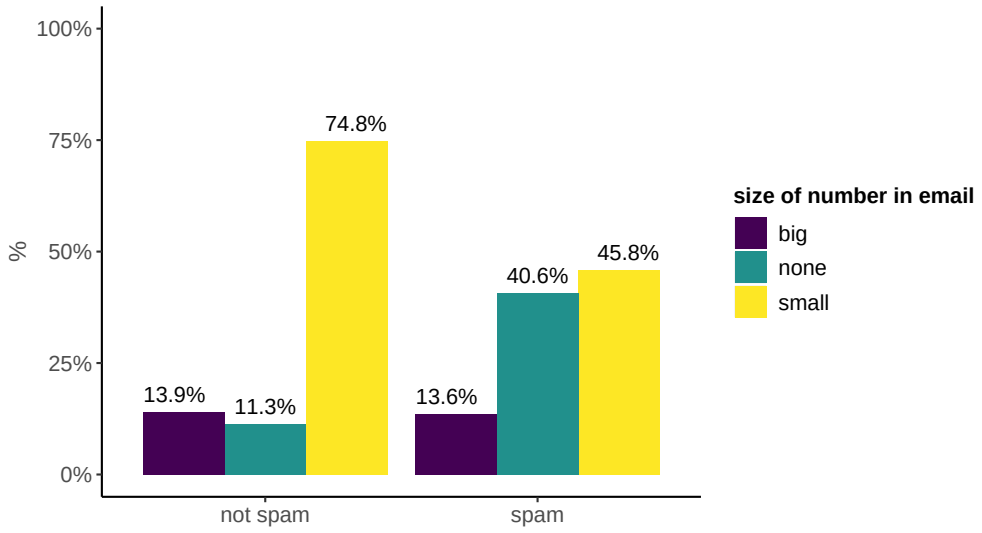
Column margin: When `margin=2`, the **column percents** are shown. The percentages down each column add up to 1. The interpretation puts the columns as the denominator; e.g., 5.9% of emails with small numbers in them are flagged as spam.

```
table(email$spam, email$number) %>% prop.table(margin=2) %>% round(3)

##
##       big  none small
##   0 0.908 0.729 0.941
##   1 0.092 0.271 0.059
```

Similar to univariate barcharts, creating side-by-side barplots of proportions requires some data management to calculate the desired proportions, then those percents are directly plotted.

Row percents: Here we want to plot the row percents. First, calculate the row proportions, as we did above, and convert the table to a data frame that I call `rowProps`. The row variable (`spam`) is now called `Var1`, the column variable `number` is now called `Var2`, and the proportion numbers that were in the body of the table are called `Freq`. Then we create a barplot using this new data set (not the original `email`).

```
rowProps <- table(email$spam, email$number) %>% prop.table(margin=1) %>% data.frame()

ggplot(rowProps, aes(x=Var1, y=Freq, fill=Var2)) +
  geom_col(position=position_dodge()) +
  # Everything below here is optional
  geom_text(aes(y=Freq+.04, label=paste0(round(Freq,3)*100, "%")),
                            position = position_dodge(width=1)) +
  scale_y_continuous(limits=c(0,1), labels = percent, name="%") +
  scale_x_discrete(name="", breaks=c("0","1"), labels=c("not spam", "spam")) +
  scale_fill_discrete(name="size of number in email")
```



This chart is a way to visualize the row proportions that we calculated earlier; e.g., 74.8% of non-spam emails contain small numbers and 13.6% of spam emails contain big numbers. *Note that any code below* `geom_col` *are optional and are NOT required to create a basic plot. If you are having trouble making this plot work, don't include it!*

Column percents: Similarly, let's plot the column percents. Since the `spam` variable is binary, it is not necessary to plot both non-spam and spam proportions (do you understand why?). Here we are only interested in the proportion of spam emails (when `spam` takes on a value of 1) within each `number` category.

We do this by calculating the column proportions as before, but then filter the data on `Var1 == 1` before saving the result as a new object that I call `colProps`. This simplifies the plot by only keep the rows where `Var1 == 1` (recall `spam` gets renamed to `Var1` when we save it as a `data.frame`).

```
colProps <- table(email$spam, email$number) %>% prop.table(margin=2)  %>%
            data.frame() %>% filter(Var1 == 1)

ggplot(colProps, aes(x=Var2, y=Freq)) +
  geom_col(position=position_dodge()) +
   # Everything below this line is optional
  theme_bw() +
  geom_text(aes(y=Freq+.04, label=paste0(round(Freq,3)*100, "%")),
                       position = position_dodge(width=1)) +
  ylab("%") + xlab("number") +
  scale_y_continuous(limits=c(0,1), labels=percent) +
  ggtitle("Proportion of emails flagged as spam")
```

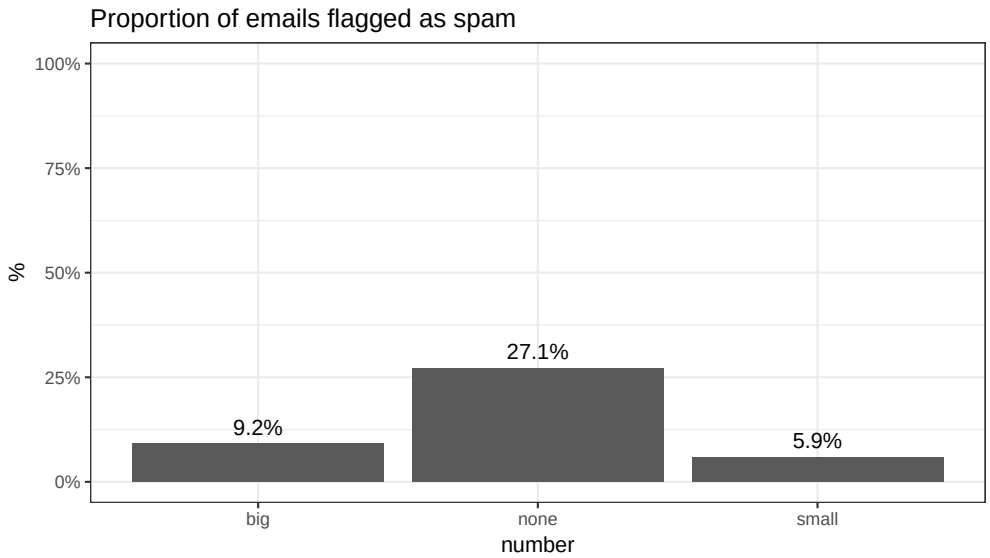


Figure 2.2: Example of plotting column percents for one group only

This plot visually represents the column proportions that we calculated earlier; e.g., 27.1%

of emails containing no numbers are flagged as spam and 5.9% of emails with small numbers in them are flagged as spam.

Example 2.6: Case Study: Using stents to prevent strokes

Consider an experiment that studies effectiveness of stents in treating patients at risk of stroke.[a] 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?

[a]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.

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

- Treatment group: Patients in the treatment group received a stent and medical management. The medical management included medications, management of risk factors and help in lifestyle modification.
- Control group: Patients in the control group received the same medical management as the treatment group, but they did not receive stents.

Researchers randomly assigned 224 patients to the treatment group and 227 to the control group. In this study, the control group provides a reference point against which we can measure the medical impact of stents in the treatment group. Researchers studied the effect of stents at two time points: 30 days after enrollment and 365 days after enrollment.

	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 2.1: Descriptive statistics for the stent study.

Table 2.1 summarizes the data into a **frequency table** in a more helpful way. In this table, we can quickly see what happened over the entire study. Let X_t be the number of patients in the treatment group who had a stroke by the end of their first year and N_t the number of patients in the treatment group overall.

Q: Using the data from Table 2.1, calculate the **sample proportion** $\hat{p}_t = {}^{X_t}/_{N_t}$ of the treatment group who had a stroke and the sample proportion $\hat{p}_c = {}^{X_c}/_{N_c}$ for the control group.

Q: Does the use of stents reduce the risk of stroke? Justify your answer in a complete English statement using the proportions you calculated as evidence to support your conclusion.

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?

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 on this specific stent type, and who may not be representative of all stroke patients. However, this study does leave us with an important lesson: we should keep our eyes open for surprises.

2.3.2 Continuous v. Categorical: Grouped Boxplots/Histograms

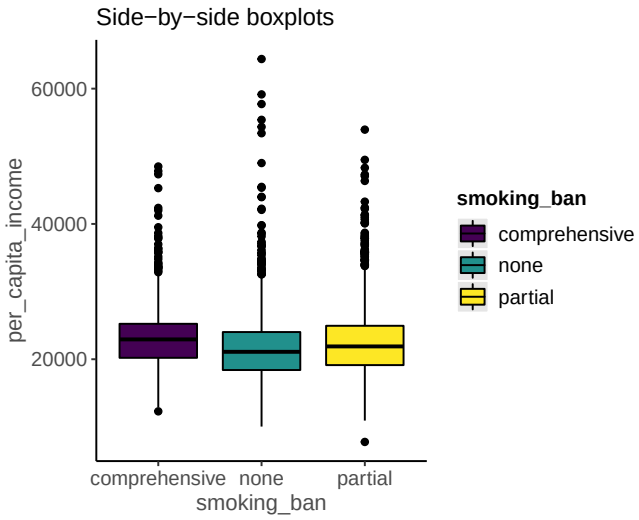
When comparing distributions of a continuous variable across levels of a categorical variable, it is **essential** to plot them on the same axis.

- Boxplot: For the side-by-side boxplots we just set the categorical variable on the x-axis and fill the boxes by that variable too.
- Density plots: Overlay (superimpose) density plots by coloring the density lines by the categorical variable.
- Histograms: Superimposed histograms are hard to read, so we need to put the histogram of the continuous variable in its own plot for each level of the categorical variable. This is called *paneling* or *faceting*. We can control the panel placement by setting the `ncol=1` argument.

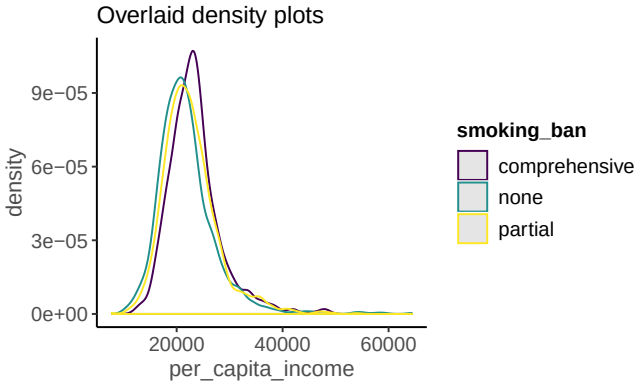
Example 2.7: Smoking bans and average income

We want to know if counties that enact one or more smoking bans have a higher per capita income. What is the continuous variable? What is the categorical variable?

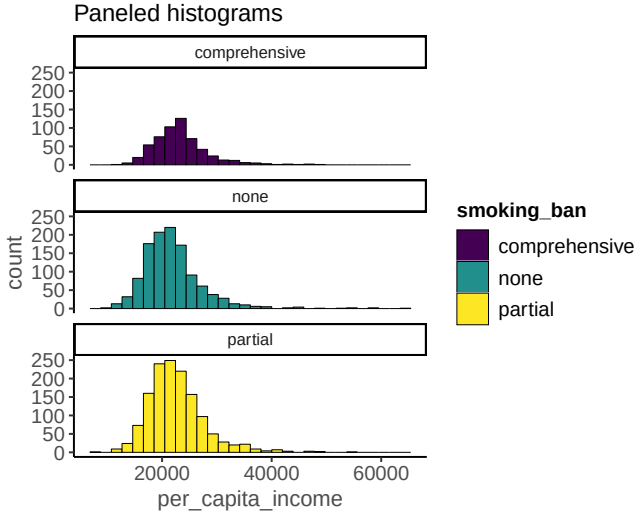
```
ggplot(county,
       aes(y=per_capita_income,
           x=smoking_ban,
           fill=smoking_ban)) +
    geom_boxplot() +
    ggtitle("Side-by-side boxplots")
```




```
ggplot(county,
       aes(x=per_capita_income,
           color=smoking_ban)) +
  geom_density() +
  ggtitle("Overlaid density plots")
```




```
ggplot(county,
       aes(x=per_capita_income,
           fill=smoking_ban)) +
  geom_histogram() +
  facet_wrap(~smoking_ban, ncol=1) +
  ggtitle("Paneled histograms")
```



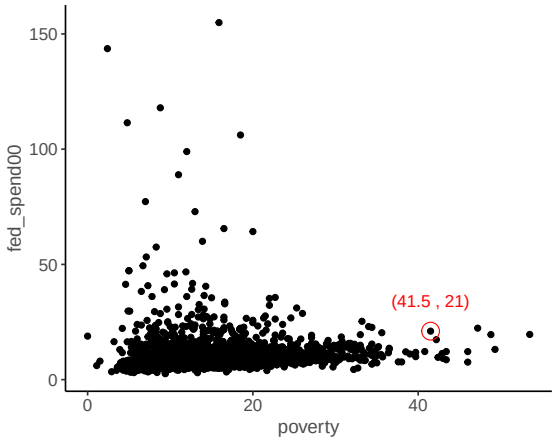
2.3.3 Continuous v. Continuous: Scatterplots & Correlation

Reading Assignment

OpenIntro Section 1.6.1

Scatterplots are the primary method used to study the relationship between two numerical variables. The scatterplot below compares the amount of federal spending per capita (`fed_spend00`), and the poverty rate (`poverty`) from the `county` data set.

Each point on the plot represents a single county. For instance, the highlighted dot corresponds to row 1073: Owsley County, Kentucky, which had a poverty rate of 41.5% and federal spending of $21.04 per capita.



The scatterplot suggests a relationship between the two variables: counties with a high poverty rate also tend to have slightly more federal spending. We might brainstorm as to why this relationship exists and investigate each idea to determine which is the most reasonable explanation.

There are three main features to talk about when describing the relationship between two continuous variables:

- Direction: What is the direction of association? Positive or negative?
- Strength: Is the relationship between y and x strong or weak?
- Form: Is the relationship linear? Quadratic? Clustered?

The correlation coefficient can provide information about the first two, but as we'll see in a moment, only by looking at the data can you learn about the third.

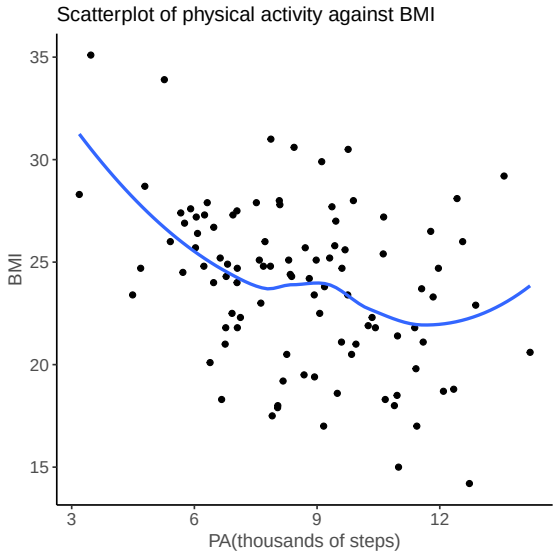
Reading Assignment

OpenIntro Section 7.1.4

The **correlation** measures the direction and strength of the linear relationship between two quantitative variables x and y. The symbol for the population correlation is ρ, and the sample correlation is just r.

Suppose we want to predict a person's BMI based on the level of physical activity (PA) they get. The most common display for examining the association (or relationship) between two quantitative variables is the scatterplot. This example uses the `PABMI` data set on physical activity and BMI value.

```
ggplot(bmi,
       aes(x=PA,
           y=BMI)) +
  geom_point() +
  geom_smooth(se=FALSE) +
  ylab("BMI") +
  xlab("PA(thousands of steps)") +
  ggtitle("Scatterplot of physical
      activity against BMI")
```



This scatterplot shows us that there is a negative association between BMI and PA; BMI decreases as physical activity increases. This trend is highlighted by the use of a locally weighted scatterplot smoothing **lowess** line. This is added using the `geom_smooth()` layer. We will not get into details about how lowess lines are calculated but they can be thought of as a moving average using only a small number of points each time.

Q: What is an example of a positive association?

The sample correlation coefficient r is an estimate for the population correlation ρ and is calculated as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/(n-1)}{s_x s_y} \tag{2.4}$$

Facts about ρ, the correlation coefficient:

1. The correlation coefficient is always between -1 (perfect negative correlation) and 1 (perfect positive correlation).
2. The correlation coefficient is unitless.
3. The sample correlation coefficient is symmetric in X and Y, so it does not depend on which variable you call the response and which variable you call the explanatory variable.
4. Correlation coefficients of 1 or -1 correspond to a perfect linear association (observations fall on a straight line) and a correlation coefficient of 0 implies no linear association between two variables.

The `cor()` command calculates the correlation coefficient r between two continuous numerical variables.

```
cor(bmi$BMI, bmi$PA)

## [1] -0.3854091
```

Exploring correlations. The following plots demonstrate the strength and direction of an association as measured by the magnitude and sign of the correlation coefficient r.

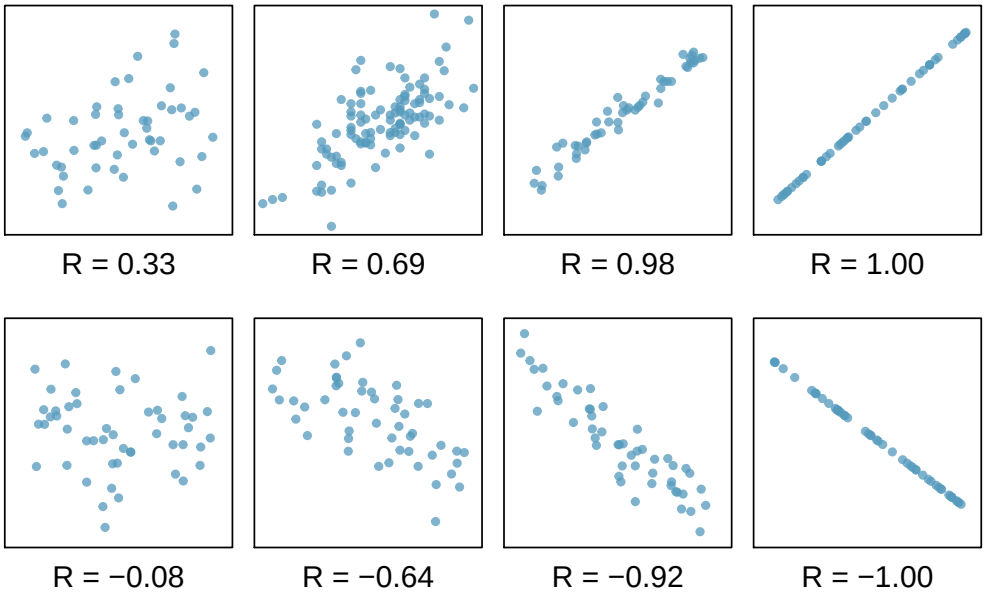


Figure 2.3: Sample scatterplots and their correlations. The first row shows variables with a positive relationship, represented by the trend up and to the right. The second row shows variables with a negative trend, where a high value in one variable is associated with a low value in the other.

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 2.4.

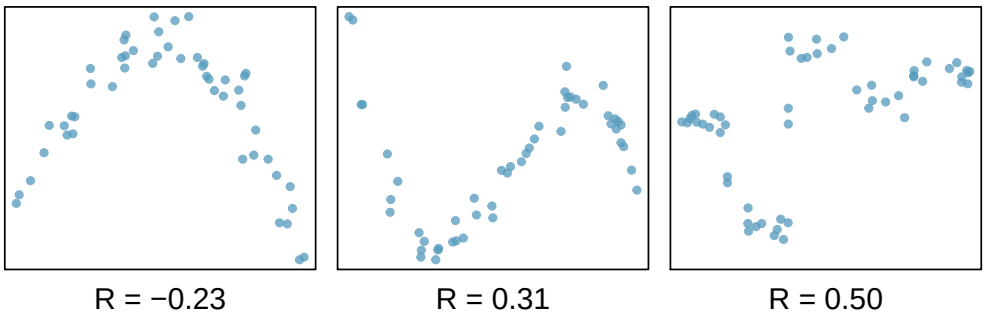


Figure 2.4: Sample scatterplots and their correlations. In each case, there is a strong relationship between the variables. However, the correlation is not very strong, and the relationship is not linear.

Closing remarks on exploratory data analysis Pictures speak a thousand words, but numerical summaries provide the specific details and allow for comparisons. The bare minimum description includes a measure of location and a measure of spread.

- Understand the background and context of the data.
- Always plot your data.
- Look for the overall pattern and for deviations such as outliers.
- Calculate an appropriate numerical summary to briefly describe the center and spread.
- When the distribution of the data is skewed, the five-number summary provides a better description of the data than simply the mean and standard deviation.

Chapter 3

Distributions of Random Variables

This chapter delves more into the distribution of data and the connection between the shape of the data distribution and the probability of an event occurring. We will introduce probability distributions using a discrete random variable, and then discuss the most common distribution used, the **Normal Distribution**. We will then see how we can use these distributions to describe characteristics of observed data.

3.1 Probability

Reading Assignment

OpenIntro Section 2.1, 2.1.1

- Probability: The probability of an outcome is the proportion of times an outcome would occur if we observed the random process an infinite number of times.
- Random Variable: A random variable is a variable which takes on a value based on the outcome of a random phenomenon (something whose outcome is uncertain).

Example 3.1: Examples of Random Variables

Q: Your instructor just passed out _____cards, _____of which are exploding kittens. What is the probability that you will draw an exploding kitten (and thus . . . explode)?
Q: What is the probability your neighbor exploded?

Example 3.2: Calculating Probability

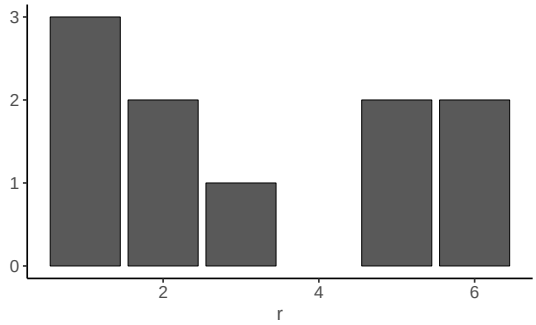
Q: Consider rolling a fair 6-sided die. What is the probability of rolling a 6?
Q: If you roll the same die a second time, what is the probability of rolling a 1 or 2?

3.2 Mathematical Models of Random Variables

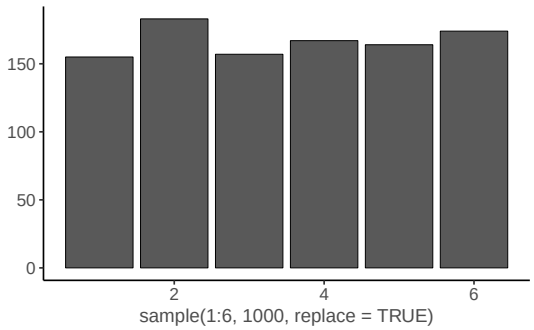
You are likely familiar with random processes if you have ever flipped a coin or rolled a die. These activities can be formalized using mathematical models in an easy way. Take for example the 6-sided die roll mentioned previously. If we roll the same die 10 times, a possible set of outcomes would look like the following, with the **frequency distribution** plotted to the right.

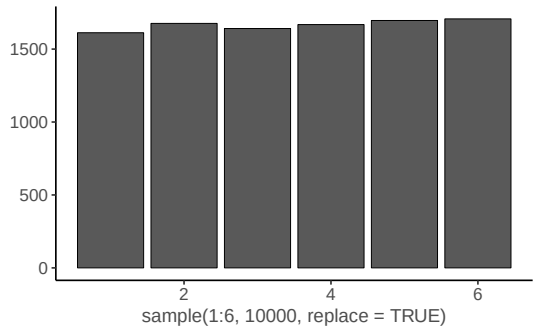
```
r <- sample(1:6, 10, replace=TRUE)
r

##  [1] 5 1 5 2 6 1 2 1 3 6
```

```
qplot(r, geom='bar')
```



When we examine the behavior over repeated sampling, we see that the frequency distribution levels out across all possible outcomes and becomes more **uniform**.



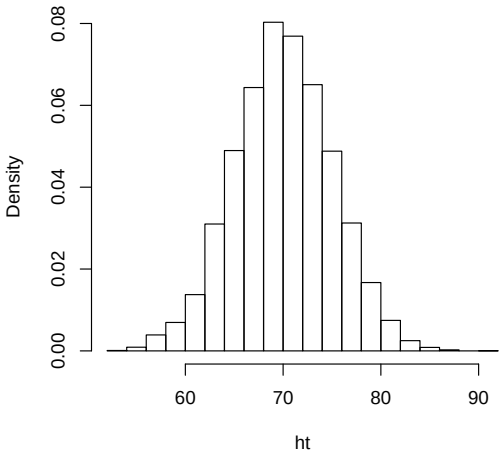


We can describe this distribution mathematically as

$$f(x) = P(X = x) = \frac{1}{6}$$

This says that if we let X be the outcome on a 6-sided die, then the probability of observing any one x is $\frac{1}{6}$. Note that probability distributions are simply functions of a specified random variable. This is an example of a **discrete** distribution. We'll see an example of a **continuous** probability distribution next.

Height is a continuous, numerical variable that tends to have a symmetric, bell-shaped distribution. Let's consider a hypothetical sample of measured heights of 10,000 males. A histogram their heights `[ht]` has been plotted to the right.



Example 3.3: Finding Empirical Probabilities

Using the information presented above, how can we find the probability that a male is over 80 inches tall? You cant calculate the exact probability given the information that has been presented so far. Brainstorm how *would* you go about finding the answer? What information do you need?

We can find the proportion of times the heights are over 80 inches by calculating how many heights are greater to or equal to 80 inches, and then dividing by the total number of records in the sample.

```
set.seed(8675309)
ht <- rnorm(10000, 70, 5)
over80 <- ht>=80
table(over80)

## over80
## FALSE  TRUE
##  9766   234
```

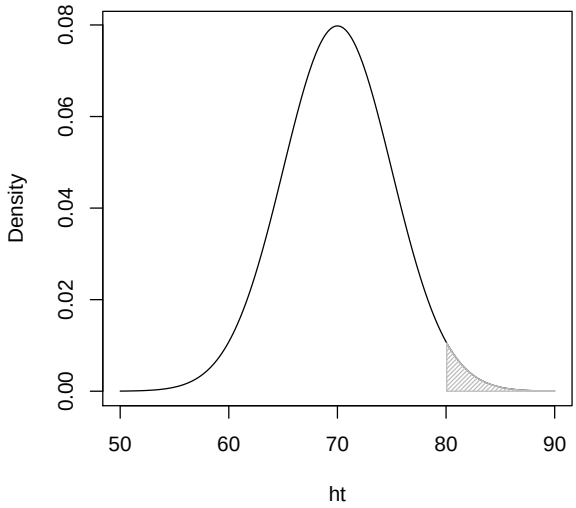
Q: Using the connection between the mean of a binary 0/1 indicator variable and a proportion, what is another way we could have calculated this proportion?

If a single individual is selected at random from the 10,000 observations, the probability of that individual of being over 80 inches is 234 / 10000 = 0.0234.

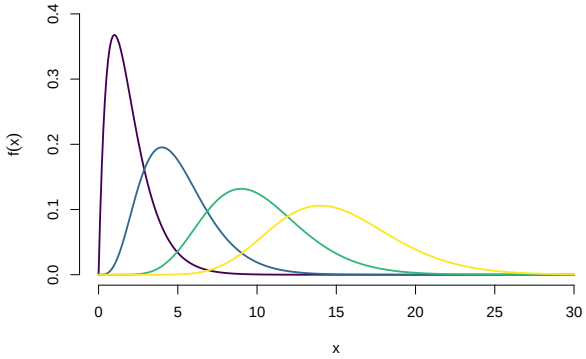
We just found the **empirical**, or observed probability of an event occurring. We will see later that if we know the equation for the probability distribution, then we can calculate a **distribution based** probability without observing all the individual data points. In fact, that's what this whole section is about!

A **density curve** describes the overall pattern of a distribution. Suppose we had an extremely large set of measurements and we constructed a histogram using many intervals, each with a very narrow width. The histogram for the set of measurements would be, for all practical purposes, a smooth curve. The calculated area under that smooth curve and above any range of values is the proportion of all observations that fall in that range.

The relative frequencies are proportional to areas over the class intervals and these areas possess a probabilistic interpretation. We can also represent this probability by shading this area in on the density plot.



There are curves of many shapes that can be used to represent a population relative frequency distribution for measurements associated with a random variable. Many of these distributions are already in use. If we know the distribution of a population, then it is easy to calculate the probability of something occurring by calculating the area under the curve.



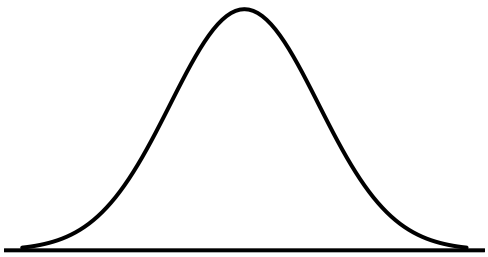
3.3 The Normal Distribution

Reading Assignment

OpenIntro Section 3.1

The Normal distribution is a symmetric, unimodal, bell curve that is ubiquitous throughout statistics. It is also known as the Gaussian distribution after Carl Friedrich Gauss, the first person to formalize its mathematical expression. Here is its mathematical description and its density curve:

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (3.1)$$



Facts about the normal distribution (or **normal model**):

1. A normal model is completely specified by the mean μ and variance σ^2 of the model.
2. $X \sim \mathcal{N}(\mu, \sigma^2)$ reads in English "the random variable X is distributed normally with parameters μ and σ^2. That means it has the function described in equation (3.1).
3. It is symmetric around μ and its tails extend to infinity.
4. The area under this distributional curve is 1.
5. It is a bell-shaped curve that can be used to approximate the histogram of a distribution of a quantitative variable.
6. Normal curves have the same shape regardless of the values of μ and σ^2.
7. A model is not useful unless it is flexible enough to be used in a variety of situations. In other words, by choosing different values of μ and σ^2, we can use the normal distribution to represent SAT scores, weights of newborn babies or heights of American adult women, as long as the distributions are unimodal and symmetric.
8. If a distribution of data follows a $\mathcal{N}(\mu, \sigma^2)$ distribution, then if we standardize all of the data values using $z = \frac{x-\mu}{\sigma}$, the standardized values will follow a $\mathcal{N}(0, 1)$ distribution. This **Standard normal distribution** is also called the Z-distribution, and it is a measure of the number of standard deviations a value is from the mean.

Example 3.4: Fact number 6: Normal curves have the same shape regardless of the values of μ and σ^2

Let's simulate some data to look into this fact.

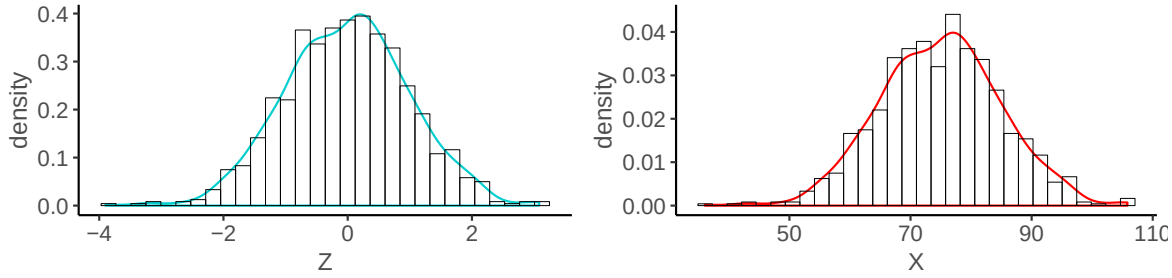
1. Use the `rnorm()` function to draw $n = 1000$ random variables from a normal distribution with mean $\mu = 75$ and variance $\sigma^2 = 100$ and store it as the vector `X`.

```
# rnorm(n, mu, s)

X <- rnorm(____, __, __)
```

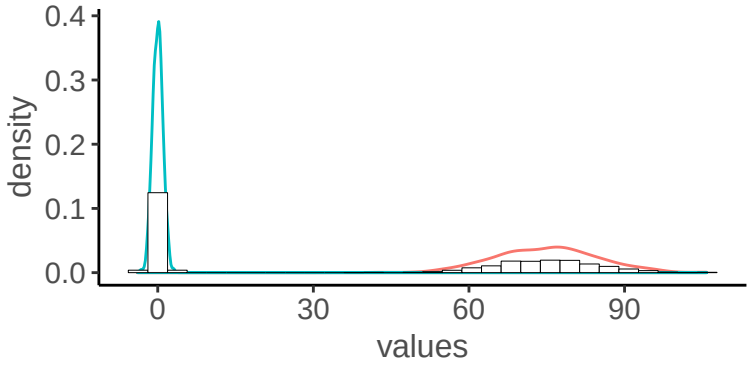
2. Calculate a new vector `Z` where $Z = \frac{x-\mu}{\sigma}$.

```
Z <- (X - __) / __
```

3. Plot histograms with overlying densities of these two variables.



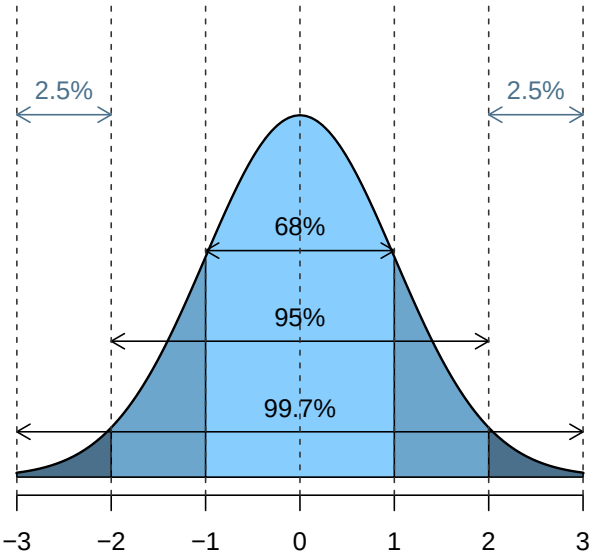
But notice the horizontal axis scales – quite different. Here's what it looks like if we plot them on the same axis.



The Z distribution **centers** the distribution of data around 0 and **scales** the standard deviation to a value of 1. This reiterates the importance of plotting the distributions on a common axis to properly compare them. (*Code not shown for this plot.*)

Q: Label the two density curves in the plot above as corresponding to either X or Z.

Empirical Rule One useful result for the normal model is that approximately 68% of the values fall within 1 sd of μ, 95% of the values fall within 2 sd of μ, and 99.7% of the values fall within 3 sd of μ.



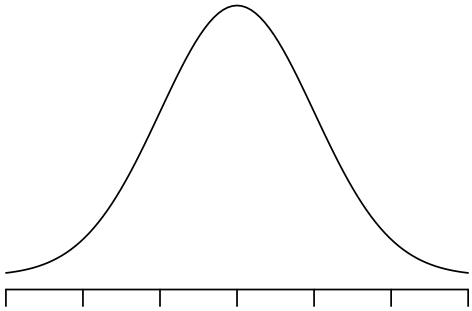
3.3.1 Calculating Probabilities with the Normal Distribution

- Normal models for the distributions of quantitative variables represent proportions by areas under the normal curve. The area under the entire normal curve is 1.
- Historically this was done by looking up values in tables, but that is not how it is done in the modern world. With the ease of using computers, and for the purpose of reproducible research, we will be doing this in `R`.
- We can find areas under normal curves using the `R` function `pnorm(z)`. This function takes a number and returns the area under the density curve to the **left** of that score. Let's look at `?pnorm` for more details.
- For another app-based demonstration of how this works, visit `https://gallery.shinyapps.io/dist_calc/`.

Example 3.5: Snowfall in Missoula

Suppose the annual snowfall amounts (inches) in Missoula are well-modeled by a normal distribution with mean of 46 inches and standard deviation of 18 inches. According to the normal model, approximately what proportion of years have between 25 and 50 inches of snowfall?

Step 1: Draw a picture and label the x-axis with μ in the center and each tick mark one σ away.



Amount of snowfall (in)

Step 2: Shade the region of interest. Write this as a probability statement.

$$P(\_\_\_\_\_ < X < \_\_\_\_\_)$$

Step 3: Calculate the z-scores of interest. Rewrite the shaded region as a probability statement in terms of z.

$$P(\_\_\_\_\_ < Z < \_\_\_\_\_)$$

Step 4: Find the area under the curve by providing the z-score to the `pnorm()` function.

```
pnorm(_____)
```

Repeat this, but use the x-score, mean and sd instead of the standardized z-score.

```
pnorm(q = __ , mean = __ , sd = __)
```

According to the normal model, about _____% of years in Missoula have between 25 and 50 inches of snowfall.

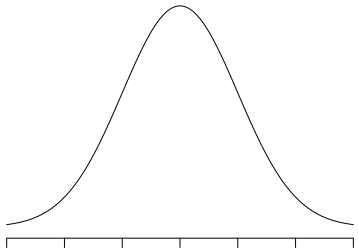
3.3.2 Inverse Normal Distribution

Sometimes, instead of wanting to know what proportion of values are in some interval, we want to know what value corresponds to some proportion (the inverse, or reverse question).

Example 3.6: Is it done snowing in Missoula yet?

According to the normal model, what is the 10^{th} percentile of snowfall amounts? That is, what is the amount of snowfall such that only 10% of snowfalls are smaller than this amount?

Step 1: Draw a picture and label the x-axis with μ in the center and each tick mark one σ away.



Step 2: Make a demarcation line on the plot that separates out the bottom 10% and label this z (take your best guess!). Shade the region to the left of that line.

$$P(Z < ??) = .10$$

Step 3: Use the `R` function `qnorm(`p`)` to find the z-score when $p = .10$.

```
qnorm(___)
```

Step 4: Back-calculate to find out what the x-score is that corresponds to the z-score: $x = z\sigma + \mu$.

Similarly, you can provide the mean and sd directly as arguments to `qnorm()` to find what the x value is that separates the bottom 10%.

```
qnorm(p = __, mean = __, sd = __)
```

According to the normal model, only 10% of snowfalls in Missoula are under _____inches.

3.4 Assessing Normality

Reading Assignment

OpenIntro Section 3.2

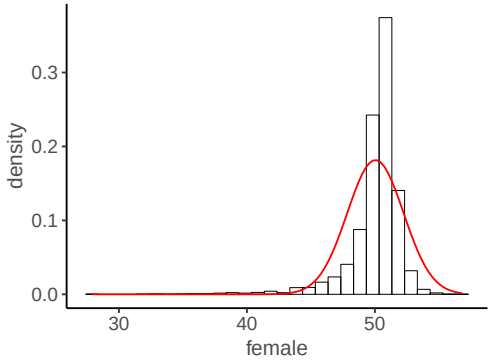
There are two visual methods for checking the assumption of normality, which can be implemented and interpreted quickly.

1. Histogram with overlaid normal density curve: 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.
2. Normal probability plot: Another more common method is examining a normal probability plot (also commonly called a **quantile-quantile plot or Q-Q plot**). The closer the points are to a perfect straight line, the more confident we can be that the data follow the normal model.

Example 3.7: Assessing Normality

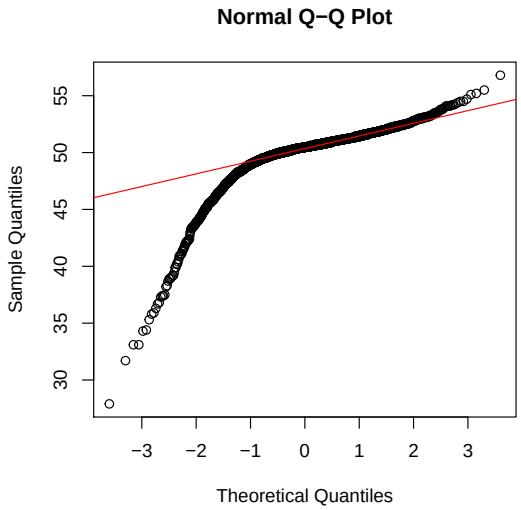
Using the `countyComplete` data set, assess the normality of the percent of residents that are female.

```
ggplot(county, aes(x=female)) +
 geom_histogram(aes(y=..density..),
  color="black", fill=NA) +
 stat_function(fun = dnorm,
               color="red",
  args = list(mean = mean(county$female),
              sd = sd(county$female)))
```



The distribution of the percent of residents that are female is skewed left.

```
qqnorm(county$female)

qqline(county$female, col="red")
```



There is no hard and fast rule to guide you on what can be considered "normal". What you are looking for is gross deviations away from the reference line or plot. The textbook has a more examples for you to review.

If you are ever overly concerned about violations of the normality assumption, you could perform the test you are interested in (for example a two-sample T-test) that relies on the assumption of normality, and a corresponding non-parametric test (such as Mann-Whitney) that does not have this assumption of a normal distribution. We will not cover non-parametric statistical procedures in this class, but it is important to know that they exist, and that you have options if your data are not normal.

Many statistical procedures are **robust** to deviations from normality. You are likely to come to the same conclusion using a non-parametric as you find using a parametric procedure.

3.5 The T-distribution

Reading Assignment

OpenIntro Section 5.1

The T distribution is a more flexible distribution when sample sizes are small compared to the Z distribution. Most statistical analysis programs only report a t statistic, so we introduce it here to emphasize the similarities.

There is a different t distribution for each sample size. A particular t distribution is specified by its **degrees of freedom** (df), typically calculated as $n-1$. The density curves for the t distribution are similar to the normal curve; however, the tails of the distribution are thicker. They are symmetric about 0 and bell-shaped. As the degrees of freedom increase the t-distribution becomes more and more similar to the normal distribution.

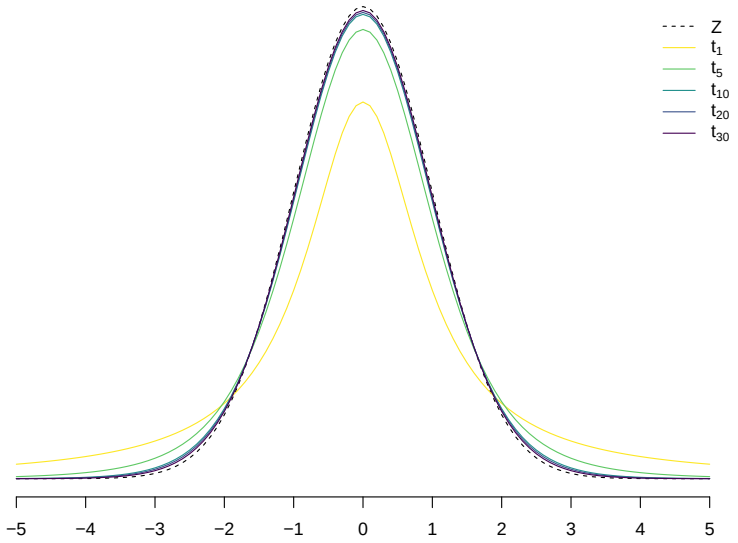


Figure 3.1: Comparison of T and Z distributions

The method to calculate the probability and the inverse probability for the family of t distributions is similar to the one for the normal distribution.

Probability To find the area under the curve to the left of a quantile q, we find $P(t_1 < 2) =?$:

```
pt(q=2, df=1)
```

```
## [1] 0.8524164
```

Inverse Probability To find the value of t_2 that has .05 in the upper tail we find $P(t_2 >?) = .05$ or $P(t_2 <?) = .95$:

```
qt(.95, df=2)
```

```
## [1] 2.919986
```

Chapter 4

Foundations for Inference

Reading Assignment

OpenIntro Chapter 4

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. Understanding this chapter will make the rest of this book, and indeed the rest of statistics, seem much more familiar (no pressure!).

4.1 Parameters and Statistics

A number that describe the population is called a **parameter**, a number that describes the sample is called a **statistic**. The value of various population parameters are rarely known, but there exist corresponding statistics that can be **calculated on data** from a sample that are used to **estimate** these parameters. For example, We want to estimate the **population mean** based on the sample. The most intuitive way to go about this is to simply take the **sample mean**. Symbols for parameters are typically Greek letters, where statistics use Roman letters.

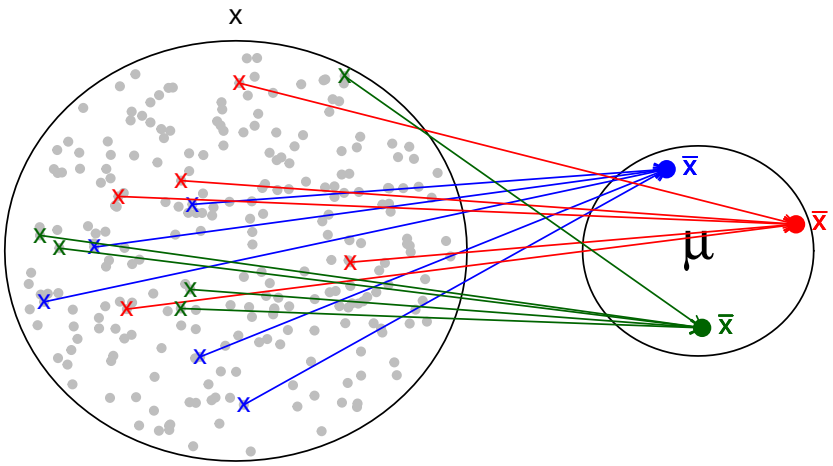
Measurement	Statistic	Calculation	Parameter
Proportion	$\hat{p}$	x/n	p
Mean	$\bar{x}$	$\frac{1}{n}\sum_{i=1}^{n} x_i$	μ
Variance	s^2	$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	σ^2

4.2 Point Estimates

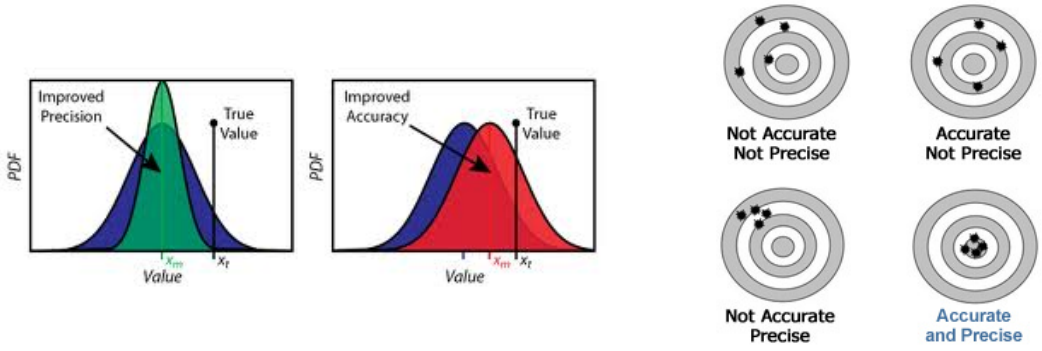
Sample statistics can also be called **point estimates**. A statistic calculated from a sample is considered a **random variable** and is subject to random variation because it is based on a sample from the population.

Q: What do we mean by random variation?

We do not expect $\bar{x}$ to be equal to μ every single time we draw a sample. The point estimates vary across samples; the size of this **sampling variation** gives us an idea of how close our estimate may be to the parameter.



We can see that these point estimates are not exact: they vary from one sample to another. So what makes a good estimate? Here are two ways to visualize it, both involving the concept of how close you are to hitting your target.



A statistic is a good estimate for a population parameter if it's accurate and precise. **Accuracy** tells you how close to the location of the parameter you are. **Precision** tells you about the variation in the estimate. Low variation means improved precision.

4.3 Sampling Distribution

The **sampling distribution** represents the distribution of the point estimates based on samples of a fixed size n from a certain population. For the diagram in section 4.2, each $\bar{x}$ was calculated using $n = 5$ data points. That is a pretty small sample – not a lot of information from 5 data points – but as we'll see you can still make inference with small samples!

It is useful to think of a particular point estimate as being drawn from a distribution of point estimates. Understanding the concept of a sampling distribution is central to understanding statistical inference.

The standard deviation associated with the estimate is called the **standard error**. It describes the typical level of uncertainty associated with that estimate. Specifically for the point estimate $\bar{x}$ we quantify the uncertainty using $\sigma_{\bar{x}}$, also called the standard error (SE).

But when you take a single sample, you can only calculate one $\bar{x}$. So how can you calculate the standard deviation from a single point? A bit of statistical theory provides a helpful tool to address this issue. Specifically the **Law of Large Numbers** states that the larger the

sample you draw, the closer the sample mean will be to the true population mean.

$$\mu_{\bar{x}} = \mu_x \tag{4.1}$$

$$\sigma^2_{\bar{x}} = \frac{\sigma^2_x}{n} \tag{4.2}$$

The mean of the sampling distribution is the same as from the original population (4.1), but the variance shrinks by a factor of n (4.2) (the number of data points used to calculate that point estimate). In summary:

1. Point estimates from a sample may be used to estimate population parameters.
2. Point estimates are not exact: they vary from one sample to another.
3. The standard error is the uncertainty of the sample mean, and gets smaller the more data you use to calculate the point estimate.

Example 4.1: Sampling Distributions, LLN, and the CLT

In class simulation

4.4 The Central Limit Theorem

Reading Assignment

OI Section 4.4

The **Central Limit Theorem** is a key assumption that is commonly relied on for many statistical procedures. Informally it says that as the sample size increases, the sampling distribution converges to a normal distribution. We saw this empirically by simulating data from different distributions.

Formally and mathematically, the **Central Limit Theorem** states that if $X_1, \ldots, X_n$ is a random sample from a distribution with mean μ and positive variance σ^2, then if you standardize $\bar{X}$ by subtracting its mean and dividing by the standard error,

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\frac{\sigma_x}{\sqrt{n}}} \tag{4.3}$$

the resulting random variable Z has a $\mathcal{N}(0, 1)$ distribution as $n \to \infty$ (gets really large). We recognize this distribution as the $\mathcal{Z}$ distribution. In practice, $n > 30$ is considered "large enough" for the CLT to hold.

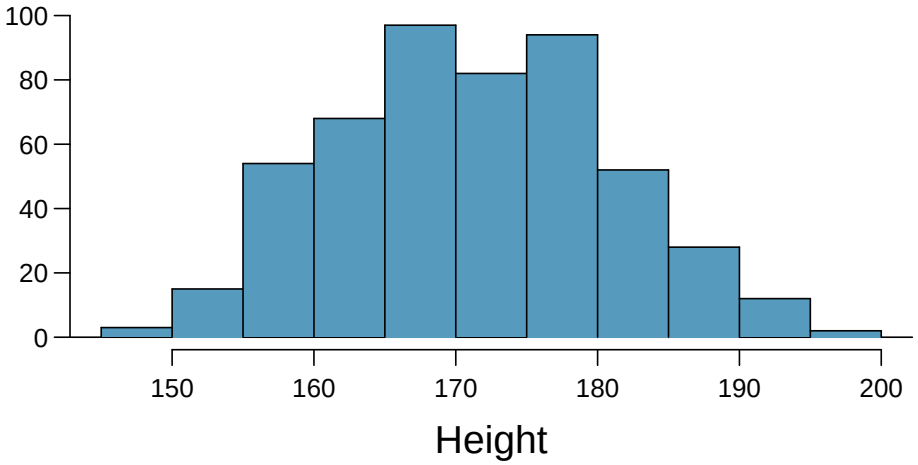
4.4.1 Using the CLT and LLN to calculate probabilities about an average

The CLT is an important theorem because it allows us to assume approximate normality of the sample means given a large enough n, regardless of the distribution of the individual random variables. Commonly used statistical procedures concerned with making a statement about the population based on a sample require the assumption of a normal distribution for the means.

Here we are going to demonstrate that as long as you know the point estimate for the mean, and the standard deviation of that point estimate, and you can assume a normal distribution, then you can calculate probabilities under the normal model the same way you did in Section 3.3.1.

Example 4.2: OpenIntro Problem 4.4: Heights of Adults

Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender, for 507 physically active individuals. The histogram below shows the sample distribution of heights in centimeters.



Min	147.2
Q1	163.8
Median	170.3
Mean	171.1
SD	9.4
Q3	177.8
Max	198.1

1. What is the point estimate for the average height of active individuals? What about the median?

2. What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR?

3. Is a person who is 1m 80cm (180 cm) tall considered unusually tall? And is a person who is 1m 55cm (155cm) considered unusually short? Explain your reasoning.

4. The researchers take another random sample of physically active individuals. Would you expect the mean and the standard deviation of this new sample to be the ones given above? Explain your reasoning.

5. The sample means obtained are point estimates for the mean height of all active individuals. Calculate the standard deviation of this sample mean using (4.2).

6. What is the probability of a single height being below 160cm?

7. What is the probability that the *average* height of 40 randomly sampled individuals is below 160cm?

4.5 Confidence Intervals

Reading Assignment

OpenIntro Section 4.2

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 see 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. A confidence interval is a balanced interval that centers on the point estimate and ranges out to each side a distance called the **margin of Error** (ME).

$$\texttt{point estimate} \pm \texttt{Margin of Error}$$

Q: If we want to be very certain we capture the population parameter, should we use a wider interval or a smaller interval?

Recall the Empirical Rule stated that 95% of the data falls within two standard deviations of the mean. We can use this information to construct a plausible confidence interval where we can say that we can be roughly 95% **confident** that we have captured the true parameter in this interval.

$$\bar{x} \pm 2 * SE \tag{4.4}$$

But what does "95% confident" mean? Suppose we took many samples and built a confidence interval from each sample using Equation (4.4). Then about 95% of those intervals would contain the actual mean, μ.

Figure 4.1 shows this process with 25 samples, where 24 of the resulting confidence intervals contain the average time for all the runners, $\mu = 94.52$ minutes, and one does not.

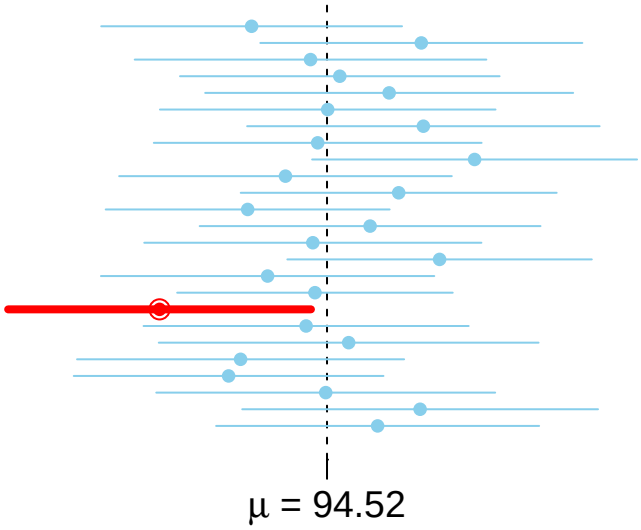


Figure 4.1: Confidence intervals for twenty-five samples of size $n = 100$ were created to try to capture the average 10 mile time for the population.

Example 4.3: Calculating an Approximate Confidence Interval

A forester wishes to estimate the average number of "count trees" per acre (trees larger than a specified size) on a 2000-acre plantation. She can then use this information to determine the total timber volume for trees in the plantation. A random sample of $n = 50$ 1-acre plots is selected and examined. The average number of count trees per acre is found to be 27.3, with a standard deviation of 12.1 trees per acre. Use this information to construct an approximate 95% confidence interval for the mean number of count trees per acre for the entire plantation.

We are approximately 95% confident that the mean number of count trees per acre for the plantation is contained in the interval (_____, _____).

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 one of the most common errors: 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 especially important consideration of confidence intervals is that they *only try to capture the population parameter*. Our intervals say 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.

Reading Assignment

`http://www.r-bloggers.com/when-discussing-confidence-level-with-others/`
`http://www.r-bloggers.com/the-95-confidence-of-nate-silver/`

The rule where about 95% of observations are within 2 standard deviations of the mean is only approximately true. It is a good "quick and dirty" way to get an idea of an interval when you don't have a calculator on hand.

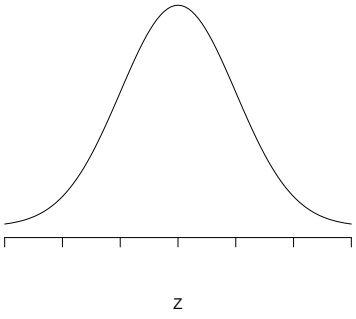
However we know that for large sample sizes, the CLT allows us to say that the sample mean can be approximated by a normal distribution. The equation for a 95% confidence interval then becomes

$$\bar{x} \pm 1.96 * SEM \tag{4.5}$$

But what if we want a wider net? Let's increase the confidence level to 99%. The equation is now

$$\bar{x} \pm 2.58 * SEM \tag{4.6}$$

Q: How did I know to use 1.96 and 2.58? I'm just brilliant like that?



The generic equation of a (100-α)% confidence interval then can be written as

$$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \tag{4.7}$$

The $z_{\frac{\alpha}{2}}$ is called the **critical value** and is the value of z from the standard normal distribution that has a tail area to the right of $\frac{\alpha}{2}$. Let's look at some common values of z

```
round(qnorm(.995),3) # 99% Confidence, alpha = .01
## [1] 2.576
round(qnorm(.975),3) # 95% Confidence, alpha = .05
## [1] 1.96
round(qnorm(.950),3) # 90% Confidence, alpha = .10
## [1] 1.645
```

4.6 Assumptions

Statistical tools rely on conditions. When the conditions are not met, these tools are unreliable and drawing conclusions from them is treacherous. The conditions for these tools typically come in two forms.

- **The individual observations must be independent.** A random sample from less than 10% of the population ensures the observations are independent. In experiments, we generally require that subjects are randomized into groups. If independence fails, then advanced techniques must be used, and in some such cases, inference may not be possible.
  - If the observations are from a simple random sample and consist of fewer than 10% of the population, then they are independent.
  - Subjects in an experiment are considered independent if they undergo random assignment to the treatment groups.

  - If a sample is from a seemingly random process, e.g. the lifetimes of wrenches used in a particular manufacturing process, checking independence is more difficult. In this case, use your best judgment.

- **Other conditions focus on sample size and skew.** For example, if the sample size is too small, the skew too strong, or extreme outliers are present, then the normal model for the sample mean will fail.
  - When there are prominent outliers present, the sample should contain at least 100 observations, and in some cases, many more.
  - This is a first course in statistics, so you won't have perfect judgment in assessing skew. That's okay. If you're in a bind, either consult a statistician or learn about the studentized bootstrap (bootstrap-t) method.

Verification of conditions for statistical tools is always necessary. Whenever conditions are not satisfied for a statistical technique, there are three options. The first is to learn new methods that are appropriate for the data. The second route is to consult a statistician. The third route is to ignore the failure of conditions. This last option effectively invalidates any analysis and may discredit novel and interesting findings.

Finally, we caution that there may be no inference tools helpful when considering data that include unknown biases, such as convenience samples. For this reason, there are books, courses, and researchers devoted to the techniques of sampling and experimental design.

4.7 Hypothesis Testing

READING ASSIGNMENT

OpenIntro Chapter 4.3.1

Confidence intervals are one way to make inference regarding a population parameter. The second type of inference-making procedure is hypothesis testing. There are two competing hypotheses involved. The first is the **research hypothesis** (or **alternative hypothesis**) and the second is the negation of this hypothesis, called the **null hypothesis**.

We will be covering several new point estimates and types of comparisons. We will continue to practice making inference on a single population mean, μ, then quickly move through the other cases.

These course notes introduce and discuss hypothesis testing slightly out of order compared to the Open Intro textbook. It is *highly* recommended that you read the textbook section all the way through first before continuing on with these notes.

4.7.1 The 5 steps to a good statistical test

A hypothesis test is composed of the following 5 steps: As you learn to formulate and test hypotheses in a clear and consistent manner, some of these steps will be combined or not explicitly numbered. However you will be required to clearly address all five points so that important steps like checking assumptions do not get forgotten.

1. Identify and define the parameter(s) to be tested.
2. Translate the English hypothesis into a statistical statement using symbols.
3. Determine the appropriate statistical method and check the assumptions of that method.
4. Compare your data to the hypothesized value by calculating a test statistic and a p-value. Make a decision to reject the null hypothesis in favor of the alternative or not.
5. State your conclusion in a full English sentence in the context of the research hypothesis using no symbols or statistical jargon.

I. Identify and define the parameter(s) to be tested.

Explicitly define the population parameter of interest.

- Let μ be the true mean ...
- Let p_1 be the true proportion of ... in (define group 1) and p_2 be the true proportion of ... in (define group 2).

II. Translate the English hypothesis into a statistical comparison.

- Null Hypothesis: The null hypothesis H_0 represents the hypothesis of "no effect", "no change", "status quo" or "nothing unusual." It is always expressed as an equality. This often represents a skeptical position or a perspective of no difference.
- Alternative Hypothesis: The alternative hypothesis H_A represents an alternative claim under consideration and is often represented by a range of possible parameter values. The alternative hypothesis represents the hypothesis of "an effect," "a change" or that the status quo does not hold.

The different ways the hypothesis can be written mathematically are as follows, where μ_0 is a quantifiable number defined in the research question.

# of tails	Side	H_A
One Tail	Upper Tail	$\mu > \mu_0$
	Lower Tail	$\mu < \mu_0$
Two Tail		$\mu \neq \mu_0$

III. Determine the most appropriate statistical method and check the assumptions of that method. Based on type of data and hypothesis, what is the appropriate analysis technique? What are the assumptions for this test and are they met?

IV. Compare your data to the hypothesized value by calculating a test statistics and a p-value. Make a decision to reject the null hypothesis or not. We want to see whether the sample data collected tend to support or contradict the null hypothesis. This is done by calculating a **test statistic (t.s.)** using the point estimate of the **parameter** we

are wanting to measure, and a **p-value**. All test statistics are calculated by subtracting the hypothesized value in the null hypothesis (θ_0) from the point estimate ($\hat{\theta}$), divided by the standard error of that point estimate ($SE_{\hat{\theta}}$).

$$t.s. = \frac{\hat{\theta} - \theta_0}{SE_{\hat{\theta}}} \tag{4.8}$$

This p-value quantifies the amount of evidence against the null hypothesis. The p-value of a test is the probability of obtaining a test statistic at least as extreme as the one observed assuming the null hypothesis is true. How exactly this p-value is calculated depends on the probability distribution of the statistic being tested (Eq 4.8), and the sampling method.

There are two basic views toward drawing conclusions for a hypothesis test.

Decision Making View. The p-value is compared to a pre-determined significance level (α) and one of two outcomes occur.

1. The p-value $< \alpha$, Reject H_0. There is sufficient evidence to support H_A.
2. The p-value $> \alpha$, Do not reject H_0. There is insufficient evidence to support H_A.

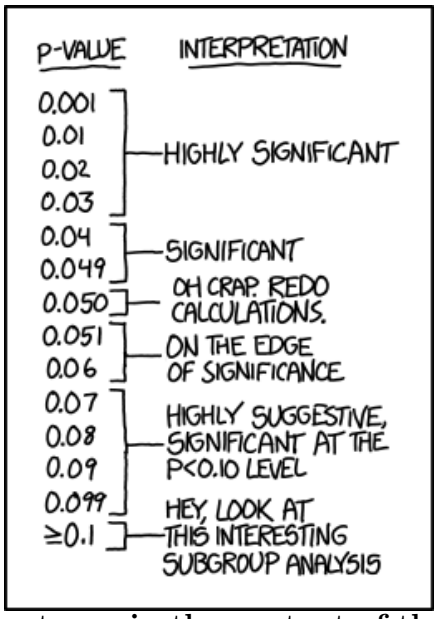
This significance level α is typically set at the arbitrary value of 0.05 and represents the chance, or probability that the researcher is willing to be wrong by incorrectly rejecting H_0.

Subjective View. Here we don't conclude that H_0 is correct or incorrect; we simply try to determine the weight of evidence against H_0. The smaller the p-value is the more evidence we have against the null hypothesis in favor of the alternative hypothesis. Some guidelines are:

- If .05 < p-value<.10 we have some evidence against H_0.
- If .01 < p-value<.05 we have moderate evidence against H_0.
- If .001< p-value<.01, we have strong evidence against H_0.
- If p-value < .001, we have very strong evidence against H_0.

A popular webcomic XKCD helps us out with some suggestions on other terms we can use to describe the size of the p-value.

(Ref: `https://xkcd.com/1478/` *)*



V. State your conclusion in a full English sentence in the context of the research hypothesis using no symbols.

Things to keep in mind when writing this conclusion:

- The null hypothesis represents a skeptic's position or a position of no difference. We reject this position only if the evidence strongly favors H_A.
- A small p-value means that if the null hypothesis is true, then 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 favor of the alternative.
- A large p-value does not "prove" the null hypothesis; it only means that the data are consistent with the null hypothesis.
- You do not "accept" the alternative hypothesis. This is not a test of the truthfulness of the alternative but a rejection of the null.
- Always state the conclusion of the hypothesis test in plain language so non-statisticians can also understand the results.

4.7.2 Practical vs Statistical Significance

Reading Assignment

Open Intro 4.5.5

- It is common practice to use a significance level of .05. However, there is no sharp border between "significant" and "not significant"; there is only increasingly strong evidence as the p-value decreases.
- When sample sizes are very large, even small deviations from the null hypothesis will be significant. If this small difference (say a mean weight loss of .5 lbs) does not really make a difference in the situation, we say that the result is statistically significant but not practically significant.
- This is one reason why it is important to report both a p-value and a confidence interval. The p-value will give statistical significance while the confidence interval will give the size of the effect.

Example 4.4: CHF in males

A study was conducted of 90 adult male patients following a new treatment for congestive heart failure. One of the variables measured on the patients was the increase in exercise capacity (in minutes) over a 4-week treatment period. The previous treatment regime had produced an average increase of $\mu = 2$ minutes. The researchers wanted to evaluate whether the new treatment had increased the mean in comparison to the previous treatment. The data yielded a sample mean $\bar{x}$ of 2.17 minutes and standard deviation $s = 1.05$ minutes. At the 5% significance level, what conclusions can you draw about the research hypothesis?

I. Let _____be _____

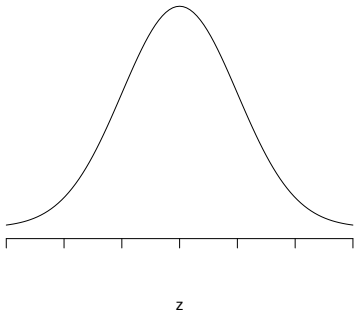
II. H_0: _____, H_a: _____.

III. Our sample size is _____so we can use the _____to assume that _____is _____distributed. We then can use a _____.

IV. The p-value of _____, is _____than $\alpha =$ _____but provides _____evidence to _____H_A

$$z^* = \frac{\hat{\theta}-\theta_0}{SE_{\hat{\theta}}} = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} =$$

```
# Use R to find the p-value
```



V. Using the *decision view* we would conclude:

Using the *subjective view* we would conclude:

4.7.3 Using a Confidence Interval to make inference

- Calculate a 95% confidence interval for the true mean μ increase in exercise capacity for males in this study.

- Plot this interval on your picture above. Does this interval cover the true mean?
- Interpret the CI in context of the problem.

- What can you conclude about H_A given the results from this CI?

- Will this always be the case? Why?

Example 4.5: Something's wrong here

Here are several situations where there is an incorrect application of the ideas presented in this section.

1. A change is made that should improve student satisfaction with the way grades are processed at Chico State. The null hypothesis, that there is an improvement, is tested versus the alternative, that there is no change.
2. A significance test rejected the null hypothesis that the sample mean is .25.
3. A report on a study says that the results are statistically significant and the p-value is .95.

4.8 Decision Errors

Reading Assignment

Open Intro 4.3.3

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 4.1.

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	okay	Type 1 Error
	H_A true	Type 2 Error	okay

Table 4.1: Four different scenarios for hypothesis tests.

- Type I error: A Type I error is committed if we reject the null hypothesis when it is true. The probability of making a Type I error is α.
- Type II error: A Type II error is committed if we fail to reject the null hypothesis when the null hypothesis is false. The probability of making a Type II error is β.

Example 4.6: The US Judiciary System

According to US law, an individual is innocent until proven guilty.
Q: What are the null and alternative hypotheses?
Q: Define a Type I and Type II error in this context.
Q: In light of the above answer, what would be the type of error we would want to minimize or control?

Chapter 5

Univariate inference

The prior chapter introduced concepts and techniques used to make inference about a population parameter using data from a sample. The examples used were all about making inference on a single mean from a large sample.

In this chapter we will extend those ideas to a single mean from a small sample, a single proportion, and learn how to use R to perform these analyses for us given a set of data.

5.1 Small sample inference for a single mean

Reading Assignment

OpenIntro Chapter 5.1

So far we have required a large sample in order to do inference 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? As we saw in the sampling distribution lab, if the population data are nearly normal, then $\bar{x}$ will also follow a normal distribution, which addresses the first problem. However, 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?

In conclusion, 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.

The accuracy of the standard error is trickier, and for this challenge we'll use 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.

5.2 Using R for Hypothesis Testing

The most used function we will be using in this class is the `t.test()`. It has a lot of arguments that make it very versatile. Be sure to read `?t.test` for more help on the options. The most commonly used arguments are:

```
t.test(x, alternative = c("two.sided", "less", "greater"),
       mu = 0, conf.level = 0.95, data, subset)
```

Example 5.1: Average age at first marriage

According to Wikipedia, the average age for a woman in the US to get married is 28 years (`http://en.Wikipedia.org/wiki/Age_at_first_marriage`). The average age at first marriage of 5,534 US women who responded to the National Survey of Family Growth (NSFG) conducted by the CDC in the 2006 and 2010 cycle was 23.4.

Q: Is there reason to believe that women who respond to the NSFG survey marry significantly earlier than the average woman?

Let's set up a full 5 step-hypothesis test to answer the research hypothesis.

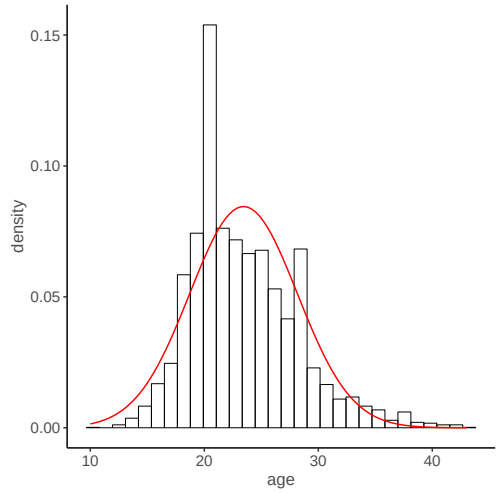
Notice that R is **NOT** used until step 3!

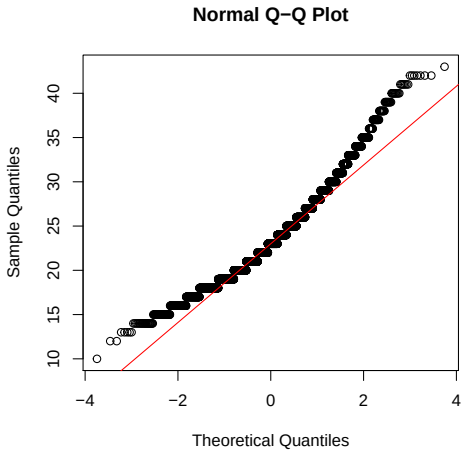
I.

II.

III.

```
qqnorm(marriage$age)
qqline(marriage$age, col="red")
```





IV.

```
t.test(marriage$age, mu=28, alternative="less")

##
##  One Sample t-test
##
## data:  marriage$age
## t = -71.845, df = 5533, p-value < 2.2e-16
## alternative hypothesis: true mean is less than 28
## 95 percent confidence interval:
##     -Inf 23.5446
## sample estimates:
## mean of x
##  23.44019
```

V.

5.2.1 Calculating Confidence Intervals using R

Since confidence intervals are two-sided, the `t.test` needs to be run again with the correct argument for the alternative.

```
t.test(marriage$age, mu=28, alternative="two.sided")

##
##  One Sample t-test
##
## data:  marriage$age
## t = -71.845, df = 5533, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 28
## 95 percent confidence interval:
##  23.31577 23.56461
## sample estimates:
## mean of x
##  23.44019
```

We can be 95% confident that the true mean age at marriage is contained in the interval (_____, _____) years.

5.3 Inference for a single proportion

Reading Assignment

OpenIntro Chapter 6.1

Inference on proportions is needed to answer many questions such as *"What proportion of the American public approves of the job the Supreme Court is doing?"* Any measurement that can be expressed as a binary categorical variable falls into this category.

Recall that a proportion p is a summary statistic for categorical data and can be calculated as the number of items in the category of interest, divided by the sample size. So the point estimate ($\hat{p}$) and standard error of that estimate ($SE_{\hat{p}}$) are

$$\hat{p} = \frac{\sum x}{n} \qquad SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{5.1}$$

The methods we learned in the previous chapter will continue to be useful in these settings. For example, sample proportions are well-characterized by a nearly normal distribution when certain conditions are satisfied, making it possible to employ the usual confidence interval and hypothesis testing tools.

5.3.1 Conditions for the sampling distribution of $\hat{p}$ being nearly normal

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

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

If these conditions are met, then the following test statistic can be approximated with the normal distribution,

$$z^* = \frac{\hat{\theta} - \theta_0}{SE_\theta} = \frac{\hat{p} - p_0}{\sqrt{\frac{p(1-p)}{n}}}, \tag{5.2}$$

and $100 - \alpha\%$ confidence interval calculated as

$$\text{point estimate} \pm \text{Margin of Error}$$
$$\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

5.3.2 Hypothesis testing with R for a proportion

There are two approaches we can take here to construct confidence intervals and hypothesis tests for a proportion, depending on if we have access to the full data set or if we have summary statistics only. The first example uses summary data only, while the second demonstrates how we can analyze proportions when we have access to the raw data.

Example 5.2: Proportion of spam emails - summary data only

Consider the `email` data set that contains information about the contents in email, and whether or not the email was flagged as spam. It was found that 367 out of the 3921 emails were flagged as spam. Can we say that the proportion of emails flagged as spam is less than 10%?

Q: What is the sample proportion of spam?

I.

II.

III.

IV.

```
prop.test(x=367, n=3921, p=.1, alternative="less")

##
##  1-sample proportions test with continuity correction
##
## data:  367 out of 3921, null probability 0.1
## X-squared = 1.7149, df = 1, p-value = 0.09518
## alternative hypothesis: true p is less than 0.1
## 95 percent confidence interval:
##  0.0000000 0.1016645
## sample estimates:
##          p
## 0.09359857
```

V.

Q: Construct a 90% CI for the true proportion of spam. Interpret in context of the problem.

```
prop.test(x=367, n=3921, p=.1, conf.level=.90, alternative="two.sided")

##
##  1-sample proportions test with continuity correction
##
## data:  367 out of 3921, null probability 0.1
## X-squared = 1.7149, df = 1, p-value = 0.1904
## alternative hypothesis: true p is not equal to 0.1
## 90 percent confidence interval:
##  0.08610245 0.10166449
## sample estimates:
##          p
## 0.09359857
```

Example 5.3: Proportion of spam emails - full data available

Consider the `email` data set that contains information about the contents in email, and whether or not the email was flagged as spam.

```
table(email$spam)

##
##    0    1
## 3554  367
```

Since the variable `spam` is coded as a 0/1 binary indicator variable, we can take advantage of the relationship between the mean of a binary variable and the proportion. We can use the `t.test()` function like we did in the previous section. For large sample sizes, we do not need the continuity correction, so set `correct=FALSE`.

```
t.test(email$spam, mu=.1, alternative="two.sided", correct=FALSE)

##
##  One Sample t-test
##
## data:  email$spam
## t = -1.376, df = 3920, p-value = 0.1689
## alternative hypothesis: true mean is not equal to 0.1
## 95 percent confidence interval:
##  0.08447775 0.10271940
## sample estimates:
##  mean of x
## 0.09359857
```

There is insufficient reason to believe that the proportion of emails flagged as spam is not equal to 10%, $\hat{p}$ =_____, 95% CI: (_____, _____).

Chapter 6

Bivariate Inference

So far we have been concerned with making inference about a single population parameter. Many problems deal with comparing a parameter across two or more groups. Research questions include questions like:

- Does the average life span differ across subspecies of a particular turtle?
- Who has a higher percentage of the female vote - Democrats or Republicans?

Table 6.1 shows which statistical analyses procedures are appropriate depending on the combination of explanatory and response variable.

Table 6.1: Choosing Appropriate Statistical Analysis Procedures.

	Response	
Explanatory	**Binary**	**Quantitative**
Binary	Chi-squared	T-Test
Categorical	Chi-squared	ANOVA
Quantitative	Logistic Regression	Linear Regression and Correlation

Sometimes these variable types are referred to using the first letter, e.g. **Q** for quantitative, **B** for binary, and **C** for categorical. Thus a T-test is a (Q $\sim$ B) analysis, and a correlation analysis is (Q $\sim$ Q) analysis.

Since this chapter is about comparing two variables (bivariate analysis), we will leave Logistic Regression for later in the Multivariable analysis chapter.

The primary assumption of most standard statistical procedures is that the data are independent of each other. However, there are many examples where measurements are made on subjects before and after a certain exposure or treatment (pre-post), or an experiment to compare two cell phone packages might use pairs of subjects that are the same age, sex and income level. One subject would be randomly assigned to the first phone package, the other in the pair would get the second phone package. But for the purposes of this class, we will only concern ourselves with independent groups.

6.1 2-sample T-test for a difference in means (Q ∼ B)

Reading Assignment

OpenIntro Section 5.3

It is common to compare means from different samples. For instance, we might investigate the effectiveness of a certain educational intervention by looking for evidence of greater reading ability in the treatment group against a control group. That is, our research hypothesis is that reading ability of a child is associated with an educational intervention.

The null hypothesis states that there is no relationship, or no effect, of the educational intervention (binary explanatory variable) on the reading ability of the child (quantitative response variable). This can be written in symbols as follows:

$$H_0 : \mu_1 = \mu_2, \text{ which can be written as a difference:} \qquad H_0 : \mu_1 - \mu_2 = 0$$

where μ_1 is the average reading score for students in the control group (no intervention) and μ_2 be the average reading score for students in the intervention group.

The alternative hypothesis H_A states that there is a relationship:

$$H_A : \mu_1 \neq \mu_2 \qquad \text{or} \qquad H_A : \mu_1 - \mu_2 \neq 0$$

Assumptions

- The data distribution for each group is approximately normal.
- The scores are independent within each group.
- The scores from the two groups are independent of each other (i.e. the two samples are independent).

Sampling Distribution for the difference

We use $\bar{x}_1 - \bar{x}_2$ as a point estimate for $\mu_1 - \mu_2$, which has a standard error of

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{SE_{\bar{x}_1}^2 + SE_{\bar{x}_2}^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{6.1}$$

So the equations for the CI (6.2) and test statistic (6.3) then look like:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, df} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{6.2}$$

$$t^* = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\left(\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)} \tag{6.3}$$

Typically it is unlikely that σ_1^2 and σ_2^2 are known so we will use s_1^2 and s_2^2 as estimates.

Example 6.1: Smoking and BMI

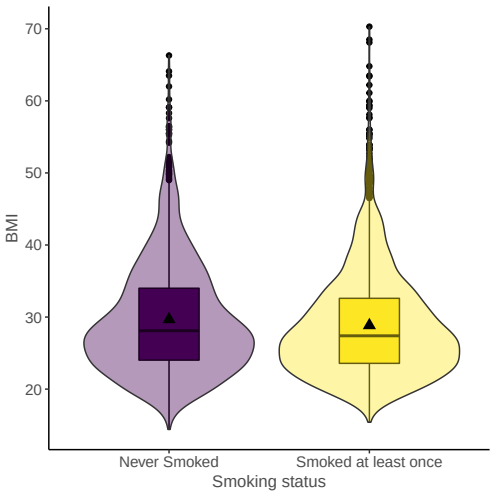
We would like to know, is there convincing evidence that the average BMI differs between those who have ever smoked a cigarette in their life compared to those who have never smoked? This example uses the Addhealth dataset that was introduced in Chapter 2.

1. Identify response and explanatory variables.

- The quantitative response variable is BMI (variable `BMI`)
- The binary explanatory variable is whether the person has ever smoked a cigarette (variable `eversmoke_c`)

2. Visualize and summarize bivariate relationship.

```
plot.bmi.smoke <- addhealth %>%
  select(eversmoke_c, BMI) %>%
  na.omit()

ggplot(plot.bmi.smoke,
       aes(x=eversmoke_c,
           y=BMI,
           fill=eversmoke_c)) +
     geom_boxplot(width=.3) +
     geom_violin(alpha=.4) +
     labs(x="Smoking status") +
     stat_summary(fun.y="mean",
                  geom="point",
                  size=3, pch=17,
     position=position_dodge(width=0.75))
```




```
plot.bmi.smoke %>% group_by(eversmoke_c) %>%
 summarise(mean=mean(BMI, na.rm=TRUE),
             sd = sd(BMI, na.rm=TRUE),
             IQR = IQR(BMI, na.rm=TRUE))

## # A tibble: 2 x 4
##   eversmoke_c           mean    sd   IQR
##   <fct>                <dbl> <dbl> <dbl>
## 1 Never Smoked          29.7  7.76  9.98
## 2 Smoked at least once  28.8  7.32  9.02
```

Smokers have an average BMI of 28.8, smaller than the average BMI of non-smokers at 29.7. Nonsmokers have more variation in their BMIs (sd 7.8 v. 7.3 and IQR 9.98 vs 9.02), but the distributions both look normal, if slightly skewed right.

3. Write the relationship you want to examine in the form of a research question.

- Null Hypothesis: There is no difference in the average BMI between smokers and nonsmokers.
- Alternate Hypothesis: There is a difference in the average BMI between smokers and nonsmokers.

4. Perform an appropriate statistical analysis.

I. Let μ_1 denote the average BMI for nonsmokers, and μ_2 the average BMI for smokers.

II. $\mu_1 - \mu_2 = 0$ There is no relationship between BMI and smoking

$\mu_1 - \mu_2 \neq 0$ There is a relationship between BMI and smoking

III. We are comparing the means between two independent samples. A Two-Sample T-Test for a difference in means will be conducted. The assumptions that the groups are independent is upheld because each individual can only be either a smoker or nonsmoker. The difference in sample means $\bar{x_1} - \bar{x_2}$ is normally distributed – this is a valid assumption due to the large sample size and that differences typically are normally distributed. The observations are independent, and the variability is roughly equal.

IV. We use the `t.test` function, but use model notation of the format `outcome` $\sim$ `category`. Here, `BMI` is our continuous outcome that we're testing across the (binary) categorical predictor `eversmoke_c`.

```
t.test(BMI ~ eversmoke_c, data=addhealth)

##
##  Welch Two Sample t-test
##
## data:  BMI by eversmoke_c
## t = 3.6937, df = 3395.3, p-value = 0.0002245
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3906204 1.2744780
## sample estimates:
##         mean in group Never Smoked mean in group Smoked at least once
##                           29.67977                           28.84722
```

Reject the null hypothesis, $p = 0.0002$.

5. Write a conclusion in the context of the problem. On average, nonsmokers have a significantly higher BMI by 0.83 (0.39, 1.27) compared to nonsmokers ($p = 0.0002$).

Always check the output against the direction you are testing. `R` always will calculate a difference as group 1 - group 2, and it defines the groups alphabetically. For

example, for a factor variable that has gender Female/Male, R will automatically calculate the difference as Female - Male. In this example it is Nonsmoker - Smoker.

6.1.1 Interpreting stratified CI's

The standard t-test gives us a confidence interval for the difference, and if that interval covers 0 then there is no reason to believe that there is a significant difference between group means. However, how does this compare to the confidence intervals for the mean of each group?

Let's use `dplyr` to calculate one-sample confidence intervals for each group separately. *Code comment: The code below reads: "Take the `plot.bmi.smoke` data table which was created in a previous example, convert it to a data frame and then `filter` on the nonsmoker group. Then extract the variable `BMI`, run a t-test on it, save as an object and then print only the `conf.int` values. Repeat for the smoker group.*

```
nonsmoker <- as.data.frame(plot.bmi.smoke) %>%
             filter(eversmoke_c=="Never Smoked") %>%
             select(BMI) %>% t.test()
smoker <- as.data.frame(plot.bmi.smoke) %>%
          filter(eversmoke_c=="Smoked at least once") %>%
          select(BMI) %>% t.test()

nonsmoker$conf.int[1:2]

## [1] 29.31575 30.04380

smoker$conf.int[1:2]

## [1] 28.59647 29.09797
```

Note that they don't overlap. This is another indication that the means are significantly different from each other. This is not a two way assumption! Overlapping means DO NOT mean that there is no difference. You cannot make a conclusion one way or another if they overlap!

6.2 Analysis of Variance (ANOVA) (Q ~ C)

Reading Assignment

OpenIntro Section 5.5

Frequently, a researcher wants to compare the means of an outcome across three or more treatments in a single experiment. We might initially think to do pairwise comparisons (1v2, 1v3, 2v3) for a total of three comparisons. However, this strategy can be treacherous. If we have many groups and do many comparisons, it is likely that we will eventually find a difference just by chance, even if there is no difference in the populations.

When we analyze a conventional two-treatment experiment, we are prepared to run a 1 in 20 risk of an apparently significant result arising purely by accident (the 5% chance of a Type I error). We regard such a risk as being fairly unlikely and feel justified in accepting with confidence any significant results we obtain.

Analyzing a single experiment as a series of 10 treatment pairs is a very different proposition. The chance of an apparently significant result arising purely by chance somewhere in the 10 analyses increases dramatically. Using a 5% error rate, the chance of NOT making at Type I error is .95. To not make a Type I error 10 times is $.95^{10} = .6$. That means there is a 40% of making a Type I error! *See:* `https://xkcd.com/882/`.

Example 6.2: Visual Comparison

Examine Figure 6.1. Compare groups I, II, and III. Can you visually determine if the differences in the group centers is due to chance or not? What about groups IV, V, and VI?

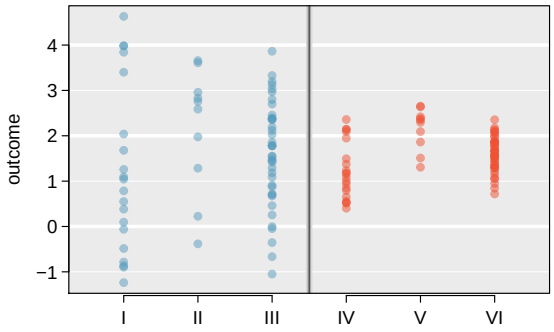


Figure 6.1: Side-by-side dot plot for the outcomes for six groups.

So we need some method of comparing treatments for more than two groups at a time.

H_0: The mean outcome is the same across all groups. $\mu_1 = \mu_2 = \cdots = \mu_k$

H_A: At least one mean is different.

You may not think that all k population means are equal, but if they don't test to be statistically significantly different, then the differences are small enough to be ignored.

Generally we must check three conditions on the data before performing ANOVA:

- the observations are independent within and across groups,
- the data within each group are nearly normal, and
- the variability across the groups is about equal.

When these three conditions are met, we may perform an ANOVA to determine whether the data provide strong evidence against the null hypothesis that all the μ_i are equal.

Terminology

- Response Variable: The response variable in the ANOVA setting is the quantitative (continuous) variable that we want to compare among the different treatments.
- Factor/Treatment: A property or characteristic (categorical variable) that allows us to distinguish the different populations from one another. An independent variable to be studied in an investigation such as temperature, type of plant, color of flower, location.
- Factor/Treatment level: Factors have different levels, such as 3 temperatures, 5 locations, 3 colors, etc.
- Within-sample Variation: Variation within a sample from one population. Individuals who receive the same treatment will experience identical experimental conditions. The variation within each of the treatment groups must therefore be a consequence of solely random variation. Random variation (always)
- Between-sample Variation: Variation between samples. This is the difference between the group means. If some treatments are genuinely more effective than others, then

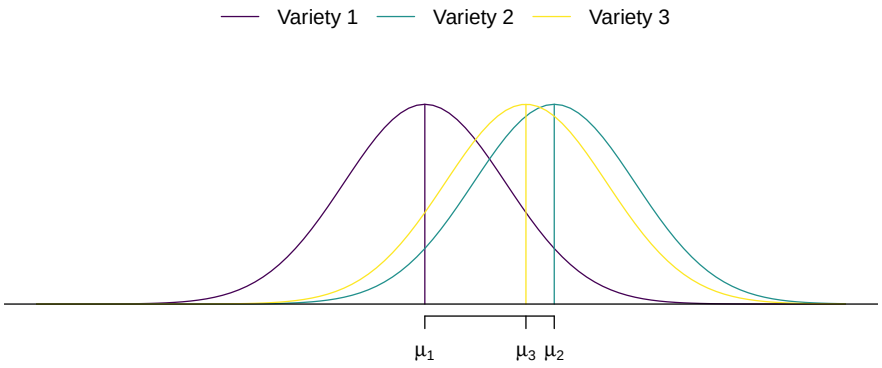
we would expect to see relatively large differences between the treatment means and a relatively large between-treatments variation. Random variation (always) + imposed variation (maybe)

Example 6.3: Phosphorous content of tree leaves

A horticulturist was investigating the phosphorous content of tree leaves from three different varieties of apple trees (Variety 1, Variety 2 and Variety 3). Random samples of five leaves from each of the three varieties were analyzed for phosphorous content.

1. What are the factors? What are their levels?
2. What is the response variable?

Typically in analysis of variance we are interested in two levels of inference. We are interested in whether there are ANY differences among the mean among several groups, and if so, where do those differences occur.



6.2.1 Formulation of the One-way ANOVA model

ANOVA is a mathematical technique which uses a model based approach to partition the variance in an experiment into different sources of variance. This technique enables us to test if most the variation in the treatment means is due to differences between the groups. The one-way ANOVA model is

$$y_{ij} = \mu_i + \epsilon_{ij} \qquad \epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \tag{6.4}$$

for $i = 1, \ldots, I$ factor levels and $j = 1, \ldots, n_i$ subjects within each factor level. The random error terms are independently and identically distributed (iid) normally with common variance. A good way to think about this is that the observed data comes from some true model with some random error.

$$\text{DATA} = \text{MODEL FIT} + \text{RESIDUAL},$$

A commonly used shorthand model notation is to write that y is a function of some (can be explained by) the group that data point belongs to. So the MODEL is just the group membership. We will soon see different forms of this model, but this is where we start.

$$y = group + \epsilon \tag{6.5}$$

The fit of the ANOVA model as being broken down into 2 parts

$$\text{Total Variation} = \text{Between Group Variation} + \text{Within Group Variation}$$

Variation is measured using the Sum of Squares (SS): The sum of the squares within a group (SSE), the sum of squares between groups (SSG), and the total sum of squares (SST).

- SSG: Measures the variation of the I group means around the overall mean.

$$SSG = \sum_{i=1}^{I} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 = n_1(\bar{y}_{1.} - \bar{y}_{..})^2 + n_2(\bar{y}_{2.} - \bar{y}_{..})^2 + n_3(\bar{y}_{3.} - \bar{y}_{..})^2 \tag{6.6}$$

- SSE: Measures the variation of each observation around its group mean.

$$SSE = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 = \sum_{i=1}^{I}(n_i - 1)Var(Y_i) \tag{6.7}$$

- SST: Measures the variation of the N data points around the overall mean.

$$SST = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 = (N - 1)Var(Y) \tag{6.8}$$

Analysis of Variance Table: The results of an analysis of variance test are always summarized in an ANOVA table. The format of an ANOVA table is as follows:

Source	SS	DF	MS	F-Test
Groups	SSG	$I-1$	SSG/df_G	MSG/MSE
Error	SSE	$N-I$	SSE/df_e	
Total	SST	$N-1$	-	

where N is the total number of observations, and I is the number of groups. Each sum of squares has an associated degree of freedom that is also a breakdown of the total DF for the model. Let σ_τ^2 be the variance of the means due to the group treatment. It can be shown that

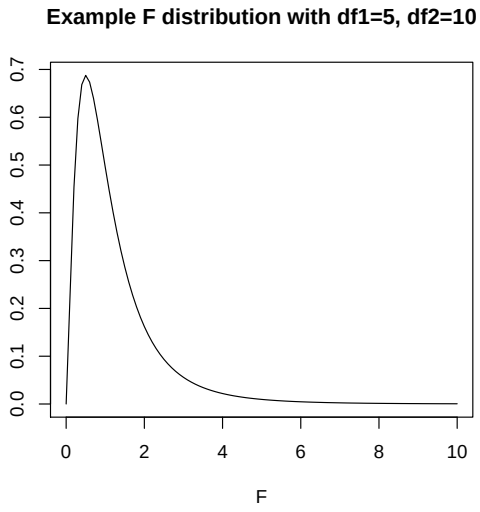
$$E(MSE) = \sigma^2$$
$$E(MSG) = \sigma^2 + n\sigma_\tau^2$$

So MSG and MSE both estimate the overall error variance σ^2. The ANOVA test compares the amount of variation between groups (MSG) to the amount of variation within groups (SSE) by calculating the test statistic $F = \frac{MSG}{MSE}$ and comparing this test statistic to a theoretical F-distribution.

If H_0 is true there is no group effect on the means and $\sigma_\tau^2 = 0$ and F is small, resulting in a large p-value. Alternatively if H_0 is false, then there is a group effect, $n\sigma_\tau^2 > 0$, and F is large and the resulting p-value is small. When H_0 is true, the F statistic has an F distribution with $I-1$ numerator degrees of freedom and $N-1$ denominator degrees of freedom. Just like how each t-distribution is different depending on its df, the shape of a F-distribution is completely determined by the numerator and denominator degrees of freedom. So "large" is subjective and dependent on the degrees of freedom.

The F-distribution:

The p-value of the test is the **area to the right** of the F statistic density curve. This is always to the right because the F-distribution is not symmetric, truncated at 0 and skewed right. This is true regardless of the df.



6.2.2 Example: Amount of nitrogen across plant species

Example 6.4: A comparison of plant species under low water conditions

The `PLANTS1` data file gives the percent of nitrogen in four different species of plants grown in a laboratory. The researchers collected these data in parts of the country where there is very little rainfall. To examine the effect of water, they varied the amount per day from 50mm to 650mm in 100mm increments. There were 9 plants per species-by-water combination. Because the plants are to be used primarily for animal food, with some parts that can be consumed by people, a high nitrogen content is very desirable.

1. What is the response variable?

2. What are the explanatory variables (factors) and levels of each?

Let's formally test to see if the nitrogen content in the plants differ across species. First we need to ensure that `species` is being treated as a factor variable and not numeric (because it was entered into the data set as simple numbers).

```
plants1$species <- as.factor(plants1$species)
```

I. Let $\mu_1, \ldots, \mu_4$ be the mean nitrogen content in plant species 1 through 4 respectively.

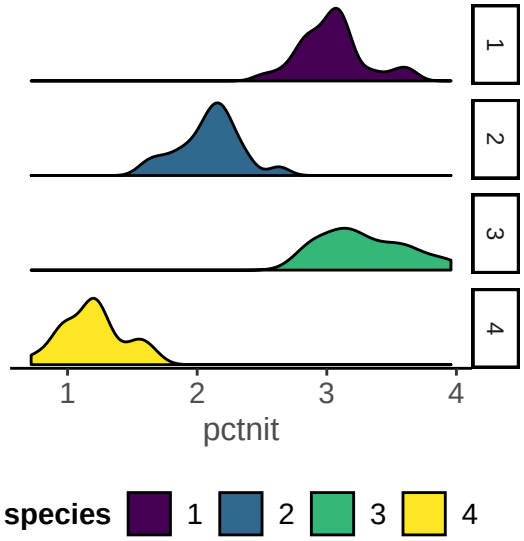
II. $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ $\quad$ H_A : At least one mean is different.

III. We are comparing means from multiple groups, so an ANOVA is the appropriate procedure. We need to check for independence, approximate normality and approximately equal variances across groups.

Independence: We are assuming that each plant was sampled independently of each other, and that the species themselves are independent of each other.

Normality: With grouped data it's easier to look at the histograms than qqplots.

```
ggplot(plants1,
       aes(x=pctnit,
           fill=species)) +
    geom_density() + ylab("")+
    facet_grid(species~.) +
    theme(legend.position="bottom")
    scale_y_continuous(breaks=NULL)
```



The distributions per group tend to follow an approximate normal distribution.

Equal variances: The quickest way to assess if the groups have approximately equal variances is by comparing the IQR across groups.

```
plants1 %>%
  group_by(species) %>%
  summarise(IQR(pctnit))
```

```
## # A tibble: 4 x 2
##   species `IQR(pctnit)`
##   <fct>           <dbl>
## 1 1               0.269
## 2 2               0.272
## 3 3               0.506
## 4 4               0.312
```

The IQRs are similar so assumption of equal variances is not grossly violated. We can proceed with the ANOVA procedure.

IV. We use the `aov(response ~ predictor)` function on the relationship between the phosphorous levels and tree variety. We then pipe in `summary()` to make the output display nicely.

```
aov(pctnit~species, data=plants1) %>% summary()

##              Df Sum Sq Mean Sq F value Pr(>F)
## species       3 172.39   57.46   827.5 <2e-16 ***
## Residuals   248  17.22    0.07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

V. The results of the ANOVA test indicate that at least one species has a different average nitrogen content than the other varieties ($p < .001$).

6.2.3 Coefficient of determination R^2

The coefficient of determination is defined as $R^2 = {}^{SSG}/_{SST}$ and can be interpreted as the % of the variation seen in the outcome that is due to subject level variation within each of the treatment groups. The strength of this measure can be thought of in a similar manner as the correlation coefficient r: $< .3$ indicates a poor fit, $< .5$ indicates a medium fit, and $> .7$ indicates a good fit.

```
172.39/(172.39+17.22)*100

## [1] 90.9182
```

A large amount (91%) of the variation seen in nitrogen content in the plant can be explained by the species of plant.

6.2.4 Multiple Comparisons

Suppose that an ANOVA test reveals that there is a difference in at least one of the means. How can we determine which groups are significantly different without increasing our chance of a Type I error?

Simple! We perform all the pairwise comparisons but using a test statistic that retains a **family-wise error rate** of 0.05 (or our chosen α). There are different methods to adjust for multiple comparisons, we will be using the **Tukey HSD (honest significant difference) test**. Continuing on with the analysis of nitrogen across plant species.

```
TukeyHSD(aov(pctnit~species, data=plants1))

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = pctnit ~ species, data = plants1)
##
## $species
##           diff        lwr        upr   p adj
## 2-1 -0.9469683 -1.0684156 -0.8255209 0.0e+00
## 3-1  0.2445556  0.1231082  0.3660029 2.4e-06
## 4-1 -1.8442222 -1.9656696 -1.7227748 0.0e+00
## 3-2  1.1915238  1.0700764  1.3129712 0.0e+00
## 4-2 -0.8972540 -1.0187014 -0.7758066 0.0e+00
## 4-3 -2.0887778 -2.2102252 -1.9673304 0.0e+00
```

The results from Tukey's HSD for all pairwise comparisons indicate that the average nitrogen content in one species is significantly different from each of the three other species. The nice benefit of this procedure is that the difference between the means of the two groups are compared, and a 95% confidence interval for each difference is included. So specifically, species 2 has on average 0.94 (0.82, 1.09) lower percent nitrogen compared to species 1 ($p < .0001$).

6.2.5 Example: Fisher's Irises

Example 6.5: Fisher's irises

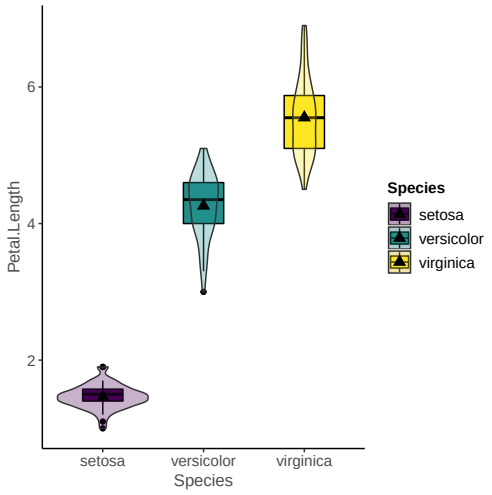
We want to know if there is a relationship between an iris flower's petal length and the species of the flower. We can answer this question using the statistician Ronald Fisher's iris data set. This data set is so special, it's built into `R` as the data frame `iris`.

1. Identify response and explanatory variables.

- The categorical explanatory variable is the species of flower (variable `species`)
- The quantitative response variable is the length of the petal (variable `Petal.Length`)

2. Visualize and summarise bivariate relationship.

```
ggplot(iris, aes(x=Species,
                 y=Petal.Length,
                 fill=Species)) +
  geom_boxplot(width=.4) +
  geom_violin(alpha=.3) +
  stat_summary(fun.y="mean",
               geom="point",
               size=3, pch=17,
               position=position_dodge(
                 width=0.75))
```




```
iris %>% group_by(Species) %>%
         summarise(mean=mean(Petal.Length),
                   sd=sd(Petal.Length),
                   n=n())

## # A tibble: 3 x 4
##   Species     mean    sd     n
##   <fct>      <dbl> <dbl> <int>
## 1 setosa      1.46 0.174    50
## 2 versicolor  4.26 0.470    50
## 3 virginica   5.55 0.552    50
```

There are clear differences in the average petal length across iris species. *Iris setosa* has an average petal length of 1.5 cm (sd 0.17), *I. versicolor* has an average petal length of 4.3 cm (sd 0.50), and *I. virginica* has the largest average petal length of 5.6 cm (sd 0.55).

3. Write the relationship you want to examine in the form of a research question.

Is there a relationship between the petal length of an iris flower and the species of flower?

- Null Hypothesis: There is no relationship between petal length and species.
- Alternate Hypothesis: There is a relationship between petal length and species.

4. Perform an appropriate statistical analysis.

I. Let μ_1 be the true mean petal length for *I. setosa*.

Let μ_2 be the true mean petal length for *I. versicolor*.

Let μ_3 be the true mean petal length for *I. virginica*.

II. $H_0 : \mu_1 = \mu_2 = \mu_3$

H_A : At least one group mean is different.

III. I will conduct an analysis of variance using ANOVA. The distribution of petal length looks approximately normal within each species group. We can assume that the group means are normally distributed due to the sample size within each group $n = 50$ being large enough for the CLT to hold. The assumption of equal variances may be violated here; the sd of *I. setosa* is less than half that of the other two species.

IV.

```
  aov(Petal.Length ~ Species, data=iris) %>% summary()

##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  437.1  218.55    1180 <2e-16 ***
## Residuals   147   27.2    0.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Reject the null hypothesis; the p-value is less than .0001.

Since the ANOVA was significant, I need to conduct a post-hoc test to identify which pairs are different.

```
  aov(Petal.Length ~ Species, data=iris) %>% TukeyHSD()

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Petal.Length ~ Species, data = iris)
##
## $Species
##                       diff     lwr     upr p adj
## versicolor-setosa    2.798 2.59422 3.00178     0
## virginica-setosa     4.090 3.88622 4.29378     0
## virginica-versicolor 1.292 1.08822 1.49578     0
```

All pairs are significantly different from each other. The p-values for post-hoc tests are all less than .0001.

5. Write a conclusion in context of the problem.

There is sufficient evidence to conclude that the average petal length of an iris flower is associated with the species of the iris ($p < .0001$). Specifically the length of the petal for species *I. virginica* is 4.1 (95% CI 3.9, 4.3) cm longer than *I. setosa*, and 1.3 (95%CI 1.1, 1.5) cm longer than *I. versicolor*. *I. versicolor* is also significantly longer than *I. setosa* (2.8, 95% CI 2.6, 3.0) cm. All pairwise comparisons were significant at the .0001 level.

6.3 χ^2 test of association (B $\sim$ C)

Reading Assignment

OpenIntro Section 6.2

We would like to make conclusions about the difference in two population proportions: $p_1 - p_2$. A reasonable point estimate based on the sample is $\hat{p}_1 - \hat{p}_2$. No surprise there.

Nearly always, we want to know if these proportions are equal, In other words we're testing the hypothesis that $p_1 - p_2 = 0$. In that case we use a **pooled proportion** to check the condition of normality, and to calculate the standard error of the estimate.

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} \tag{6.9}$$

where x_1 and x_2 represent the number of "successes" in each group and n_1 and n_2 represent the sample sizes for each group. Then the standard error of the point estimate is calculated as

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}} \tag{6.10}$$

So the equations for the CI (6.11) and test statistic (6.12) then look like:

$$(\hat{p}_1 - \hat{p}_2) \pm t_{\alpha/2, df} \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}} \tag{6.11}$$

$$t^* = \frac{(\hat{p}_1 - \hat{p}_2) - d_0}{\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}\right)} \tag{6.12}$$

Conditions for the sampling distribution to be normal

The difference $\hat{p}_1 - \hat{p}_2$ tends to follow a normal model when 1) each proportion separately follows a normal model, and 2) the two samples are independent of each other. #1 can be verified by checking the **success-failure condition** for each group. That means

- $\hat{p}n_1 \geq 10$, and
- $\hat{p}n_2 \geq 10$, and
- $\hat{q}n_1 \geq 10$, and
- $\hat{q}n_2 \geq 10$,

where, if I've forgotten to mention it yet, $q = 1 - p$.

Testing differences in proportions using R

Example 6.6: Are Mammograms effective? OI Section 6.3.2

A 30-year study was conducted with nearly 90,000 female participants. [a] During a 5-year screening period, each woman was randomized to one of two groups: in the first group, women received regular mammograms to screen for breast cancer, and in the second group, women received regular non-mammogram breast cancer exams. No intervention was made during the following 25 years of the study, and we'll consider death resulting from breast cancer over the full 30-year period. Results from the study are summarized in Table 6.2.

[a] Miller AB. 2014. *Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomized screening trial.* BMJ 2014;348:g366.

If mammograms are much more effective than non-mammogram breast cancer exams, then we would expect to see additional deaths from breast cancer in the control group. On the other hand, if mammograms are not as effective as regular breast cancer exams, we would expect to see an increase in breast cancer deaths in the mammogram group.

	Death from breast cancer?	
	Yes	No
Mammogram	500	44,425
Control	505	44,405

Table 6.2: Summary results for breast cancer study.

```
prop.test(x=c(500, 505), n=c(500+44425, 505+44405), correct=FALSE)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(500, 505) out of c(500 + 44425, 505 + 44405)
## X-squared = 0.026874, df = 1, p-value = 0.8698
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.001490590  0.001260488
## sample estimates:
##     prop 1     prop 2
## 0.01112966 0.01124471
```

An astute student would notice that `prop.test` reports something called the `X-squared`, pronounced *chi-squared*, text statistic, and that the p-value is *slightly* off from the one reported in the textbook. Statistical theory that is outside the scope of this course provides the proof that using χ^2 distribution to test for a difference in proportions is equivalent to using the normal model **as long as the normal conditions apply**.

6.3.1 Example: Smoking and General Health

Example 6.7: Smoking and general health

Using the Addhealth data set, what can we say about the relationship between smoking and a person's perceived general level of general health?

1. Identify response and explanatory variables.

- The binary explanatory variable is whether the person has ever smoked an entire cigarette (variable `eversmoke_c`)
- The categorical explanatory variable is the person's general health (variable `genhealth`) and has levels "Excellent", "Very Good", "Good", "Fair", and "Poor".

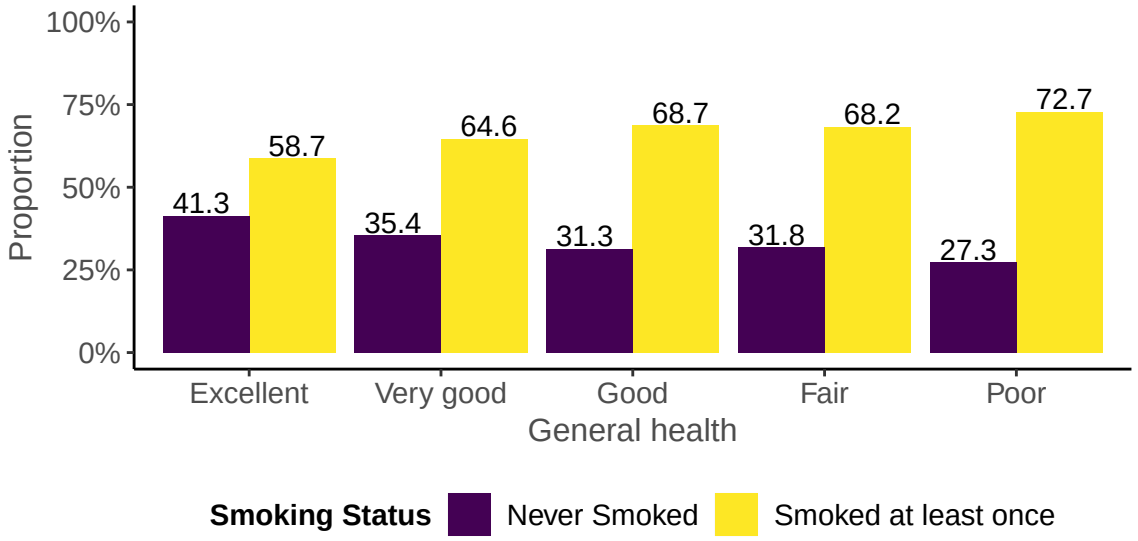
2. Visualize and summarise bivariate relationship. First we create the summary tables of frequencies and proportions.

```
table(addhealth$eversmoke_c, addhealth$genhealth)
```

	Excellent	Very good	Good	Fair	Poor
Never Smoked	403	694	525	136	15
Smoked at least once	573	1265	1154	292	40

```
table(addhealth$eversmoke_c, addhealth$genhealth) %>% prop.table(margin=2)
```

	Excellent	Very good	Good	Fair	Poor
Never Smoked	0.413	0.354	0.313	0.318	0.273
Smoked at least once	0.587	0.646	0.687	0.682	0.727



The percentage of smokers seems to increase as the general health status decreases. Almost three-quarters (73%) of those reporting poor health have smoked an entire cigarette at least once in their life compared to 59% of those reporting excellent health.

3. Write the relationship you want to examine in the form of a research question.

Is the proportion of those who have ever smoked equal across all levels of general health?

- Null Hypothesis: The proportion in each general health category is the same.
- Alternate Hypothesis: At least one proportion is different.

4. Perform an appropriate statistical analysis.

I. Let p_i be the true proportion of lifetime smokers within in general health category i.

II. $H_0 : p_1 = p_2 = p_3 = p_4 = p_5$

H_A : At least one proportion is different.

III. I will conduct a χ-squared test of association. There is at least 5 observations in each combination of smoking status and general health.

IV.

```
chisq.test(addhealth$genhealth, addhealth$eversmoke_c)
```

```
##
##  Pearson's Chi-squared test
##
## data:  addhealth$genhealth and addhealth$eversmoke_c
## X-squared = 30.795, df = 4, p-value = 3.371e-06
```

Reject the null; the p-value is less than .0001.

Since the test was significant, I will conduct a post-hoc test to identify which pairs are different. The `RVAideMemoire` package provides the `fisher.multcomp` function to conduct post-hoc pairwise comparisons after a significant chi-squared test of association. See the `RVAideMemoire` package reference (`https://rdrr.io/cran/RVAideMemoire/`) for more information on how to use this function. Of the many methods to control for multiple comparisons, the default is to use the *fdr* or "false discovery rate".

```
RVAideMemoire::fisher.multcomp(table(addhealth$genhealth, addhealth$eversmoke_c),
                               p.method="fdr")
```

```
##
##         Pairwise comparisons using Fisher's exact test for count data
##
## data:  table(addhealth$genhealth,addhealth$eversmoke_c)
##
##           Excellent Very good   Good   Fair
## Very good 6.965e-03         -      -      -
## Good      2.047e-06   0.02062      -      -
## Fair      4.163e-03   0.26903 0.8611      -
## Poor      9.495e-02   0.36087 0.7303 0.6753
##
## P value adjustment method: fdr
```

Hint: Put the categorical variable first in the table.

The p-values for all tests comparing Excellent to other groups are all significant at $< .01$. The p-value comparing Very Good to Good is also ($p = 0.02$).

5. Write a conclusion in context of the problem. We can conclude that there is an association between ever smoking a cigarette in their life and perceived general health ($\chi^2 = 30.8$, df=4, $p < .0001$). The proportion of those who have smoked at least one cigarette in their life in the Excellent health group (59%) is significantly lower than any other group (65% for Very Good to 73% for Poor).

6.4 Correlation (Q $\sim$ Q)

Recall from Section 2.3.3 the definition of **correlation** between two continuous variables. The **correlation coefficient** is designated by r for the sample correlation, and ρ for the population correlation. The correlation is a measure of the strength and direction of a linear relationship between two variables.

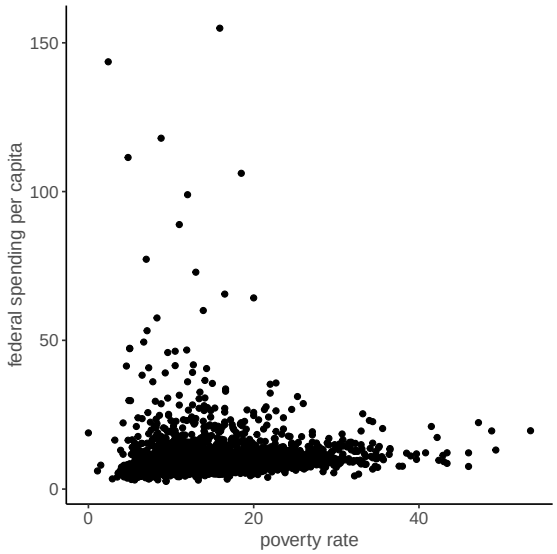
The correlation ranges from +1 to -1. A correlation of +1 means that there is a perfect, positive linear relationship between the two variables. A correlation of -1 means there is a perfect, negative linear relationship between the two variables. In both cases, knowing the value of one variable, you can perfectly predict the value of the second.

Here are rough estimates for interpreting the strengths of correlations based on the magnitude of r.

- $|r| \geq 0.7$: Very strong relationship
- $0.4 \leq |r| < 0.7$: Strong relationship
- $0.3 \leq |r| < 0.4$: Moderate relationship
- $0.2 \leq |r| < 0.3$: Weak relationship
- $|r| < 0.2$: Negligible or no relationship

For example, let's return to the example of federal spending per capita and poverty rate in the `county` dataset introduced in Chapter 2.

```
ggplot(county,
       aes(x=poverty,
           y=fed_spend00)) +
  geom_point() +
  ylab("federal spending per capita") +
  xlab("poverty rate")
```




```
cor(county$poverty, county$fed_spend00, use="complete.obs")

## [1] 0.03484461
```

There is a negligible, positive, linear relationship between poverty rate and per capita federal spending ($r = 0.03$). Let ρ denote the true correlation between poverty rate and federal spending per capita. Our null hypothesis is that there is no correlation between poverty rate and federal spending ($\rho = 0$), and the alternative hypothesis is that they are correlated ($\rho \neq 0$). We can use the `cor.test()` function to analyze the evidence in favor of this alternative hypothesis.

```
cor.test(county$poverty, county$fed_spend00)

##
##  Pearson's product-moment correlation
##
## data:  county$poverty and county$fed_spend00
## t = 1.9444, df = 3110, p-value = 0.05194
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.0002922843  0.0698955658
## sample estimates:
##        cor
## 0.03484461
```

We conclude from this that there was a non-statistically significant, negligible correlation between poverty and federal spending ($r = 0.03(-0.0003, .069), p = 0.05$).

6.4.1 Example: Fisher's Irises II

Example 6.8: Fisher's irises II

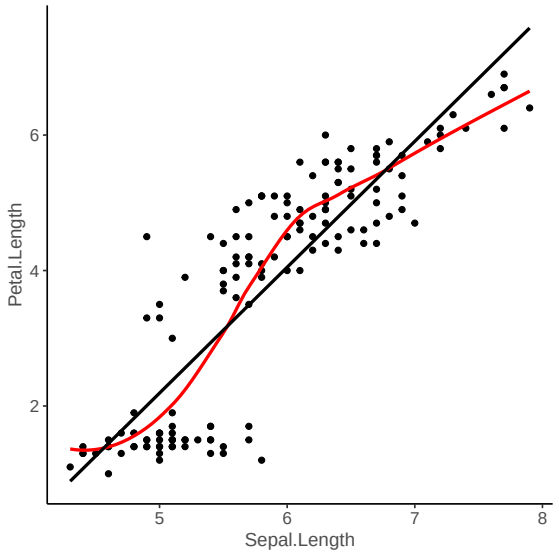
We would like to know if there is a correlation between the length of the sepal of an iris flower and the length of the flower.

1. Identify response and explanatory variables.

- The quantitative explanatory variable is the sepal length (variable `Sepal.Length`)
- The quantitative response variable is the petal length (variable `Petal.Length`)

2. Visualize and summarise bivariate relationship.

```
ggplot(iris, aes(x=Sepal.Length,
                 y=Petal.Length)) +
  geom_point() +
  geom_smooth(se=FALSE, col="red") +
  geom_smooth(method="lm", col="black",
              se=FALSE)
```




```
cor(iris$Sepal.Length, iris$Petal.Length)

## [1] 0.8717538
```

There is a strong, positive, linear relationship between the sepal length of the flower and the petal length ($r = 0.87$).

3. Write the relationship you want to examine in the form of a research question.

Is there a correlation between the sepal length of an iris flower and the petal length?

- Null Hypothesis: There is no correlation between length of sepal and petal.
- Alternate Hypothesis: Sepal and petal lengths are correlated.

4. Perform an appropriate statistical analysis.

I. Let ρ be the true correlation between sepal and petal length.

II. $H_0 : \rho = 0$

$H_A : \rho \neq 0$

III. Both variables are quantitative, so a correlation analysis will be conducted.

IV.

```
cor.test(iris$Petal.Length, iris$Sepal.Length)

##
##  Pearson's product-moment correlation
##
## data:  iris$Petal.Length and iris$Sepal.Length
## t = 21.646, df = 148, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8270363 0.9055080
## sample estimates:
##       cor
## 0.8717538
```

Reject the null hypothesis; the p-value is $< .0001$.

5. Write a conclusion in context of the problem.

There was a statistically significant and very strong correlation between the sepal length of an iris and the petal length, $r(148) = 0.87$, $p < .0001$. The significant positive correlation shows that as the sepal length increases, so does the petal length. These results suggest that 76% (95% CI: 68.9-82.8) of the variance in petal length can be explained by the length of the sepal.

6.5 Linear Regression (Q ~ Q)

Reading Assignment

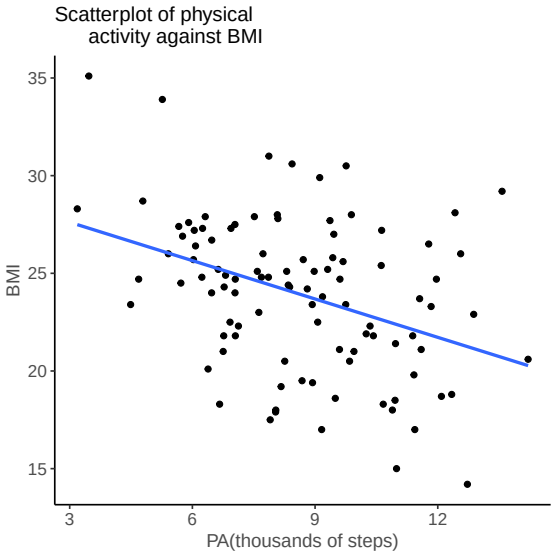
OpenIntro Chapter 7

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. Regression analysis is used to describe the distribution of values of the response variable y as a function of the other explanatory variables x_i. Linear models can be used for prediction or to evaluate whether there is a linear relationship between two numerical variables.

When only one explanatory variable x is considered, this is termed **simple linear regression** (SLR). The basic idea of SLR is to use data to fit a straight line that relates the response Y to the predictor X. The blue line in the scatterplot represents this "best fit" or linear regression line.

Here is an example of the relationship between physical activity as measured in 1,000 of steps and BMI.

```
ggplot(bmi,
       aes(x=PA,
           y=BMI)) +
  geom_point() + ylab("BMI") +
  geom_smooth(se=FALSE, method="lm") +
  xlab("PA(thousands of steps)") +
  ggtitle("Scatterplot of physical
      activity against BMI")
```



This regression line is a mathematical relationship between the mean of the response variable and the explanatory variable.

Assuming the relationship is linear, we can write the model the population average value of the response variable μ_y as a linear function of x:

$$\mu_y = \beta_0 + \beta_1 x \tag{6.13}$$

The intercept parameter, β_0, represents where the line crosses the y-axis when $x = 0$. The slope parameter, β_1, represents the change in μ_y per 1 unit x.

However, we know that there is always random noise in real data (DATA = MODEL FIT + RESIDUAL) so we introduce a random error term, ϵ_i and assume the model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{6.14}$$

This model states that the random variable y to be made up of a predictable part (a linear function of x) and an unpredictable part (the random error, ϵ_i). The error (residual) term includes the effects of all other factors, known or unknown. In the example of BMI against physical activity, there are also other factors that are unaccounted for such as gender, income, diet, race, etc. The unaccounted information from extraneous variables gets absorbed into the error.

6.5.1 Least Squares Regression Estimation

The method of least squares is the most common method of fitting a straight line to two variables.

Using a sample of data we can calculate point estimates b_0 and b_1 (sometimes written as $\hat{\beta}_0$ and $\hat{\beta}_1$) to estimate the population parameter values β_0 and β_1 respectively. Then the estimated mean function is

$$\hat{y}_i = b_0 + b_1 x_i \tag{6.15}$$

The estimated value for point i, $\hat{y}_i$, is called the fitted value.

The difference between the observed and the fitted value is called the residual

$$\epsilon_i = y_i - \hat{y}_i. \tag{6.16}$$

The residual sum of squares (RSS a.k.a SSE) sums the squared distances from each observed value to each fitted value and is represented mathematically as $\sum(y_i - \hat{y}_i)^2$. Thus, it is a measure of how well the line fits the data. Least squares regression finds values for b_0 and b_1 that minimizes those squared residuals ϵ_i^2.

The point estimates b_1 and b_0 are calculated as

$$b_0 = \bar{y} - b_1\bar{x}, \qquad b_1 = r\frac{s_y}{s_x}, \tag{6.17}$$

where r is the correlation coefficient between x and y.

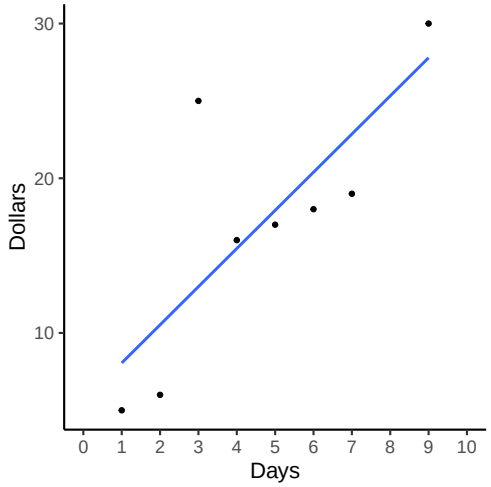
Example 6.9: Hospital expenses

Let's examine a fictitious data set that lists hospital expenses (in thousands of dollars) and the length of stay (days) in the hospital for a random sample of 8 patients.

```
# Make up data and put it into a data frame for analysis
insurance <- data.frame(Days = c(1,2,3,4,5,6,7,9),
                        Dollars = c(5, 6,25, 16, 17, 18, 19, 30))
```

1. First plot the data, and add a least-squares line.

```
ggplot(insurance,
       aes(x=Days, y=Dollars)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE)
```



2. How would you describe the association between length of stay and expenses covered by the insurance company?

3. Which is the response variable and which is the explanatory variable?

4. Draw the residual distances on the plot above.
5. Fit a linear regression model and report the regression equation.

```
cost <- lm(Dollars~Days, data=insurance)
cost

##
## Call:
## lm(formula = Dollars ~ Days, data = insurance)
##
## Coefficients:
## (Intercept)         Days
##       5.594        2.466
```

$$\hat{y} = \_\_\_\_\_ + \_\_\_\_\_ x$$

6. Interpret the intercept b_0 in context of the problem.

7. Interpret the slope b_1 in context of the problem.

8. What is the expected cost for a 8 day stay at a hospital? Plot this point on your graph.

9. Suppose person i stayed at a hospital for 8 days and was billed \$18,000. Plot the observed value on the graph and annotate the residual. What is the residual $\epsilon_i = (y_i - \hat{y}_i)$ for this value?

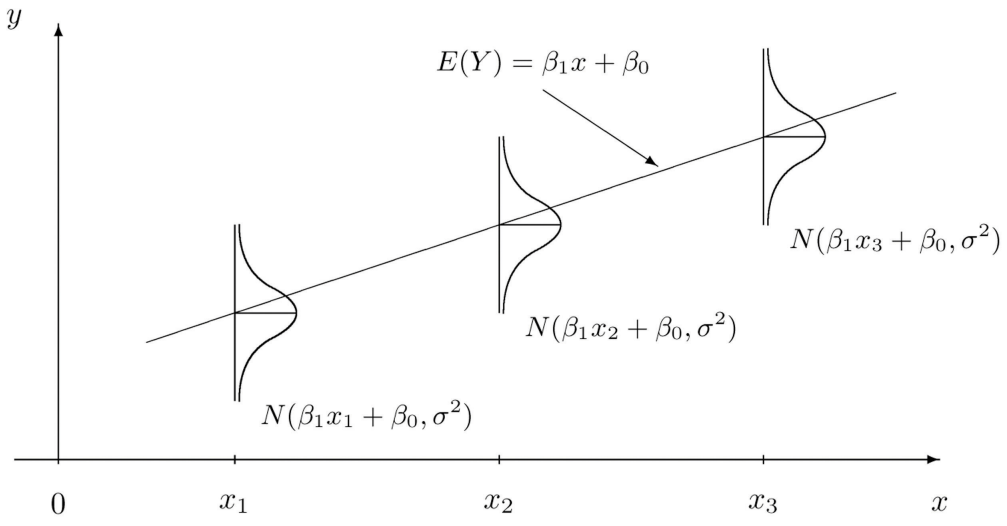
Facts about least-squares (LS) regression

1. A change of one standard deviation in x corresponds to a change of r standard deviations in y. Recall: $b_1 = r^{s_y}/_{s_x}$.
2. If the correlation is 0, the slope of the LS line is 0. A test of $\beta_1 = 0$ is equivalent to a test of $\rho = 0$.
3. The LS line always passes through the point $(\bar{x}, \bar{y})$.
4. The distinction between explanatory and response variables is essential in regression. Reversing x and y results in a different LS regression line.

6.5.2 Model Assumptions

This section presents some informal graphical tools for assessing the lack of fit. The assumptions for linear regression are:

- Linearity: The mean of the response variable y changes linearly as x changes.
- Independence: Each observation y_i is independent of all other $y_j, i \neq j$.
- Normality of Residuals: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- Constant Variance: Individual responses y with the same x vary according to a normal distribution with common constant variance σ (homoscedasticity).



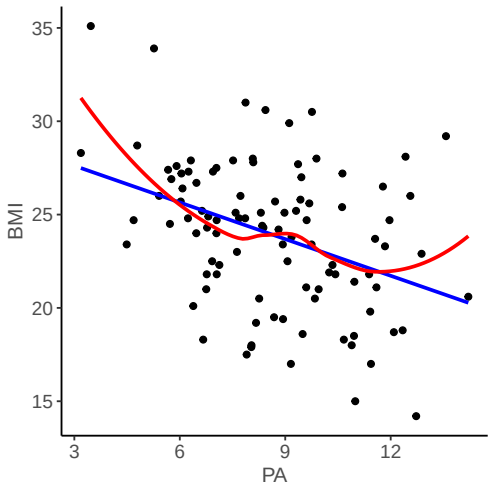
Example 6.10: Assessing BMI model fit

Let's graphically check the model assumptions for the linear model of BMI on physical activity.

```
bmi.model <- lm(BMI~PA, data=bmi)
```

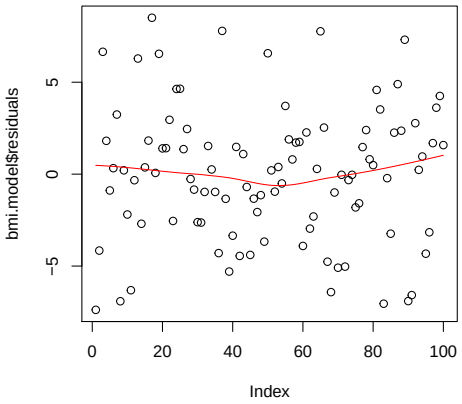
Assumption: The relationship is linear. Fitting a locally weighted scatterplot smoothing **lowess** line on the data will help see the underlying trend of the data.

```
ggplot(bmi, aes(x=PA, y=BMI)) +
  geom_point() +
  geom_smooth(se=FALSE, method="lm",
              col="blue") +
  geom_smooth(se=FALSE, col="red") +
  xlab("PA") + ylab("BMI")
```



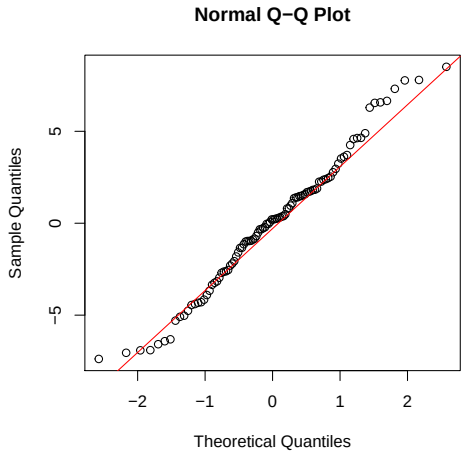
Assumption: The data are independent. Plotting the residuals against row number (order in which the data was collected) can help you see if any discernible pattern arises, such as early or late observations having higher or lower than average responses. Knowledge of the method of data collection here is also paramount!

```
plot(bmi.model$residuals)
lines(lowess(bmi.model$residuals),
      col="red")
```

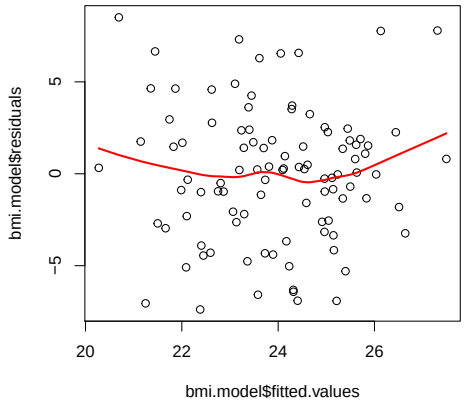


Assumption: The residuals are normal. Points on a normal probability plot should fall close on the red reference line. We tend to worry more about residuals that are far above, or below the line.

```
qqnorm(bmi.model$residuals)
qqline(bmi.model$residuals, col="red")
```



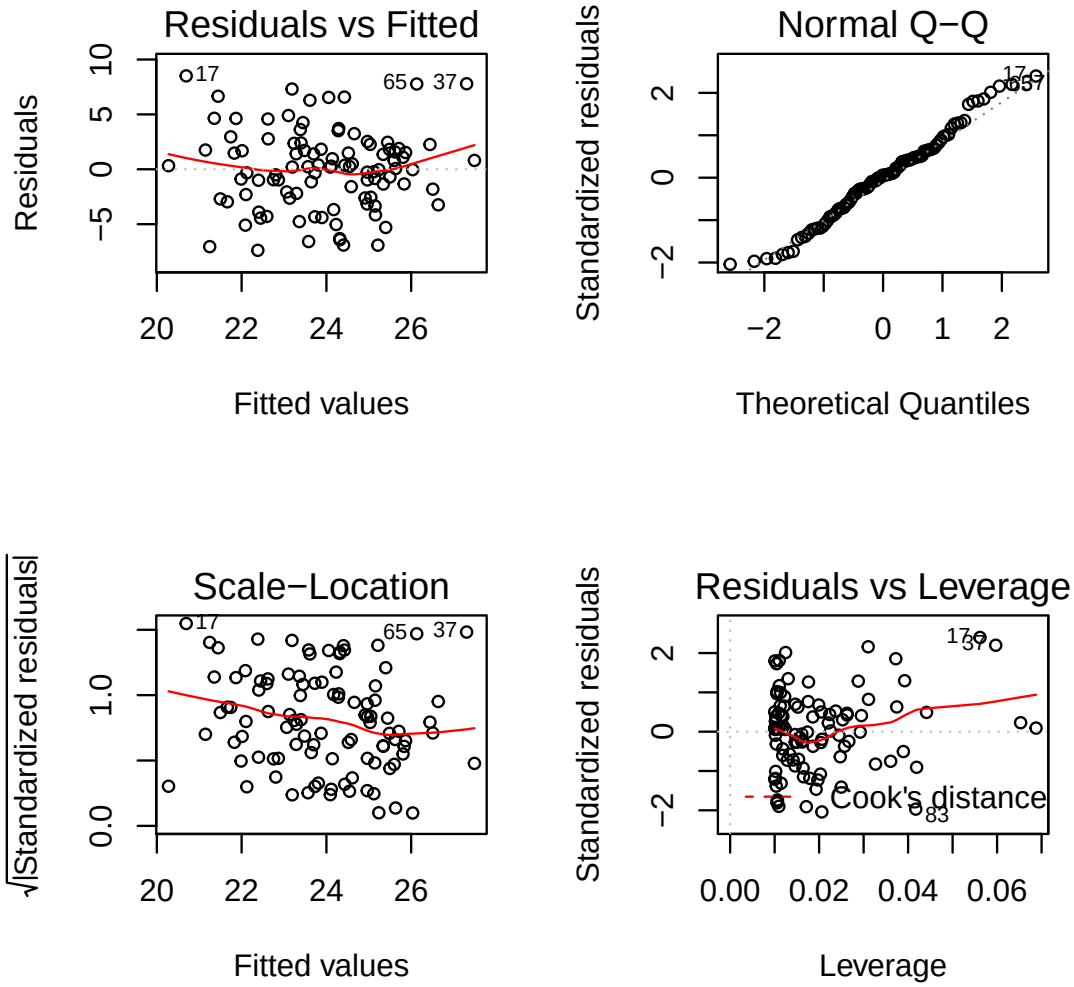
Assumption: Constant variance. Plotting the residuals against the fitted values ($\hat{y}$) allows us to determine if the variance remains relatively constant for all values of x. We see that the residuals increase slightly both at the low and high values. This could mean that a curved relationship between BMI and PA would better fit the data, or it could be chance variation.




```
plot(bmi.model$residuals~bmi.model$fitted.values)
lines(lowess(bmi.model$residuals~bmi.model$fitted.values),
      col="red", lwd=2)
```

Plotting the model fit in R R has a similar set of model diagnostics that are available by simply plotting the model object. The advantage of doing this is that R will identify the row number of observations that it *thinks* are potential outliers that should be investigated and considered for removal.

```
par(mfrow=c(2,2), # Create a 2 x 2 grid of plots
    oma=c(0,0,0,0)) # Remove the outer margins
plot(bmi.model, cex=.8) # cex changes the point sizee
```



6.5.3 Model Predictions

We may want to use the regression line to make predictions about the average (or expected) value of the response ($\hat{\mu}_y$) when x is at a specific value x^*, and a confidence interval for that prediction.

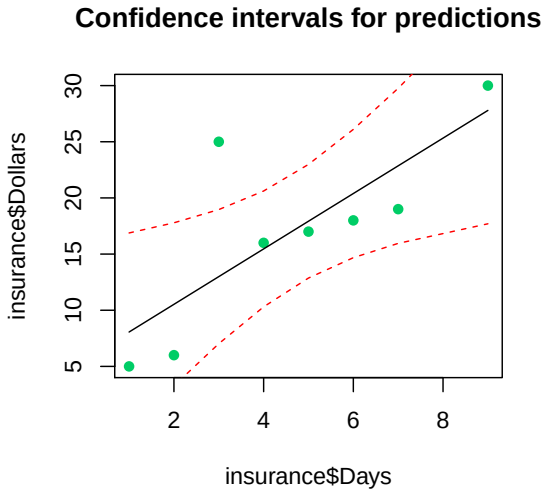
Prediction of the mean response $\hat{\mu}_y$ when x takes on the value x^*

$$\hat{\mu}_y = b_0 + b_1 x^* \tag{6.18}$$

A 95% Confidence Interval (CI) for the prediction of a mean response is

$$\hat{\mu}_y \pm t_{.025,n-2}(SE_{\hat{\mu}}) \qquad \text{where} \qquad SE_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \tag{6.19}$$

Notice how the width of the intervals increases the further away from the bulk of the data you are. The less information you have, the more uncertain you are about where a new data point will lie.



Let's have R calculate the predicted amount of covered hospital expenses for a 3 and 8 day hospital stay, with corresponding confidence intervals for the mean response.

```
new.data <- data.frame(Days = c(3,8)) # Put new data into a data frame
predict(lm(Dollars~Days, data=insurance), new.data, interval="confidence")

##        fit       lwr      upr
## 1 12.99248  7.023261 18.96170
## 2 25.32331 16.911333 33.73528
```

Q: Interpret the CI for 3 days.

Caution on making out-of-sample predictions. Predictions made using a regression model are only valid when predicting values of x that are within the range of data used to estimate the regression model parameters. This means predicting the amount of covered hospital expenses for a 10 day stay would not be a valid prediction using the linear model have been discussing. This is because the linear model describes the data provided, so there is no guarantee that the relationship continues to have the same constant slope and linear relationship outside the window of the data used in the model.

6.5.4 Inferences on Regression Parameters

READING ASSIGNMENT

OpenIntro Section 7.4

Inferences on the parameter estimates are based on the t-ratios,

$$t^* = \frac{(b_0 - \beta_0)}{SE(b_0)} \qquad \text{and} \qquad t^* = \frac{(b_1 - \beta_1)}{SE(b_1)}$$

where t^* has a $\mathcal{T}$ distribution with $n - 2$ degrees of freedom. We are usually interested in testing whether the slope is equal to zero and also calculating a confidence interval for the estimates.

$$H_0 : \beta_1 = 0 \qquad \text{versus} \qquad H_A : \beta_1 \neq 0$$

We would reject H_0 in favor of H_A if the data provide strong evidence that the true slope parameter is something other than zero. Calling the `summary()` function on a model object provides this, and other model information.

Example 6.11: Inference on the slope

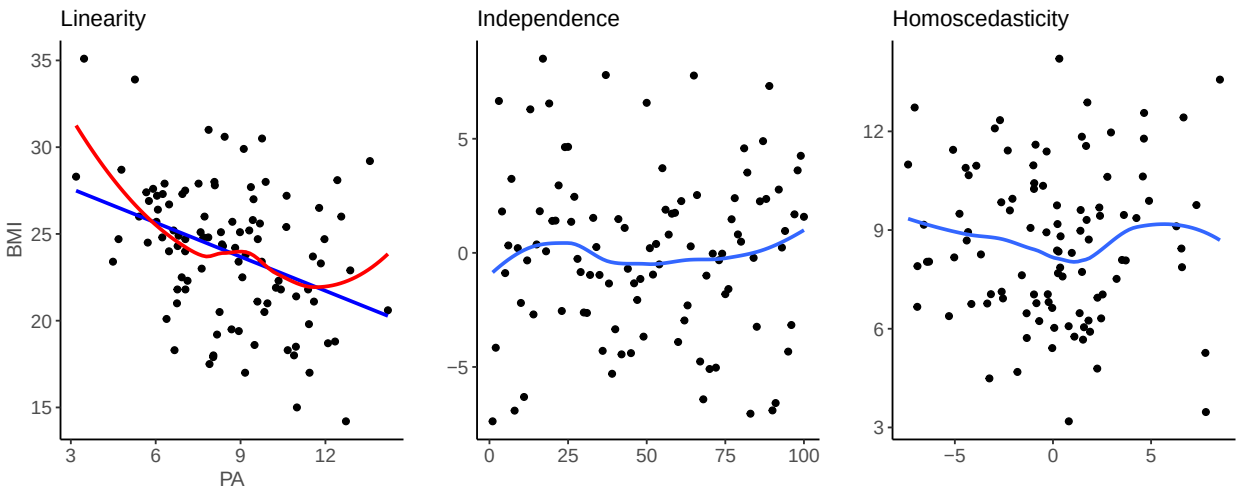
Test if there is a significant relationship between BMI and physical activity on the slope parameter β_1.

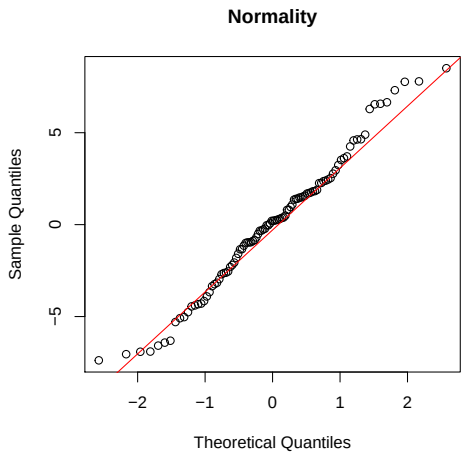
I. Let β_1 be the slope parameter that specifies the direction and strength of relationship between BMI and physical activity.

II. $H_0 : \beta_1 = 0$ There is no relationship between BMI and PA

$H_A : \beta_1 \neq 0$ There is a relationship between BMI and PA.

III. We are interested in the relationship between two continuous variables, so a linear regression model is appropriate as long as the assumptions for linear regression are not violated. You may have noticed when we discussed this earlier that most of the assumptions can only be checked post-modeling. These plots were created earlier so the code will not be shown, just the plots reproduced and discussed below.





While there is some indication of non-constant variance, the rest of the assumptions are upheld. We will proceed with inference based on this linear model, keeping in mind that a linear model may not be the best fit for this data.

IV. Run a linear model in `R` using the `lm` function.

```
summary(lm(BMI~PA, data=bmi))

##
## Call:
## lm(formula = BMI ~ PA, data = bmi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3819 -2.5636  0.2062  1.9820  8.5078
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.5782     1.4120  20.948  < 2e-16 ***
## PA           -0.6547     0.1583  -4.135  7.5e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.655 on 98 degrees of freedom
## Multiple R-squared:  0.1485,Adjusted R-squared:  0.1399
## F-statistic:  17.1 on 1 and 98 DF,  p-value: 7.503e-05
```

The results of the linear model show that the coefficient for the relationship between PA and BMI, β_1 is -0.65, and this is significantly different from zero as shown by the p-value of .000075.

V. This model indicates that physical activity has a significant negative relationship with BMI; as physical activity increases BMI significantly decreases ($p < .001$).

We can also compute the confidence intervals by

```
confint(bmi.model)

##                  2.5 %     97.5 %
## (Intercept) 26.7762222 32.3802721
## PA          -0.9688987 -0.3404729
```

Q: Interpret the 95% confidence interval for β_1.

6.5.5 Example: Fisher's irises III

Example 6.12: Fisher's irises III

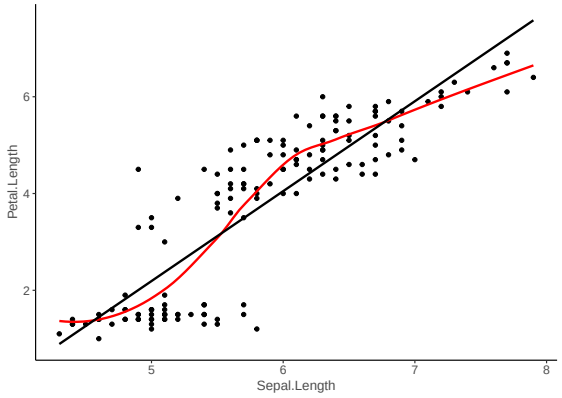
In the last section we found a significant correlation between the length of the sepal and length of petal of an iris flower. Testing that the slope coefficient $\beta \neq 0$ will give us the same result, but linar regression provides us with a measure of how much the quantitative response variable changes as the explanatory variable changes.

1. Identify response and explanatory variables.

- The quantitative explanatory variable is the sepal length (variable `Sepal.Length`)
- The quantitative response variable is the petal length (variable `Petal.Length`)

2. Visualize and summarise bivariate relationship.

```
ggplot(iris, aes(x=Sepal.Length,
                 y=Petal.Length)) +
  geom_point() +
  geom_smooth(se=FALSE, col="red") +
  geom_smooth(method="lm", col="black",
              se=FALSE)
```




```
cor(iris$Sepal.Length, iris$Petal.Length)

## [1] 0.8717538
```

There is a strong, positive, relationship between the sepal length of the flower and the petal length ($r = 0.87$). However the points look a bit clustered, but still mostly linear.

3. Write the relationship you want to examine in the form of a research question.

Does the length of the flower's sepal linearly correlate with the length of the flower's petal?

- Null Hypothesis: There is no linear relationship between length of sepal and petal.
- Alternate Hypothesis: Sepal and petal lengths are linearly related.

4. Perform an appropriate statistical analysis.

I. Let β_1 be the true measure of linear association between sepal and petal length.

II. $H_0 : \beta_1 = 0$

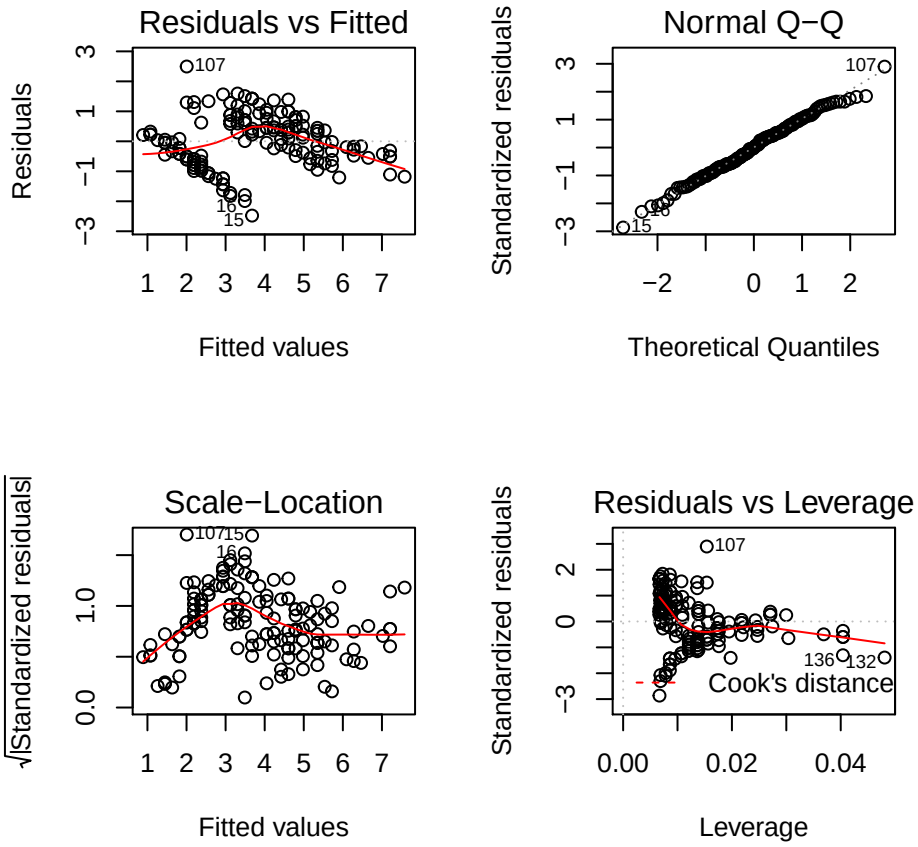
$H_A : \beta_1 \neq 0$

III. Both variables are quantitative; a linear regression analysis will be conducted. The first assumption that the relationship is linear is verified using the scatterplot in part 2. Further assumptions are that the residuals are normally distributed, centered around zero and have constant variance. All assumptions are checked after the model has been fit.

```
iris.linear.model <- lm(Petal.Length ~ Sepal.Length, data=iris)
summary(iris.linear.model)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Sepal.Length, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.47747 -0.59072 -0.00668  0.60484  2.49512
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.10144    0.50666  -14.02   <2e-16 ***
## Sepal.Length  1.85843    0.08586   21.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8678 on 148 degrees of freedom
## Multiple R-squared:   0.76,Adjusted R-squared:  0.7583
## F-statistic: 468.6 on 1 and 148 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(iris.linear.model)
```



Reject the null hypothesis; the p-value for β_1 is $< .0001$. The model assumptions appear to be upheld. The normal Q-Q plot indicates that the residuals are normally distributed and the residuals v. fitted plot indicates that the residuals are centered around zero. There might be a slight violation of the assumption of homoscedascity (constant variance): the variance of the residuals appears to decrease as the fitted values increase.

```
  confint(iris.linear.model)

##                   2.5 %    97.5 %
## (Intercept)  -8.102670 -6.100217
## Sepal.Length  1.688772  2.028094
```

5. Write a conclusion in context of the problem.

The length of a petal and sepal of an iris flower are linearly correlated. For every 1 cm longer the sepal of the flower is, the petal length is increased by 1.85cm (95% CI $1.69, 2.03, p <$ .0001).

Chapter 7

Moderation

Sometimes a third variable can change the relationship between an explanatory and response variable.

Moderation occurs when the relationship between two variables depends on a third variable.

- The third variable is referred to as the moderating variable or simply the moderator.
- The moderator affects the direction and/or strength of the relation between your explanatory and response variable.
- When testing a potential moderator, we are asking the question whether there is an association between two constructs, **but separately for different subgroups within the sample.**

This is also called a **stratified** model, or a **subgroup analysis**.

How to determine if a variable is a moderator

- What you are looking for first depends on which of the 3 scenarios below describe your original analysis (i.e., your original ANOVA test).
- Whether your third variable is a moderator depends on what you see happening in your moderator analysis (i.e., the second ANOVA test split by your third variable).
- If ANY of the 3 scenarios explained below occur in your analysis then your Third Variable IS a Moderator of the Bivariate Relationship.

Scenario 1 - Significant relationship at bivariate level (i.e., ANOVA, Chi-Square, Correlation) (saying expect the effect to exist in the entire population) then when tested for moderation, the third variable is a moderator if the strength (i.e., p-value is Non-Significant) of the relationship changes. It could just change strength for one level of the third variable, not necessarily all levels of the third variable.

Scenario 2 - Non-significant relationship at bivariate level (i.e., ANOVA, Chi-Square, Correlation) (saying do not expect the effect to exist in the entire population) then when tested for moderation, the third variable is a moderator if the relationship becomes significant (saying expect to see it in at least one of the sub-groups or levels of third variable, but not in entire population because was not significant before tested for moderation). It could just become significant in one level of the third variable, not necessarily all levels of the third variable.

Scenario 3 - Significant relationship at bivariate level (i.e., ANOVA, Chi-Square, Correlation) (saying expect the effect to exist in the entire population) then when tested for moderation the third variable is a moderator if the direction (i.e., means change order/direction) of the relationship changes. It could just change direction for one level of third variable, not necessarily all levels of the third variable.

What to look for in each type of analysis. **ANOVA** - look at the p-value, r-squared, means, and the graph of the ANOVA and compare to those values in the Moderation (i.e., each level of third variable) output to determine if third variable is a moderator or not.

Chi-Square - look at the p-value, the percents for the columns in the crosstab table, and the graph for the Chi-Square and compare to those values in the Moderation (i.e., each level of third variable) output to determine if third variable is a moderator or not.

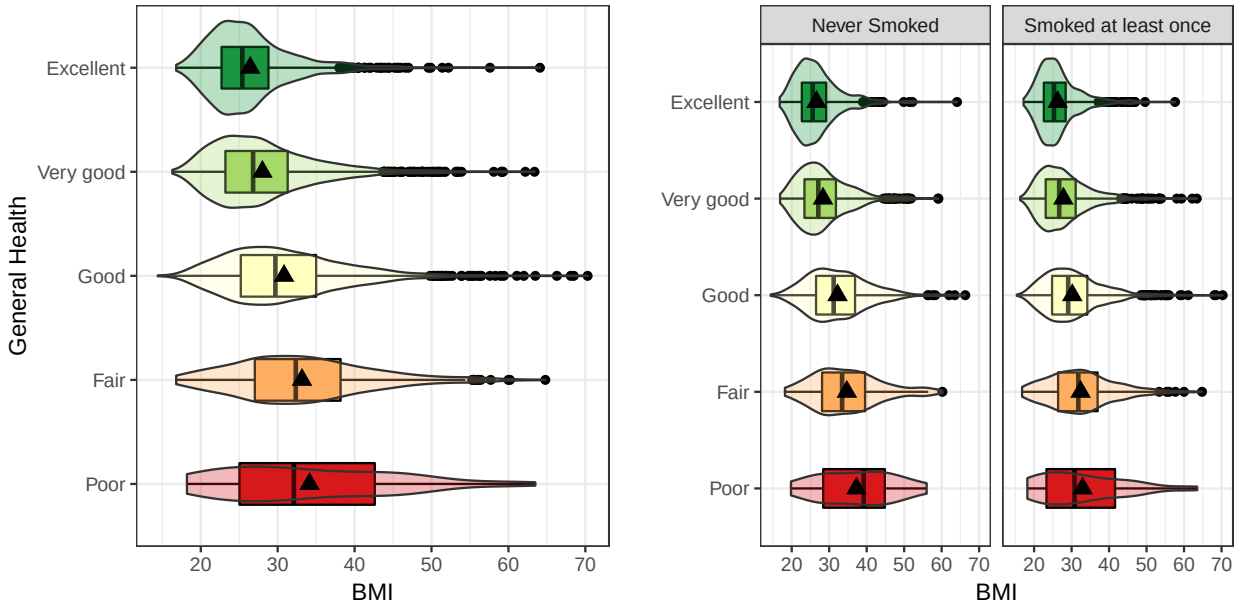
Correlation and Linear Regression - look at the correlation coefficient (r), p-value, regression coefficients, r-squared, and the scatterplot. Compare to those values in the Moderation (i.e., each level of third variable) output to determine if third variable is a moderator or not.

7.1 ANOVA

1. **Identify response, explanatory, and moderating variables**

   - Categorical explanatory variable = Perceived General Health (variable `genhealth`)
   - Quantitative response variable = Body Mass Index (variable `BMI`)
   - Categorical Potential Moderator = Ever smoked (variable `eversmoke_c`)

2. **Visualize and summarise the potential effect of the moderator**




```
bmi.plot %>% group_by(eversmoke_c, genhealth) %>% summarise(mean=mean(BMI))
```

```
## # A tibble: 10 x 3
## # Groups:   eversmoke_c [?]
##    eversmoke_c          genhealth  mean
##    <fct>                <fct>     <dbl>
##  1 Never Smoked         Excellent  26.6
##  2 Never Smoked         Very good  28.4
##  3 Never Smoked         Good       32.2
##  4 Never Smoked         Fair       34.7
##  5 Never Smoked         Poor       37.3
##  6 Smoked at least once Excellent  26.3
##  7 Smoked at least once Very good  27.8
##  8 Smoked at least once Good       30.2
##  9 Smoked at least once Fair       32.4
## 10 Smoked at least once Poor       32.9
```

The average BMI for those reporting excellent health is basically the same between smokers and non-smokers. As perceived general health decreases, the average BMI increases, but it seems to do so at a faster rate in non-smokers compare to smokers. There seem to be more high end outlying points that are smokers compared to non-smokers. Otherwise the relationship between BMI and general health appears the same within each smoking group.

3. **Write the relationship you want to examine in the form of a research question - including a statement about the modifier**

   Does ever smoking change the relationship between perceived general health and BMI? Is the distribution of BMI the same across perceived general health status, for both smokers and non-smokers?

4. **Fit both the original, and stratified models.**

The `pander()` function prints these as nice tables. This is not required, just recommended.

```
aov(BMI ~ genhealth, data=addhealth) %>%
  summary() %>% pander()
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
genhealth	4	22563.37	5640.84	109.24	0.0000
Residuals	5037	260096.03	51.64		

Table 7.1: Original Model

```
aov(BMI ~ genhealth, data=filter(addhealth, eversmoke_c=="Smoked at least once")) %>%
  summary() %>% pander()
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
genhealth	4	11363.79	2840.95	56.62	0.0000
Residuals	3271	164116.79	50.17		

Table 7.2: Model for Smokers

```
aov(BMI ~ genhealth, data=filter(addhealth, eversmoke_c=="Never Smoked")) %>%
  summary() %>% pander()
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
genhealth	4	12564.66	3141.16	59.02	0.0000
Residuals	1745	92872.75	53.22		

Table 7.3: Model for Non-Smokers

5. **Determine if the Third Variable is a moderator or not.**

Both the original ANOVA and the stratified ANOVA models for smokers and non-smokers separately are highly significant. There is not a clear difference in the relationship between BMI and general health status between smokers and non-smokers, so ever being a smoker is not a moderating variable for this relationship.

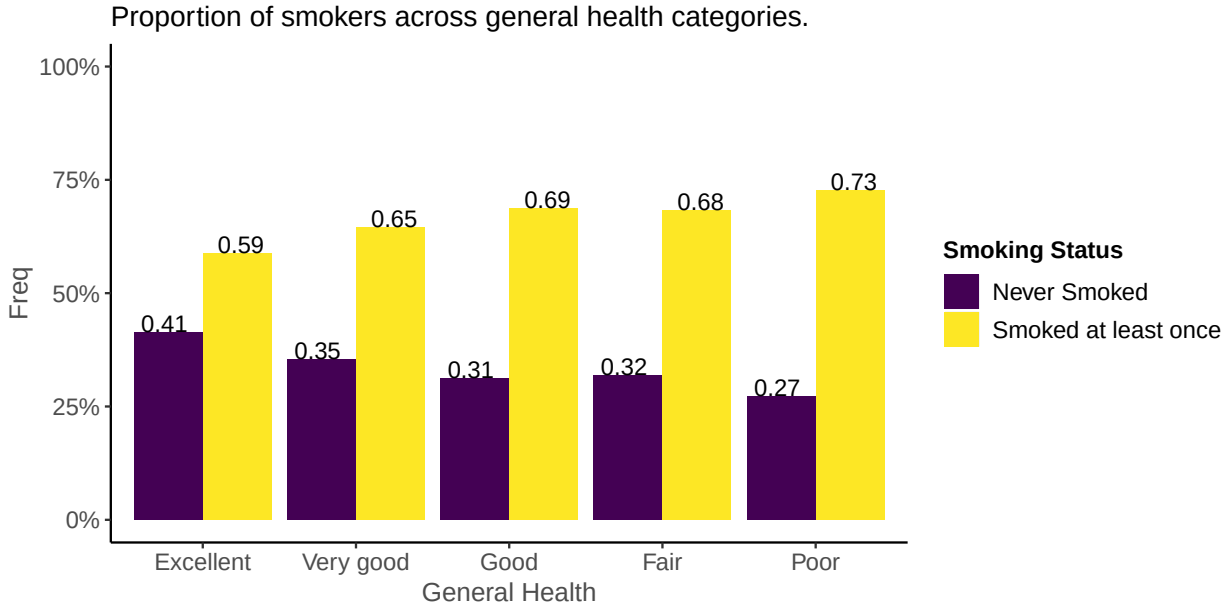
7.2 Chi-Squared

1. **Identify response, explanatory, and moderating variables**

   - Categorical response variable = Ever smoked (variable `eversmoke_c`)
   - Categorical explanatory variable = General Health (variable `genhealth`)
   - Categorical Potential Moderator = Gender (variable `female_c`)

2. **Visualize and summarise the potential effect of the moderator**

Visualize the relationship between smoking and general health across the entire sample.

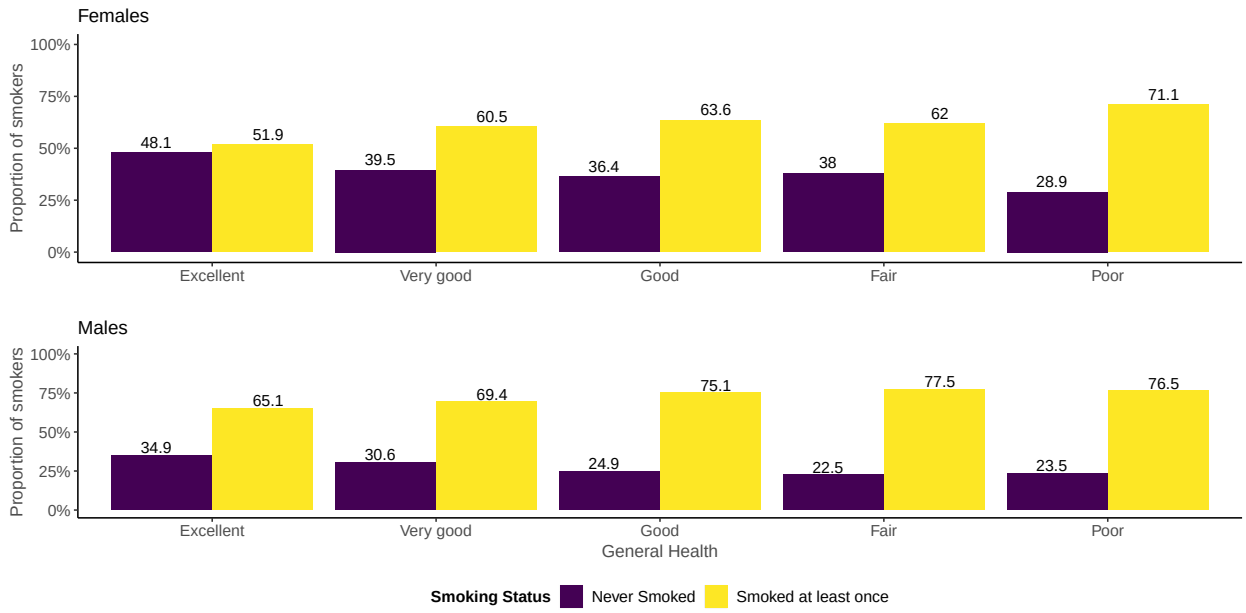


Visualize the relationship between smoking and general health for females and males separately.

```
female <- filter(addhealth, female_c=="Female")

props.f <- table(female$eversmoke_c, female$genhealth) %>%
              prop.table(margin=2)  %>% data.frame()

plot.female.smoke.health <- ggplot(props.f, aes(x=Var2, y=Freq,fill=Var1)) +
  geom_col(position=position_dodge()) +
  geom_text(aes(y=Freq+.05, label=round(Freq*100,1)),
            position = position_dodge(width=1))+
  scale_y_continuous(limits=c(0,1), labels=percent) +
  scale_fill_viridis_d(guide=FALSE) + xlab("")+
  ylab("Proportion of smokers")+ ggtitle("Females")
```

```
male <- filter(addhealth, female_c=="Male")

props.m <- table(male$eversmoke_c, male$genhealth) %>%
              prop.table(margin=2)  %>% data.frame()

plot.male.smoke.health <- ggplot(props.m, aes(x=Var2, y=Freq,fill=Var1)) +
  geom_col(position=position_dodge()) +
  geom_text(aes(y=Freq+.05, label=round(Freq*100,1)),
            position = position_dodge(width=1))+
  scale_y_continuous(limits=c(0,1), labels=percent) +
  scale_fill_viridis_d(name="Smoking Status") + ylab("Proportion of smokers")+
  xlab("General Health") + ggtitle("Males") + theme(legend.position = "bottom")
```

```
grid.arrange(plot.female.smoke.health, plot.male.smoke.health, ncol=1)
```



A general pattern is seen where the proportion of smokers increases as the level of general health decreases. This pattern is similar within males and females.

3. **Write the relationship you want to examine in the form of a research question - including a statement about the modifier**

   Does being female change the relationship between smoking and general health? Is the distribution of smoking status (proportion of those who have ever smoked) equal across all levels of general health, for both males and females?

4. **Fit both the original, and stratified models.**

Original Model

```
chisq.test(addhealth$eversmoke_c, addhealth$genhealth)

##
##  Pearson's Chi-squared test
##
## data:  addhealth$eversmoke_c and addhealth$genhealth
## X-squared = 30.795, df = 4, p-value = 3.371e-06
```

Stratified Models – **Don't change the `function` part here, nor the `x`.** Just change your variable names and data set.

```
by(addhealth, addhealth$female_c, function(x) chisq.test(x$eversmoke_c, x$genhealth))

## addhealth$female_c: Male
##
##  Pearson's Chi-squared test
##
## data:  x$eversmoke_c and x$genhealth
## X-squared = 19.455, df = 4, p-value = 0.0006395
##
## ------------------------------------------------------------
## addhealth$female_c: Female
##
##  Pearson's Chi-squared test
##
## data:  x$eversmoke_c and x$genhealth
## X-squared = 19.998, df = 4, p-value = 0.0004998
```

5. **Determine if the Third Variable is a moderator or not.**

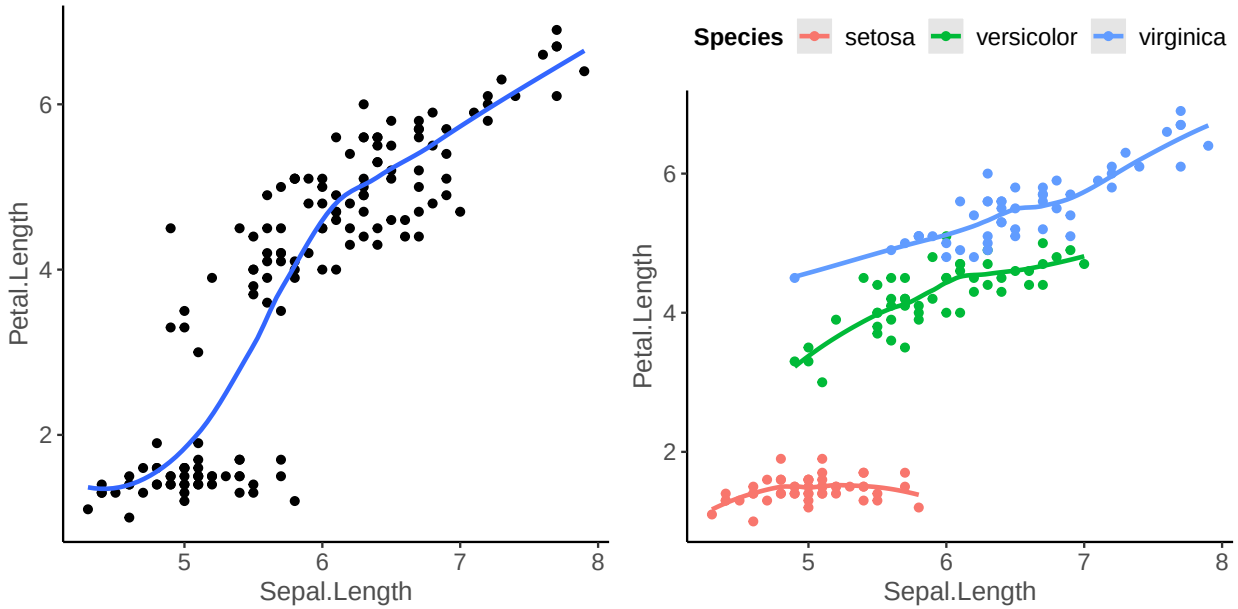
The relationship between gender and general health is significant in both the main effects and the stratified model. The distribution of females across general health categories differs greatly between smokers and non-smokers. This fits into **Scenario 3**, so Smoking is a significant moderator.

7.3 Correlation and Linear Regression

1. **Identify response, explanatory, and moderating variables**

   - Quantitative explanatory variable = Sepal Length (variable `Sepal.Length`)
   - Quantitative response variable = Petal Length (variable `Petal.Length`)
   - Categorical Potential Moderator = Species (variable `Species`)

2. **Visualize and summarise the potential effect of the moderator**



Calculate the sample correlations overall, and for each group.

```
cor(iris$Sepal.Length, iris$Petal.Length)

## [1] 0.8717538

by(iris, iris$Species, function(x) cor(x$Sepal.Length, x$Petal.Length))

## iris$Species: setosa
## [1] 0.2671758
## ------------------------------------------------------------
## iris$Species: versicolor
## [1] 0.754049
## ------------------------------------------------------------
## iris$Species: virginica
## [1] 0.8642247
```

There is a strong, positive, linear relationship between the sepal length of the flower and the petal length when ignoring the species. The correlation coefficient r for *I. virginica* and *I. versicolor* are similar to the overall r value, 0.86 and 0.75 respectively compared to 0.87. However the correlation between sepal and petal length for species *I. setosa* is only 0.26. The points are clearly clustered by species, the slope of the lowess line between *I. virginica* and *I. versicolor* appear similar in strength, whereas the slope of the line for *I. setosa* is closer to zero. This would imply that petal length for *I. setosa* may not be affected by the length of the sepal.

3. **Write the relationship you want to examine in the form of a research question - including a statement about the modifier**

   Does the species of the flower change or modify the linear relationship between the length of the flower's sepal and its petal?

4. **Fit both the original, and stratified models.**

The `pander()` function prints these as nice tables. This is not required, just recommended.

Original Model

```
lm(Petal.Length ~ Sepal.Length, data=iris) %>% pander()
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-7.1014	0.5067	-14.02	0.0000
Sepal.Length	1.8584	0.0859	21.65	0.0000

Stratified Models

```
setosa.model     <- lm(Petal.Length ~ Sepal.Length,
                       data=filter(iris, Species=="setosa"))
veriscolor.model <- lm(Petal.Length ~ Sepal.Length,
                       data=filter(iris, Species=="versicolor"))
virginica.model  <- lm(Petal.Length ~ Sepal.Length,
                       data=filter(iris, Species=="virginica"))
```

```
pander(setosa.model)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.8031	0.3439	2.34	0.0238
Sepal.Length	0.1316	0.0685	1.92	0.0607

```
pander(veriscolor.model)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1851	0.5142	0.36	0.7204
Sepal.Length	0.6865	0.0863	7.95	0.0000

```
pander(virginica.model)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6105	0.4171	1.46	0.1498
Sepal.Length	0.7501	0.0630	11.90	0.0000

5. **Determine if the Third Variable is a moderator or not.**

The estimate for sepal length in the original model is 1.85, p-value <.0001. For *I. setosa* the estimate is 0.13, p-value of 0.06. For *I. versicolor* the estimate is 0.69, p-value <.0001. For *I. virginica* the estimate is 0.75, p-value <.0001.

Within the *I. setosa* species, there is little to no relationship between sepal and petal length. For *I. versicolor* and *I. virginica* the relationship is still significantly positive. This is **Scenario 1**, so Species moderates the effect of sepal length on petal length.

Chapter 8

Multivariable Regression Modeling

So far we have examined the relationship between two variables only. Life is never as simple as that. We know that the number of steps someone takes per day is not the only thing that is related to someone's BMI. What they eat, their age, gender, climate they live in, etc. So how can we understand whether or not physical activity is associated with BMI **after controlling for** these other measures?

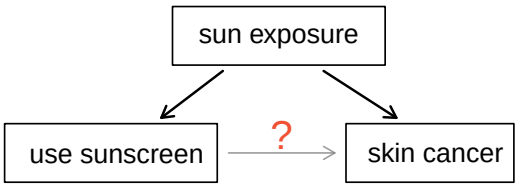
That is, for two people of the same age, same gender, living in the same climate, with the same diet, but their level of physical activity is *different*. Then we can ask how much physical activity affects someone's BMI. But can we really control all those variables?

First we'll discuss some study design concepts such as types of studies and lurking variables. Then we'll see how **multiple regression** can help us tease out how much variability in our response variable is due to an individual explanatory variable.

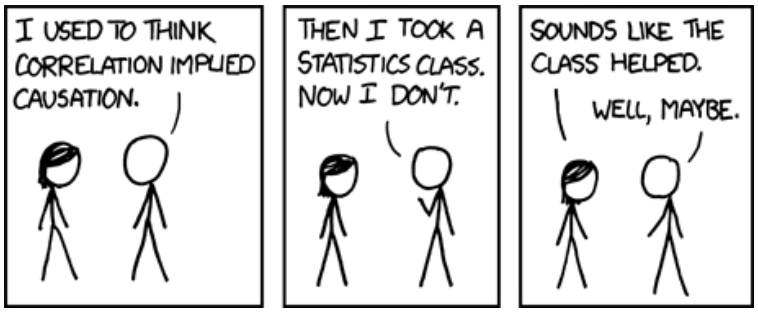
Example 8.1: Skin cancer and sunscreen

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?

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, absent piece of information is sun exposure. If someone is out in the sun all day, s/he 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** (a.k.a. **lurking variable**, **confounder** or **confounding factor**), 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.



8.1 Study Design

8.1.1 Observational studies and Experiments

Reading Assignment

OpenIntro Sections 1.4, 1.5

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

- Observational Study: The researcher simply monitors and collects data on things as they are. There is no manipulation of the study by the researcher. In general, observational studies can provide evidence of a naturally occurring association between variables, but by themselves they cannot show a causal connection.
  - For example, the Youth Risk Behavior Surveillance System (YRBSS) monitors six types of health-risk behaviors that contribute to the leading causes of death and disability among youth and adults.
  - Taking measurements on post-spawn carcasses of Chinook salmon in Butte Creek to assess the annual population health.
- Experiment: In a controlled experiment, the researcher controls the value of the explanatory variable for each unit. In other words, the researcher controls which subjects go into which treatment groups. The value of such control is that cause-and-effect relationships can be established between the response and explanatory variable.
  - For instance, we may suspect that diet and exercise reduce mortality in heart attack patients over the following year. To check if there really is a causal connection between the explanatory variable (diet and exercise) and the response (mortality), researchers will collect a sample of individuals and split them into groups. The individuals in each group are *assigned* a treatment, one group per level of the explanatory variable.
  - To study the effect of tar contained in cigarettes, researchers (Wynder 1953) painted tobacco tar on the back of some mice but not others, and observed if the painted mice had cancer at a higher rate than those not exposed to the tar.

8.1.2 Types of studies

Observational studies come in two forms: prospective and retrospective studies.

A **prospective study** identifies individuals and collects information as events unfold. For instance, medical researchers may identify and follow a group of similar individuals (i.e., a *cohort*) over many years to assess the possible influences of behavior on cancer risk. One

example of such a study is the Nurses Health Study, started in 1976 and expanded in 1989.[1] This prospective study recruits registered nurses and then collects data from them using questionnaires.

Retrospective studies collect data after events have taken place, e.g. researchers may review past events in medical records.

Some data sets, such as the county data examined earlier, may contain both prospectively- and retrospectively-collected variables. Local governments prospectively collected some variables as events unfolded (e.g. retail sales) while the federal government retrospectively collected others during the 2010 census (e.g. county population counts).

Design of Experiments

- An experiment is any study where some treatment is imposed on the experimental units (subjects/participants) in order to observe a response.
- A treatment is a specific experimental condition applied to units in the experiment.
- In an experiment, the explanatory variable is often referred to as a factor. If this variable is categorical, there will be several levels for each factor. In the smoking/lung cancer example, the factor was whether or not a person smokes, occurring at two levels (yes/no).

Essential elements for a good controlled experiment:

- Comparison: There should be at least two groups. Often, there are just two groups: the control group who receives no treatment or the standard treatment, and the experimental group who receives the new treatment. One goal is to compare the responses between groups.
- Randomization: Subjects are assigned randomly (flip of a coin) to groups. This ensures that possible confounding variables are "balanced" between treatment groups.
- Replication: We want to repeat the experiment on a large enough number of experimental units to see any treatment effects.

[1] `http://www.channing.harvard.edu/nhs/`

- Placebo: If possible, the control group receives a placebo to eliminate the possible confounding caused by one group receiving something (a pill, for example) while the other doesn't.

  In medical studies, in particular, people will claim improvement from a drug whether it really helped them or not. This is known as the placebo effect. The introduction of a placebo (or control) into the experiment allows us to measure the effects of such responses. Why? It isn't always possible to give a placebo. For example, if one is interested in comparing two medical treatments, one of which is surgery and the other is doing nothing, then it would be difficult to give the control group a "placebo" surgery.

- Blindness: The subjects in an experiment are "blind" if they do not know which treatment group they are in. The researchers are "blind" if, when measuring the response variable for a subject, they don't know which treatment group the subject is in. An experiment is **double-blind** if both the subjects and the researchers are blind. Blindness for the subjects is important because knowing which group they are in may influence their attitude or behavior. Blindness for the researchers is important when they are measuring the response, because knowing which group the subject is in may (unintentionally) bias their evaluation. It is not always possible to make an experiment double-blind.

8.2 Multiple Linear Regression Analysis

Reading Assignment

OpenIntro Chapter 8.1

The goals of Multiple Linear Regression are to extend simple linear regression to multiple predictors. This means we are trying to describe a linear relationship between a single continuous Y variable and several X variables, and draw inferences regarding these relationships.

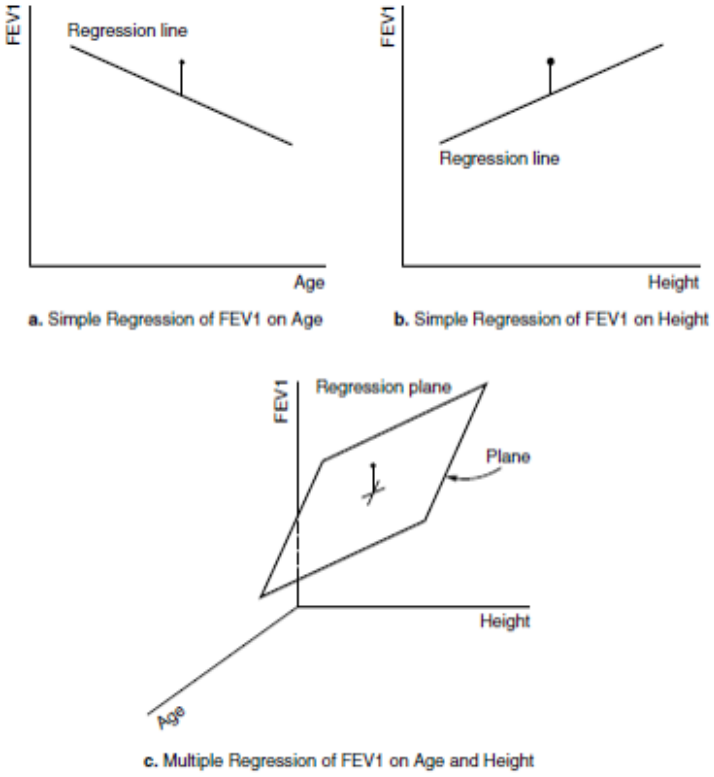


Figure 7.1: *Hypothetical Representation of Simple and Multiple Regression Equations of FEV1 on Age and Height*

Now it's no longer a 2D regression line, but a p dimensional regression plane.

8.2.1 Mathematical Model and Assumptions

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \epsilon_i \tag{8.1}$$

The same assumptions still hold here. Recall that these concern the error terms, or residuals, ϵ_i:

- They have mean zero
- They are homoscedastic, that is all have the same finite variance: $Var(\epsilon_i) = \sigma^2 < \infty$
- Distinct error terms are uncorrelated: (Independent) $\text{Cov}(\epsilon_i, \epsilon_j) = 0, \forall i \neq j$.

We have the same goal to come up with parameter estimates $b_1, \ldots, b_p$ that minimize the residual error. That is, minimize the difference between the value of the dependent variable predicted by the model and the true value of the dependent variable. The details of methods to solve these minimization functions are left to a course in mathematical statistics, however the concept is important.

Example 8.2: Lung Function in adults

Consider the lung function data from the CORD study from Southern California during 1978-1981. In this example we consider the relationship between FEV1 (forced expiratory volume in 1 second) and height on fathers (`FHEIGHT`).

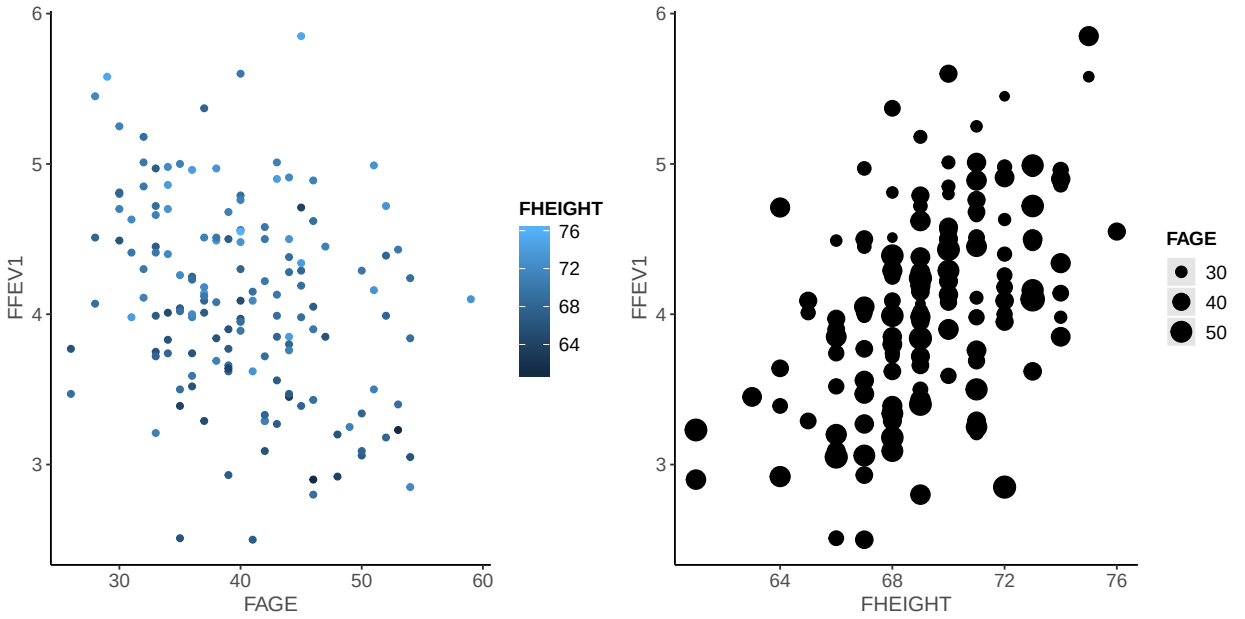
```
summary(lm(FFEV1 ~ FHEIGHT , data=fev))

##
## Call:
## lm(formula = FFEV1 ~ FHEIGHT, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56688 -0.35290  0.04365  0.34149  1.42555
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.08670    1.15198  -3.548 0.000521 ***
## FHEIGHT      0.11811    0.01662   7.106 4.68e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5638 on 148 degrees of freedom
## Multiple R-squared:  0.2544,Adjusted R-squared:  0.2494
## F-statistic:  50.5 on 1 and 148 DF,  p-value: 4.677e-11

lm(FFEV1 ~ FHEIGHT , data=fev) %>% confint() %>% round(2)

##             2.5 % 97.5 %
## (Intercept) -6.36  -1.81
## FHEIGHT      0.09   0.15
```

This model concludes that FEV1 in fathers significantly increases by 0.12 (95% CI: 0.09, 0.15) liters per additional inch in height ($p < .0001$). Looking at the multiple R^2 (correlation of determination), this simple model explains 25% of the variance seen in the outcome y.

However, FEV tends to decrease with age for adults, so we should be able to predict it better if we use both height and age (`FAGE`) as independent variables in a multiple regression equation.

First let's see different ways to graphically explore the relationship between three characteristics simultaneously. Here we plot the relationship between two variables while Controlling the color, or size of points using the third characteristic.

```
a <- ggplot(fev, aes(y=FFEV1, x=FAGE, color=FHEIGHT)) + geom_point()
b <- ggplot(fev, aes(y=FFEV1, x=FHEIGHT, size=FAGE))+ geom_point()
grid.arrange(a, b, ncol=2)
```



The scatterplot of FEV against age demonstrates the decreasing trend of FEV as age increases, and the increasing trend of FEV as height increases. The third color however is pretty scattered across the plot. There is no obvious trend observed.

Q: What direction do you expect the slope coefficient for age to be? For height?

Fitting a regression model in R with more than 1 predictor is accomplished by connecting each variable to the right hand side of the model notation with a +.

```
mv_model <- lm(FFEV1 ~ FAGE + FHEIGHT, data=fev)
summary(mv_model)

##
## Call:
## lm(formula = FFEV1 ~ FAGE + FHEIGHT, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.34708 -0.34142  0.00917  0.37174  1.41853
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.760747   1.137746  -2.427   0.0165 *
## FAGE        -0.026639   0.006369  -4.183 4.93e-05 ***
## FHEIGHT      0.114397   0.015789   7.245 2.25e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5348 on 147 degrees of freedom
## Multiple R-squared:  0.3337,Adjusted R-squared:  0.3247
## F-statistic: 36.81 on 2 and 147 DF,  p-value: 1.094e-13
```

- A father who is one year older is expected to have a FEV value 0.03 liters less than another father of the same height ($p < .0001$).
- A father who is the same age as another father is expected to have a FEV value of 0.11 liter greater than another father of the same age who is one inch shorter ($p < .0001$).

For the model that includes age, the coefficient for height is now 0.11, which is interpreted as the rate of change of FEV1 as a function of height **after adjusting for age**. This is also called the **partial regression coefficient** of FEV1 on height after adjusting for age.

Both height and age are significantly associated with FEV in fathers ($p < .0001$ each).

Example 8.3: Relationship between the time you wake up in the morning and your income, and gender

Does the early bird really get the worm? Or after accounting for gender is there still a relationship between the time someone wakes up and their income? We will use the AddHealth data to explore this question.

1. **Identify response and explanatory variables**

   - Quantitative outcome (y): Income (variable `income`).
   - Quantitative predictor (x_1): Time you wake up in the morning (variable `wakeup`)
   - Binary predictor (x_2): Gender of individual as an indicator of being female (variable `gender`, 0=male, 1=female)

2. **Write the mathematical model**

$$y_i \sim \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

3. **Fit the multivariable model**

```
lm.mod <- lm(income ~ wakeup + female_c, data=addhealth)
summary(lm.mod)

##
## Call:
## lm(formula = income ~ wakeup + female_c, data = addhealth)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -36047 -15141  -5252   8678 205610
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     48669.4     1206.9  40.325  < 2e-16 ***
## wakeup           -611.3      149.4  -4.092 4.37e-05 ***
## female_cFemale  -8527.1      789.3 -10.803  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24300 on 3810 degrees of freedom
```

```
##   (2691 observations deleted due to missingness)
## Multiple R-squared:  0.03236,Adjusted R-squared:  0.03185
## F-statistic:  63.7 on 2 and 3810 DF,  p-value: < 2.2e-16
```

```
confint(lm.mod)
```

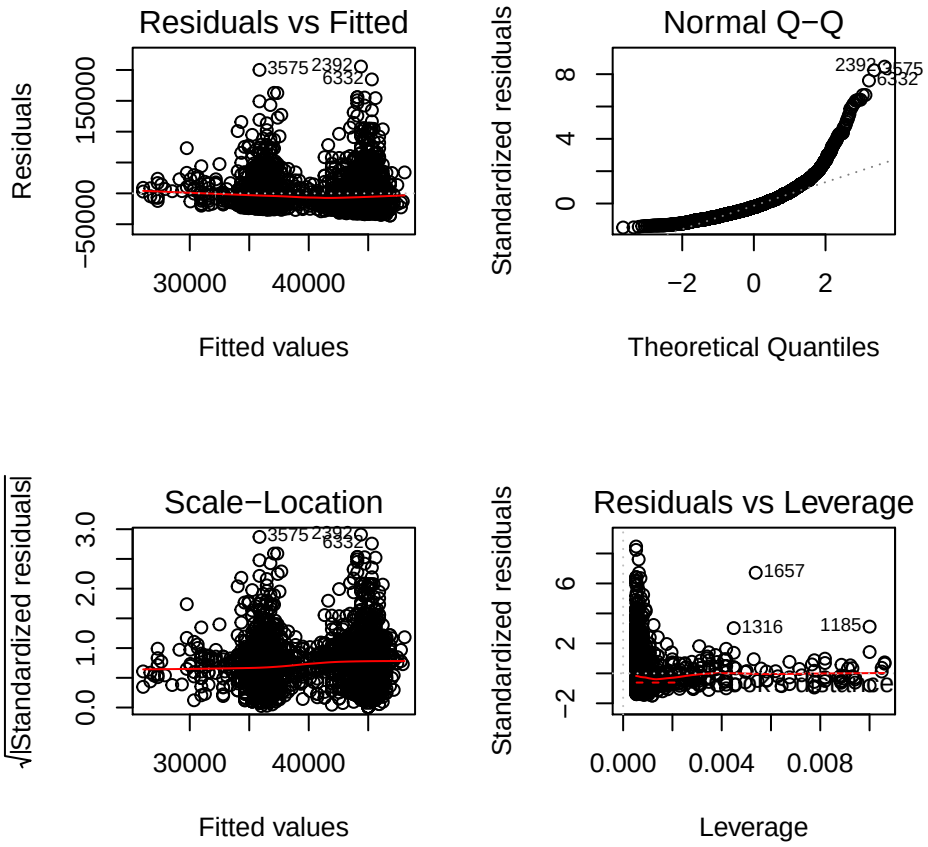
```
##                       2.5 %     97.5 %
## (Intercept)     46303.1191 51035.7225
## wakeup           -904.2448  -318.4088
## female_cFemale -10074.5798 -6979.5823
```

4. **Interpret the regression coefficients.**

- b_1: Holding gender constant, for every hour later a person wakes up, their predicted average income drops by 611 (318, 904) dollars. This is a significant association ($p <$ .01).
- b_2: Controlling for the time someone wakes up in the morning, the predicted average income for females is 8,527 (6980, 10,074) dollars lower than for males. This is a significant association ($p <$ .01).

5. **Check for violations of assumptions**

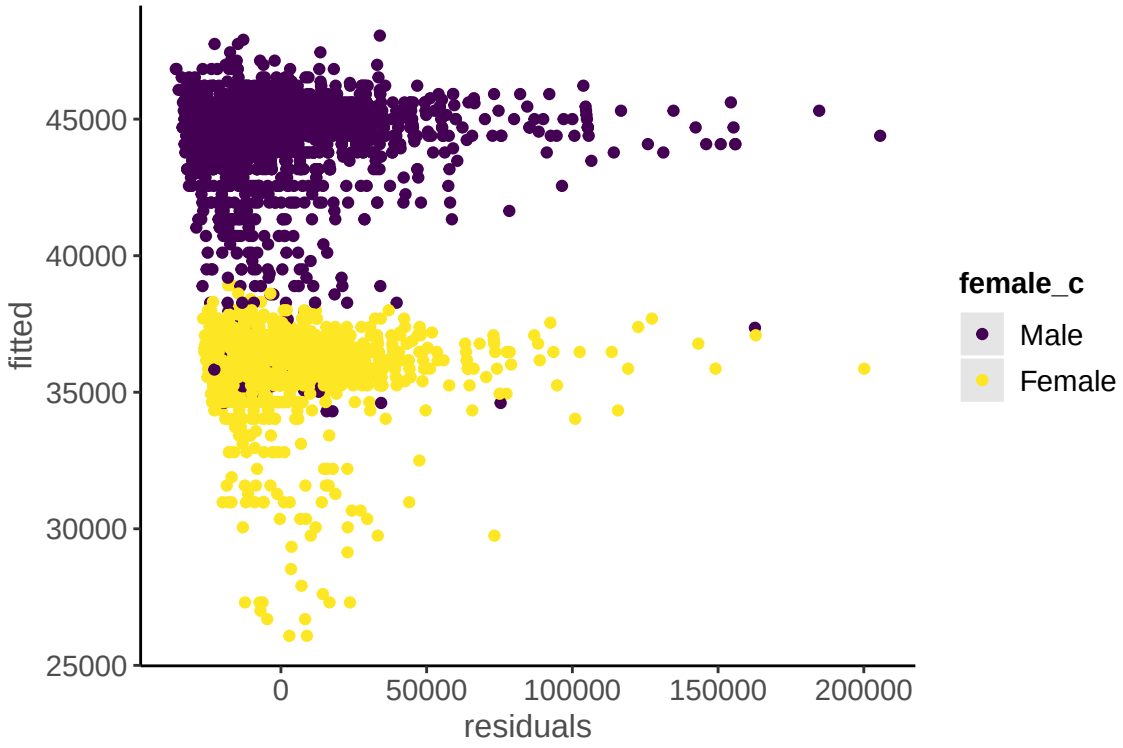
The same set of regression diagnostics can be examined to identify outliers or other problems with the linear model.



The normal probability plot indicates that the residuals are heavily skewed right. Because of the range of the residuals it is difficult to interpret much of the residuals vs fitted plot. The grouping or clustering that is apparent is due to the inclusion of a binary variable (gender) to the model. This is a common feature that may indicate that the model may consistently over (or under) estimate the predicted income for one gender only.

We can take a closer look at this by manually create a plot of the residuals. First we add the residuals `lm.mod$residuals` and fitted values (`lm.mod$fitted.values`) onto the data set (`lm.mod$model`) that was used to fit the model using the function `cbind`.

```
model.results <- cbind(lm.mod$model,
                       residuals = lm.mod$residuals,
                       fitted    = lm.mod$fitted.values)
ggplot(model.results, aes(x=fitted, y=residuals, col=female_c)) +
  geom_point() + coord_flip() + scale_colour_viridis_d()
```



We see that the residuals are not centered at zero, and that the clustering in the residuals really is due to the difference in predicted values for the separate genders. Thus the clustering portion of the pattern that is seen in these residuals can be ignored.

A skew in the residuals is not uncommon, and at this level wouldn't have that large of an effect on the model results. High residuals are seen for those with high incomes. A positive residual means the observed value is larger than the predicted. In other words, the model **underestimates** individuals with large predicted incomes.

8.3 Logistic Regression

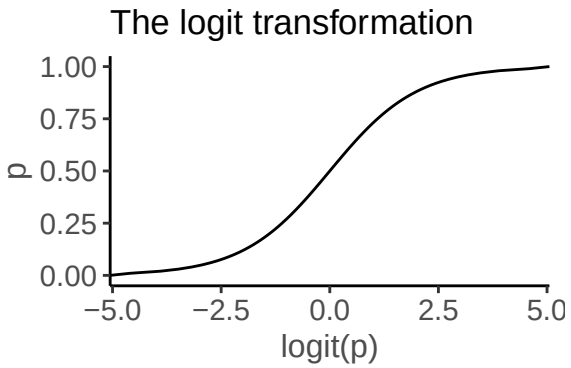
Consider an outcome variable Y with two levels: Y = 1 if event, = 0 if no event. Your outcome variable must be coded as 1 (event) and 0 (non-event). Recoding this way ensures you are predicting the presence of your categorical variable and not the absence of it.

Let $p_i = P(y_i = 1)$.

The logistic model relates the probability of an event based on a linear combination of X's.

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$$

Since the **odds** are defined as the probability an event occurs divided by the probability it does not occur: $(p/(1-p))$, the function $log\left(\frac{p_i}{1-p_i}\right)$ is also known as the *log odds*, or more commonly called the **logit**. This is the link function for the logistic regression model.



This in essence takes a binary outcome 0/1 variable, turns it into a continuous probability (which only has a range from 0 to 1) Then the logit(p) has a continuous distribution ranging from $-\infty$ to ∞, which is the same form as a Multiple Linear Regression (continuous outcome modeled on a set of covariates)

Back solving the logistic model for $p_i = e^{\beta X}/(1 + e^{\beta X})$ gives us the probability of an event.

$$p_i = \frac{e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}}}$$

8.3.1 Fitting Logistic Models

Example 8.4: The effect of gender on Depression

Consider the depression data set where we have
a binary outcome variable: Symptoms of Depression (`cases`)
a binary predictor variable: Gender (`sex`) as an indicator of being female
continuous predictor variables: age, income.

The general syntax is similar to `lm()`, with the additional required `family=` argument. Just like multiple linear regression, additional predictors are simply included in the model using a `+` symbol.

```
log.model <- glm(cases ~ age + income + sex, data=depress, family="binomial")
summary(log.model)

##
## Call:
## glm(formula = cases ~ age + income + sex, family = "binomial",
##     data = depress)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0249  -0.6524  -0.5050  -0.3179   2.5305
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.67646    0.57881  -1.169  0.24253
## age         -0.02096    0.00904  -2.318  0.02043 *
## income      -0.03656    0.01409  -2.595  0.00946 **
## sex          0.92945    0.38582   2.409  0.01600 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 268.12  on 293  degrees of freedom
## Residual deviance: 247.54  on 290  degrees of freedom
## AIC: 255.54
##
## Number of Fisher Scoring iterations: 5
```

The output looks very similar to what we saw for logistic regression.

8.3.2 Interpreting Odds Ratios

The regression coefficients b_p from a logistic regression are not interpreted directly, since a 1 unit change in x does not correspond to a b_p unit change in y. The coefficients instead must be *exponentiated* to create the **Odds Ratio**. This is done by raising the constant e to the value of the coefficient.

$$OR = e^b$$

The odds ratio is interpreted by comparing it to a value to 1. You will see one of three things:

- $OR = 1$: equal chance of response variable being YES given any explanatory variable value. You are not able to predict participants responses by knowing their explanatory variable value. This would be a non-significant model when looking at the p-value for the explanatory variable in the parameter estimate table.
- $OR > 1$: as the explanatory variable value increases, the presence of a YES response is more likely. We can say that when a participants response to the explanatory variable is YES (1), they are more likely to have a response that is a YES (1).
- $OR < 1$: as the explanatory variable value increases, the presence of a YES response is less likely. We can say that when a participants response to the explanatory variable is YES (1) they are less likely to have a response that is a YES (1).

Where does $OR = e^{\beta}$ come from?

The model is:

$$log(odds) = -0.676 - 0.02096 * age - .03656 * income + 0.92945 * gender$$

We want to calculate the Odds Ratio of depression for women compared to men.

$$OR = \frac{Odds(Y = 1|F)}{Odds(Y = 1|M)}$$

Write out the equations for men and women separately.

$$= \frac{e^{-0.676-0.02096*age-.03656*income+0.92945(1)}}{e^{-0.676-0.02096*age-.03656*income+0.92945(0)}}$$

Applying rules of exponents to simplify.

$$= \frac{e^{-0.676}e^{-0.02096*age}e^{-.03656*income}e^{0.92945(1)}}{e^{-0.676}e^{-0.02096*age}e^{-.03656*income}e^{0.92945(0)}}$$

$$= \frac{e^{0.92945(1)}}{e^{0.92945(0)}}$$

$$= e^{0.92945}$$

```
exp(.92945)

## [1] 2.533116

exp(coef(log.model)[4])

##      sex
## 2.533112
```

Confidence intervals are calculated on the coefficients first, then transformed into Odds Ratios.

```
exp(confint(log.model))

##                 2.5 %    97.5 %
## (Intercept) 0.1585110 1.5491849
## age         0.9615593 0.9964037
## income      0.9357319 0.9891872
## sex         1.2293435 5.6586150
```

The odds of a female being depressed are 2.53 (1.2, 6.7) times greater than the odds for a male after adjusting for the linear effects of age and income ($p = .016$).

Effect of a k unit change

Sometimes a 1 unit change in a continuous variable is not meaningful.

- The Adjusted odds ratio (AOR) for increase of 1 year of age is 0.98 (95% CI .96, 1.0)
- How about a 10 year increase in age? $e^{10*\beta_{age}} = e^{-.21} = .81$

```
exp(10*coef(log.model)) %>% round(3)

## (Intercept)         age      income         sex
##       0.001       0.811       0.694   10877.881

exp(10*confint(log.model)) %>% round(3)

##                2.5 %       97.5 %
## (Intercept) 0.000       79.622
## age         0.676        0.965
## income      0.515        0.897
## sex         7.884 33659019.318
```

Controlling for gender and income, an individual has 0.81 (95% CI 0.68, 0.97) times the odds of being depressed compared to someone who is 10 years younger than them. *Note: a 10-unit increase in* `sex` *is meaningless so ignore that line.*

Example 8.5: The time you wake up vs poverty level

Let's revisit the relationship between income, the time you wake up, and gender. But since the `income` variable was so right-skewed, let's dichotomize it and make a new binary indicator for whether or not the individual is living below the poverty line.

1. Identify variables

- Binary response variable (y): Poverty (variable `poverty`). This is an indicator if reported personal income is below $10,210.
- Quantitative explanatory variable (x_1): Time you wake up in the morning (variable `wakeup`)
- Binary explanatory variable (x_2): Gender (variable `female_c`)

Since the response variable is binary, we must use a logistic regression model.

2. Write the mathematical model

$$logit(y_i) \sim \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

3. Fit the multivariable model

```
log.mod <- glm(poverty~wakeup + female_c, data=addhealth, family='binomial')
summary(log.mod)

##
## Call:
## glm(formula = poverty ~ wakeup + female_c, family = "binomial",
##     data = addhealth)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0597  -0.7703  -0.5423  -0.5141   2.1124
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.18642    0.11857 -18.439  < 2e-16 ***
## wakeup          0.04587    0.01351   3.396 0.000683 ***
## female_cFemale  0.84822    0.07660  11.074  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4909.2  on 4832  degrees of freedom
## Residual deviance: 4772.7  on 4830  degrees of freedom
##   (1671 observations deleted due to missingness)
## AIC: 4778.7
##
## Number of Fisher Scoring iterations: 4
```

4. Interpret the Odds Ratio estimates Below I create a table containing the odds ratio estimates and 95% CI for those estimates using the confounding model.

```
data.frame(
  OR  = exp(coef(log.mod)),
  LCL = exp(confint(log.mod))[,1],
  UCL = exp(confint(log.mod))[,2]
  )
```

```
##                       OR        LCL       UCL
## (Intercept)    0.112318 0.08897539 0.1416664
## wakeup         1.046936 1.01919842 1.0746803
## female_cFemale 2.335481 2.01185025 2.7166401
```

- After controlling for gender, those that wake up one hour later have 1.05 (1.02, 1.07) times the odds of reporting annual earned wages below the federal poverty level compared to someone waking up one hour earlier. This is a significant association ($p < .001$), but the magnitude of the increase is very small.
- After controlling for the time someone wakes up, females have 2.34 (2.01, 2.72) times the odds of reporting annual earned wages below the federal poverty level compared to males. This is a significant association ($p < .001$).

8.4 Model Building

How do you decide between models with different explanatory variables? How do you model a categorical predictor variable? How do you include moderators in the model? What effect can outliers have on a model? We will explore some of these questions in this last chapter.

8.4.1 Variable selection

Model building methods are used mainly in exploratory situations where many independent variables have been measured, but a final model explaining the dependent variable has not been reached.

We want to choose a set of independent variables that both will yield a good fit using as few variables as possible. In many situations where regression is used the investigator may have prior justification for using certain variables (such as prior studies or accepted theory) but may be open to suggestions for the remaining variables.

The set of independent variables can be broken down into logical subsets:

- The usual demographics are entered first (age, gender, ethnicity)

- A set of variables that other studies have shown to affect the dependent variable
- A third set of variables that *could* be associated but the relationship has not yet been examined.

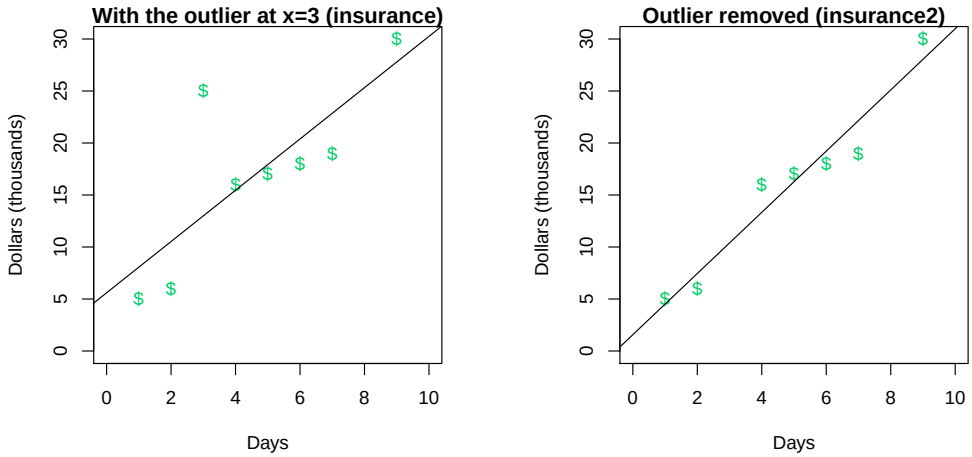
When working with multiple models, how do you choose the optimal model? Two criteria you can use is the RMSE and the Adjusted R^2.

1. RMSE (Root Mean Squared Error): How biased are the results? How "far away" are the estimates $\hat{\theta}$ from the truth θ? $\sqrt{MSE(\hat{\theta})} = \sqrt{E((\hat{\theta} - \theta)^2)}$. We would like to minimize this value.
2. Multiple R^2: If the model explains a large amount of variation in the outcome that's good right? So we could consider using **Adjusted** R^2 as a selection criteria and trying to find the model that maximizes this value. Problem: As the number of predictors added to the model increases, measurements such as the R^2 will also increase.

8.4.2 Outliers

Reconsider the plot of fictitious hospital expenses against length of stay at the hospital (left). The point at $x = 3$ is called an outlier. We can remove this observation by dropping that whole row. Observe what happens to the regression line if it is removed (right).

```
insurance2 <- insurance[-3,]
```



```
lm(Dollars~Days, data=insurance2)

##
## Call:
## lm(formula = Dollars ~ Days, data = insurance2)
##
## Coefficients:
## (Intercept)         Days
##       1.567        2.942
```

The new regression equation is

$$\hat{y} = \_\_\_\_\_ + \_\_\_\_\_ x$$

Q: How did the slope and intercept values change?

Q: What do you think would happen if we removed the observed data point at 9 days?

Q: What if there was an outlier at 8 days and \$5k expenses? What would happen to the slope if that point were removed?

8.4.3 Interpreting Categorical Predictors

Factor variable coding is also commonly known as "dummy coding". A better used term is indicator variable or reference coding.

- For a nominal X with K categories, define K indicator variables.
- Choose a reference (referent) category:
- Leave it out
- Use remaining $K - 1$ in the regression.
- Often, the largest category is chosen as the reference category.

Example 8.6: Southern California Pollution

Consider a linear model to examine the effect of geographic area on fathers FEV1 `FFEV1` while controlling for age (x_1).

```
fev$AREA <- factor(fev$AREA,
                   labels= c("Burbank", "Lancaster", "Long Beach", "Glendora"))
```

	Burbank	Lancaster	Long Beach	Glendora
1	24	49	19	58

Area has 4 levels, so we would need 3 indicator variables. `R` always uses the first level of a factor variable as the reference level.

- Let $x_2 = 1$ when `AREA='Lancaster'`, and 0 otherwise,
- let $x_3 = 1$ when `AREA='Long Beach'`, and 0 otherwise,
- let $x_4 = 1$ when `AREA='Glendora'`, and 0 otherwise.

The mathematical model would look like:

$$Y = \beta_0 + \beta_1 * x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon \tag{8.2}$$

The coefficients for the other levels of the categorical variable are interpreted as the effect of that variable on the outcome in *compared to* the reference level.

```
area.model <- lm(FFEV1 ~ FAGE + AREA, data=fev)
summary(area.model)

##
## Call:
## lm(formula = FFEV1 ~ FAGE + AREA, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.68574 -0.47983 -0.02035  0.47581  1.94084
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.138869   0.321546  15.982  < 2e-16 ***
## FAGE           -0.028658   0.007508  -3.817 0.000199 ***
## AREALancaster   0.059886   0.156290   0.383 0.702154
## AREALong Beach  0.219614   0.191466   1.147 0.253263
## AREAGlendora    0.147790   0.151289   0.977 0.330258
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6233 on 145 degrees of freedom
## Multiple R-squared:  0.1072,Adjusted R-squared:  0.08252
## F-statistic: 4.351 on 4 and 145 DF,  p-value: 0.002367

round(confint(area.model),2)

##                 2.5 % 97.5 %
## (Intercept)      4.50   5.77
## FAGE            -0.04  -0.01
## AREALancaster   -0.25   0.37
## AREALong Beach -0.16   0.60
## AREAGlendora    -0.15   0.45
```

- After controlling for geographic area, for every year older a father in the CORD study is, his lung function significantly decreases by .03 (.01, .04), $p = .0002$.
- Controlling for age, a father who lives in Lancaster has .06 (-.25, .37) higher lung function than a father who lives in Burbank.
- Controlling for age, a father who lives in Long Beach has .21 (-.16, .60) higher lung function than a father who lives in Burbank.
- Controlling for age, a father who lives in Glendora has .15 (-.15, .45) higher lung function than a father who lives in Burbank.

None of these differences are significant, so we can conclude that after controlling for age, area does not have an effect on the fathers' lung function.

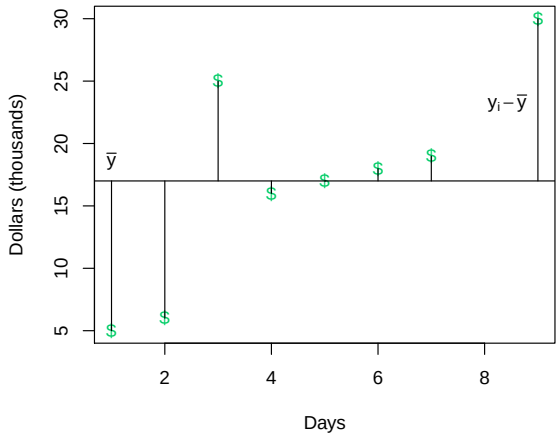
8.4.4 Model testing and fit

Recall that for ANOVAs, the table summarized the comparison of the within group variation to between group variation. The F-statistic was the ratio of the between groups to the within group variation. Recall also that the total sum of squares was partitioned into the sum of squares within groups and the sum of squares between groups. We can do a similar thing with regression to produce an ANOVA table.

The total sum of squares (SST) is calculated by computing the sum of the squared distances from y_i to $\bar{y}$:

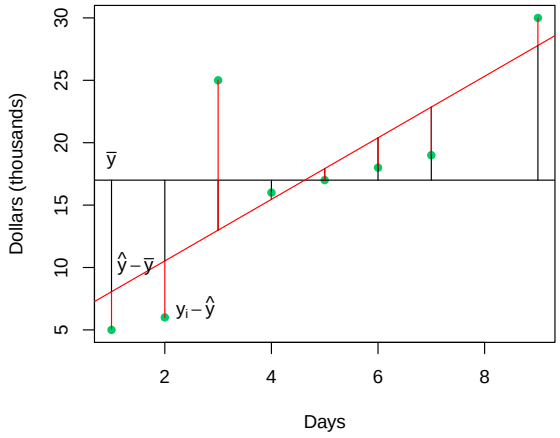
$$\sum_{i=1}^{n}(y_i - \bar{y})^2$$

and is a measure of the variation in the y_i's without any consideration of x.



The vertical distance from each observation to the mean can be partitioned into to parts: the distance from the point to the line and the distance from the line to the mean. We can partition the total variability in the y's into 2 parts:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$



Therefore, the regression line partitions the variability in y into the part that is explained

by the regression model (SSM) and the part that remains unexplained (SSE).

The ANOVA table for simple linear regression is as follows

Source	SS	DF	MS	F-Test
Regression	SSM	1	$SSM/_1$	$MSM/_{MSE}$
Error	SSE	$n-2$	$SSE/_{(n-2)}$	
Total	SST	$n-1$	-	

The regression sum of squares is the amount by which the residual sum of squares decreases when the model for the mean of y has the the added term $\beta_1 X$. Therefore, a large regression sum of squares implies that the regression line is doing an adequate job of modeling the data. If X is a good predictor of Y we would expect the regression sum of squares to be large relative to the residual (error) sum of squares. This would cause the F-statistic to be large, giving us a small p-value. The F-statistic test is a test that compares the regression model to the mean model ($\bar{y}$). If the p-value is small then there is evidence that the regression model is an improvement over the mean model.

- For Simple Linear Regression, testing the hypothesis that $\beta_1 = 0$ is equivalent to the F test MSM/MSE.

Example 8.7: ANOVA table for the analysis of BMI

Let's continue our examination of BMI and physical activity (PA).

The ANOVA table is

```
summary(aov(BMI~PA, data=bmi))
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## PA            1  228.4  228.38    17.1 7.5e-05 ***
## Residuals    98 1309.1   13.36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. Interpret the F test for overall model fit.

2. What does this tell you about the slope parameter β_1?

Coefficient of determination (R^2)

This value has the same interpretation as we saw previously, namely the percentage of the total response variation explained by the explanatory variable. This can be calculated from the ANOVA table as we did earlier in this chapter, but it is also presented as the `Multiple R-Squared` value from the summary of the linear regression object obtained in `R`. We will cover Adjusted R-squared when we discuss multiple linear regression. For the BMI example:

```
summary(lm(BMI~PA, data=bmi))

##
## Call:
## lm(formula = BMI ~ PA, data = bmi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3819 -2.5636  0.2062  1.9820  8.5078
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.5782     1.4120  20.948  < 2e-16 ***
## PA           -0.6547     0.1583  -4.135  7.5e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.655 on 98 degrees of freedom
## Multiple R-squared:  0.1485,Adjusted R-squared:  0.1399
## F-statistic:  17.1 on 1 and 98 DF,  p-value: 7.503e-05
```

We see that 14.8% of the variation in BMI can be explained by the regression equation on among of physical activity.

Recall a high R^2 does not necessarily mean that a linear model is appropriate. You should always look at scatterplots and residual plots of your data before reporting R^2.

For a global test to see whether or not the regression model is helpful in predicting the values

of y, we can use an ANOVA. This is the same as testing that all $\beta_j, j = 1, \ldots, p$ are all equal to 0. Let's look at what we get if we wrap the ANOVA function, `aov()` around the linear model results.

```
summary(aov(mv_model))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## FAGE          1   6.04   6.044   21.13 9.17e-06 ***
## FHEIGHT       1  15.01  15.013   52.49 2.25e-11 ***
## Residuals   147  42.04   0.286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We get the sums of squares (SS) for each predictor individually, not combined into a SS for regression and a SS residuals. So where do we find this global test that this model is better than using no predictors at all? At the very last line in the summary of the linear model results.

```
summary(mv_model)
```

```
##
## Call:
## lm(formula = FFEV1 ~ FAGE + FHEIGHT, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.34708 -0.34142  0.00917  0.37174  1.41853
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.760747   1.137746  -2.427   0.0165 *
## FAGE        -0.026639   0.006369  -4.183 4.93e-05 ***
## FHEIGHT      0.114397   0.015789   7.245 2.25e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5348 on 147 degrees of freedom
## Multiple R-squared:  0.3337,Adjusted R-squared:  0.3247
## F-statistic: 36.81 on 2 and 147 DF,  p-value: 1.094e-13
```