AP Statistics Notes

Mr. Chang edwardchang@berkeley.net

2023-01-25

Contents

Chapter 1

Basic notation

1.1 Lists

Most of the time, you will see x denote a list of values (i.e. a variable in a data table).

For example, if x was the list of numerical values a to g, we can write it as:

$x = [a, b, c, d, e, f, g]$

Then, x_i means the i^{th} value in the list of x, so

$x_1 = a$ and $x_2 = b$ and $x_7 = g$.

1.2 n

In regards to a data table or list of values, n stands for the number of rows or data points that are in the data table or list (we will learn this as the *sample size* later on)

So, for list x, $n = 7$.

Adding on, x_n would mean the last value in the list x (since there are only n values in x)

1.3 Summation (Σ)

You will also see the greek letter Σ in formulas. Usage of this sign means that we are using *summation notation*.

If we want the sum of all numbers from 1 to 7, we would write it as,

$$\sum_{i=1}^{7} i$$

We interpret this as,

1. Start from $i = 1$, evaluate the expression, which is i.
2. Keep our evaluated expression to the side and get ready to add the other values to it, so

$$1 + \cdots$$

3. Now go the next numbers until we get to 7 (the number on the top of the Σ) So moving onto $i = 2$, we end up with

$$1 + 2 + \cdots$$

   And with $i = 3$, we end up with

$$1 + 2 + 3 + \cdots$$

4. When we get to the end of it (when we reach $i = 7$), we have the *expanded form* the summation.

$$1 + 2 + 3 + 4 + 5 + 6 + 7$$

1.4 Other notation

- $\bar{x}$: The "line" on top of x means the *mean of* x. If we had $\bar{a}$, I would be asking for the mean of a.
  - Pronounced "x bar"
- $\hat{p}$: The "hat" on top of p means the *estimate of* p. If we had $\hat{x}$, I would be asking for the estimate of x.
  - Pronounced "p hat"

1.5 Other commonly used symbols

- p: proportion, probability, or p-value
- $\bar{x}$: sample mean
- s_x: sample standard deviation (of x), so s_y is the sample standard deviation of y
- μ: population mean (true mean)
- σ: population standard deviation (true standard deviation)
- N: population size

Chapter 2

Categorical Data Visualizations

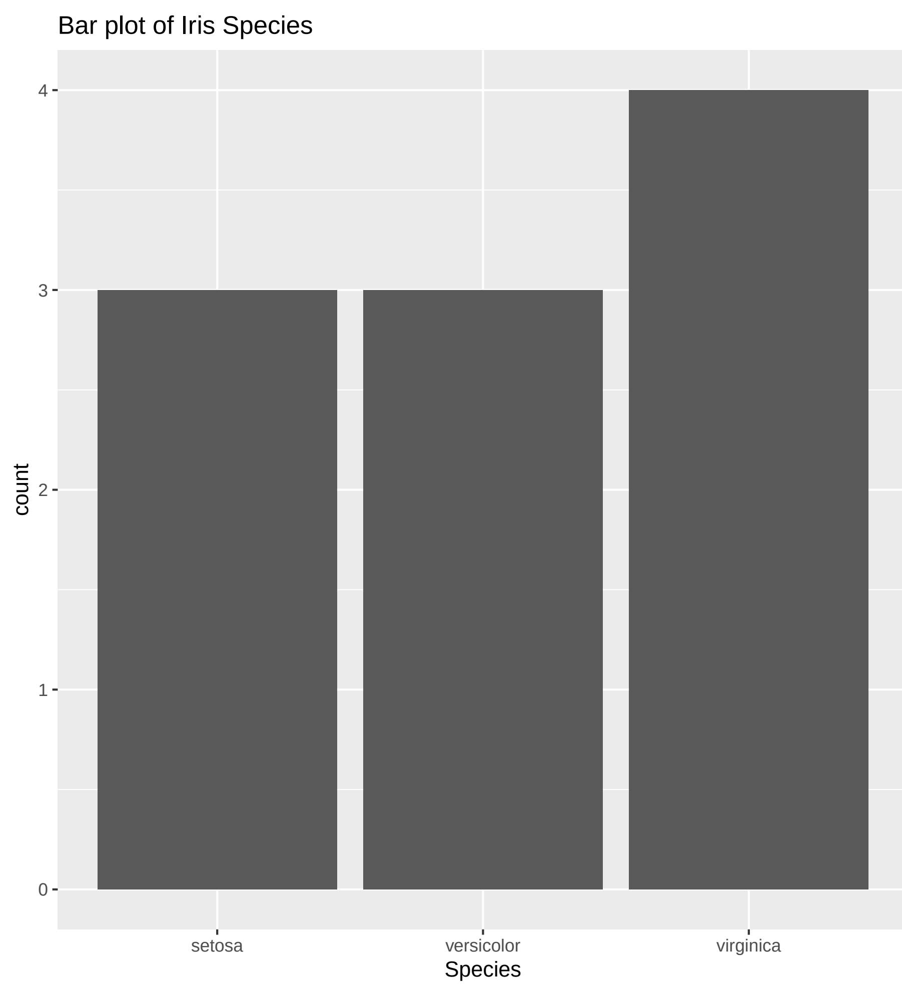
2.1 Bar plots

Represent the number or proportion of each unique value. These numbers or proportions are represented with rectangular bars with heights proportional to the values that they represent. You can plot these vertically or horizontally (i.e. categories on the x-axis or categories on the y-axis)

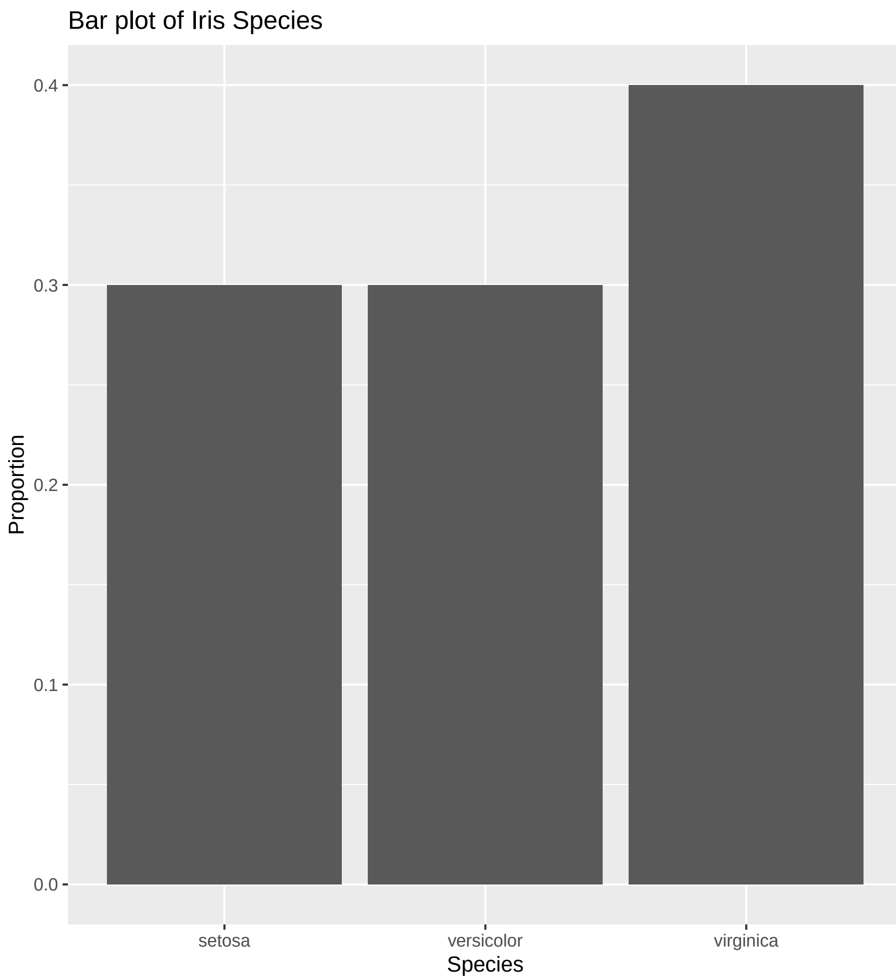
Following data from this table:

1. Count up the number of values per category (make a frequency table). Note: This table is missing the total

2. Plot the frequencies with them as the height of the bars



If needed (if you need proportions for the y-axis instead, calculate the relative frequency table for the frequency table first). Note: again, this one is missing the total



2.2 Stacked Bar Plots and Side-by-Side Bar Graphs

Stacked bar plots show two categorical variables, one on the x-axis/y-axis, and the other as the legend (colours). We will call the variable on the x-axis as the "groups" and the variable on the legend as the "categories."

When constructing these bar plots, we first want to determine which variable goes where (your choice or given choice to you). Then you calculate relative frequencies *per group*

For example, here I have a two-way table detailing the hair and eye colour of some statistics students

```
## Warning: package 'reshape' was built under R version 4.1.3
```

```
##
## Attaching package: 'reshape'
```

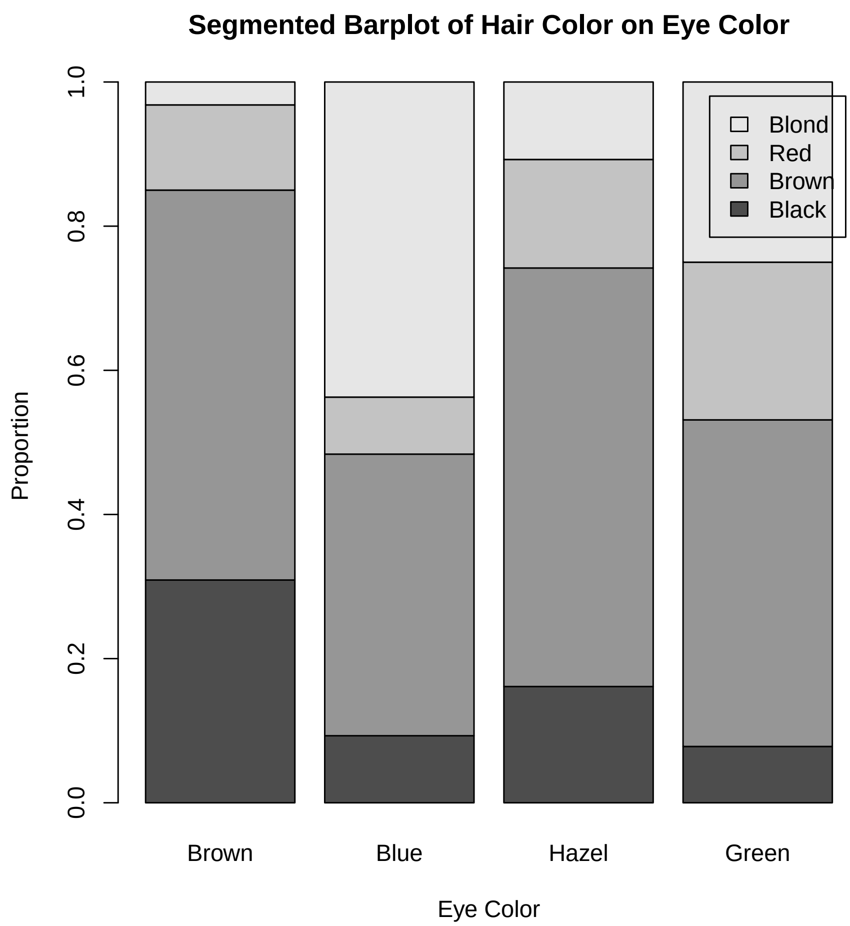
```
## The following object is masked from 'package:dplyr':
##
##     rename
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, smiths
```

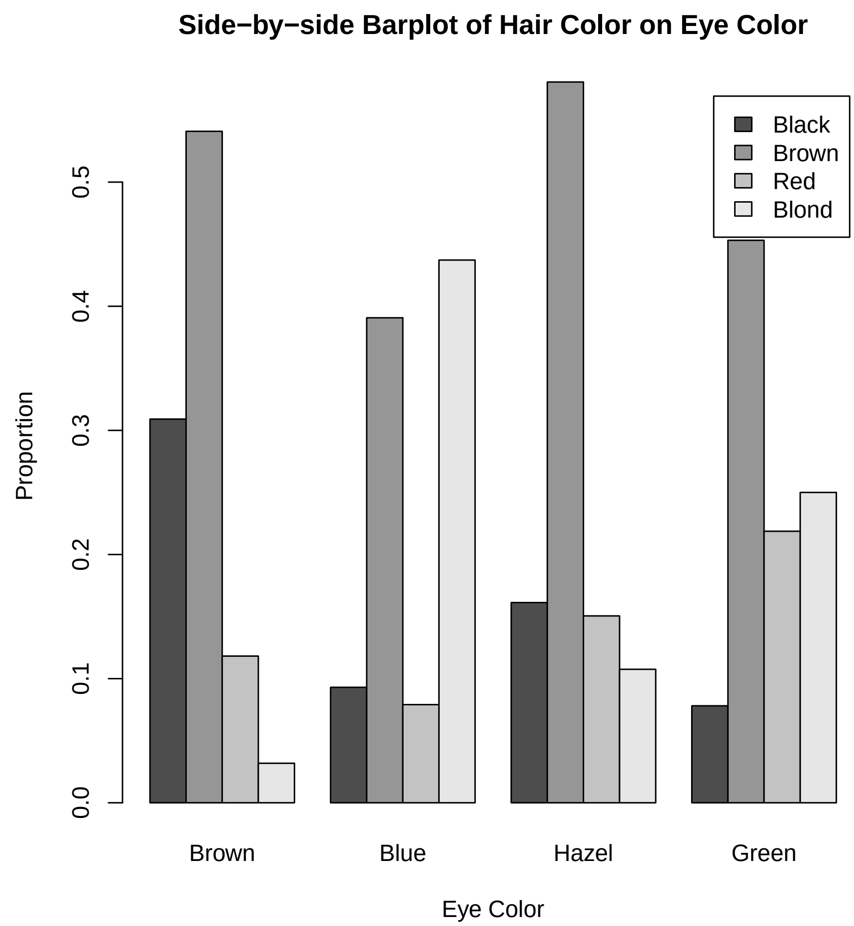
So if I want eye colour to be my groups, I would calculate the relative frequencies by column (use the total of the column and divide the whole column by it), so each group/column will add up to 1.

These numbers will be my bar heights. So for the bar(s) representing brown eyes:

- black hair will be .3091
- brown hair will be .5409
- red hair will be 0.1182
- blond hair will be 0.0318



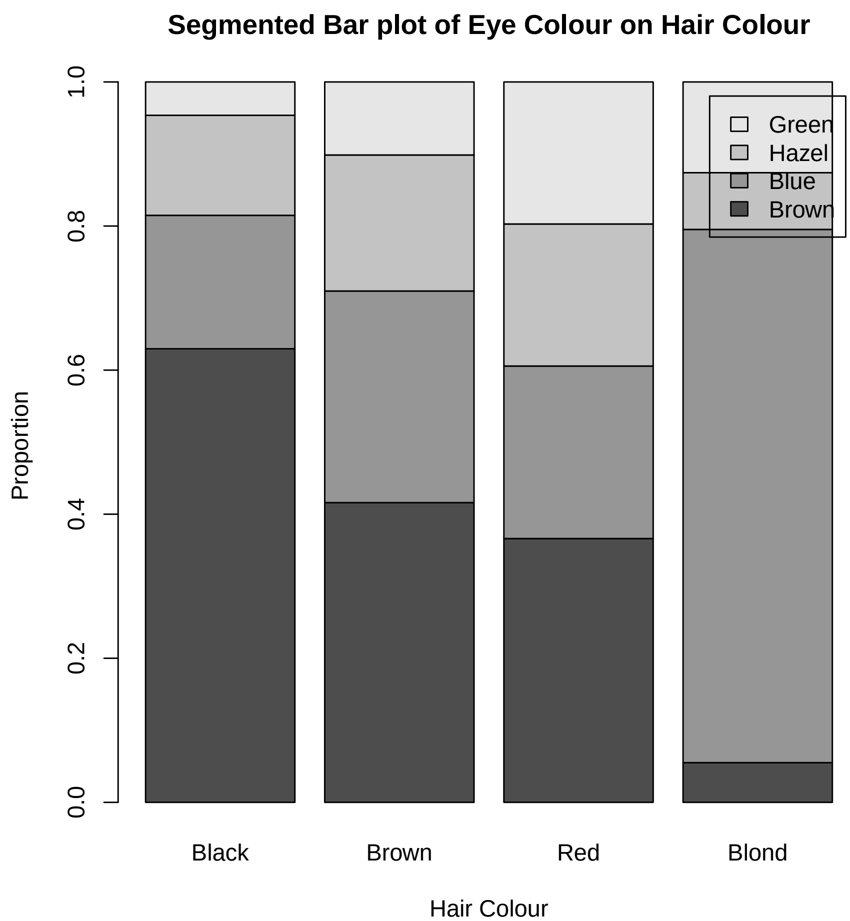
Here's the corresponding side-by-side bar plot. Note that the heights of the bars are the same as the segmented bar graph.



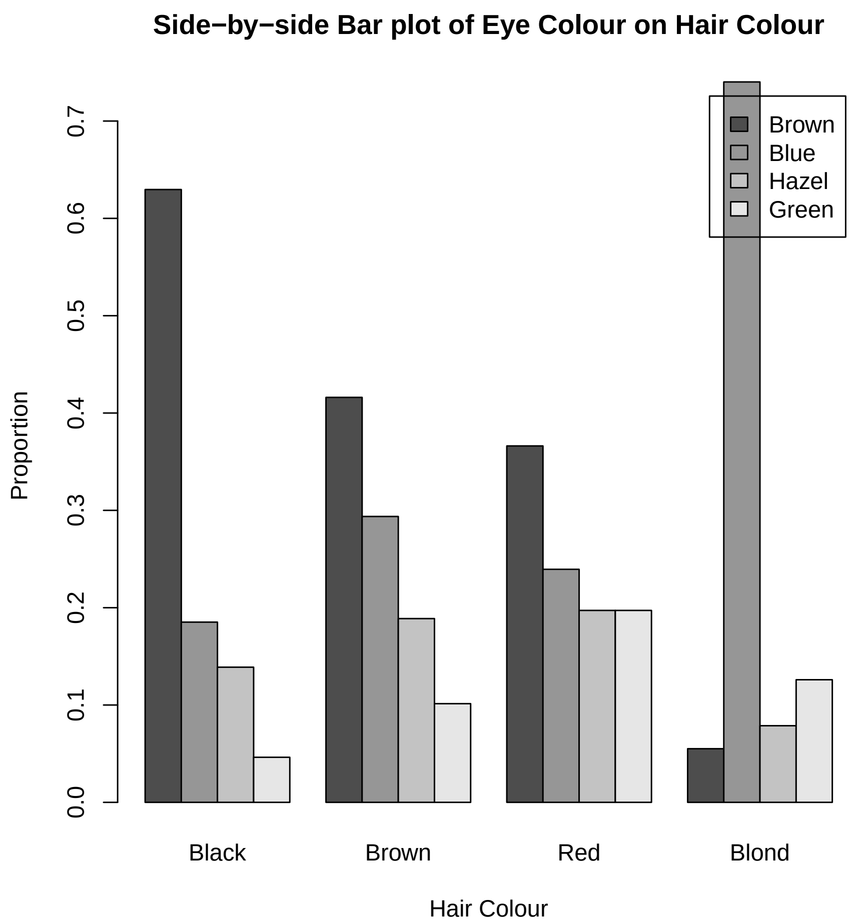
On the other hand, if I want my eye colour to be my groups, I would calculate the relative frequencies by row (use the total of the row and divide the whole row by it), so each group/row will add up to 1.

These numbers will be my bar heights. So for the bar(s) representing black hair:

- brown eyes will be 0.6296296
- blue eyes will be 0.1851852
- hazel eyes will be 0.1388889
- green eyes will be 0.0462963



Segmented Bar plot of Eye Colour on Hair Colour
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
Black
Brown
Red
Blond
Hair Colour
Green
Hazel
Blue
Brown



2.3 Mosaic Plots

Mosaic plots are the almost the same as stacked bar plots. The only difference is that the widths of the bars change according to the proportion of points in each group. In a mosaic plot, the x-axis will also measure the proportion of observations/data points within the groupings (i.e. the x-axis reflects the marginal distribution of the variable on the x-axis).

Following the same steps as the side-by-side and stacked bar charts to find the heights, we now add an additional step before plotting.

Find the widths of the bars by finding the marginal distriubtion of the variable on the x-axis (the groups)

1. For each group, find the probability of having that trait. So for our previous example, we had this table:

Using our eye colours as the groups (vertical bars), we will find:

- $P(Brown) \approx .3716$
- $P(Blue) \approx .3632$
- $P(Hazel) \approx .1571$
- $P(Green) \approx .1081$

When we plot our mosaic plot, we do the same thing, except now, we have our bars differ in widths according to the numbers that we just calculated.

Chapter 3

Quantative Data Visualizations

3.1 Dot Plots

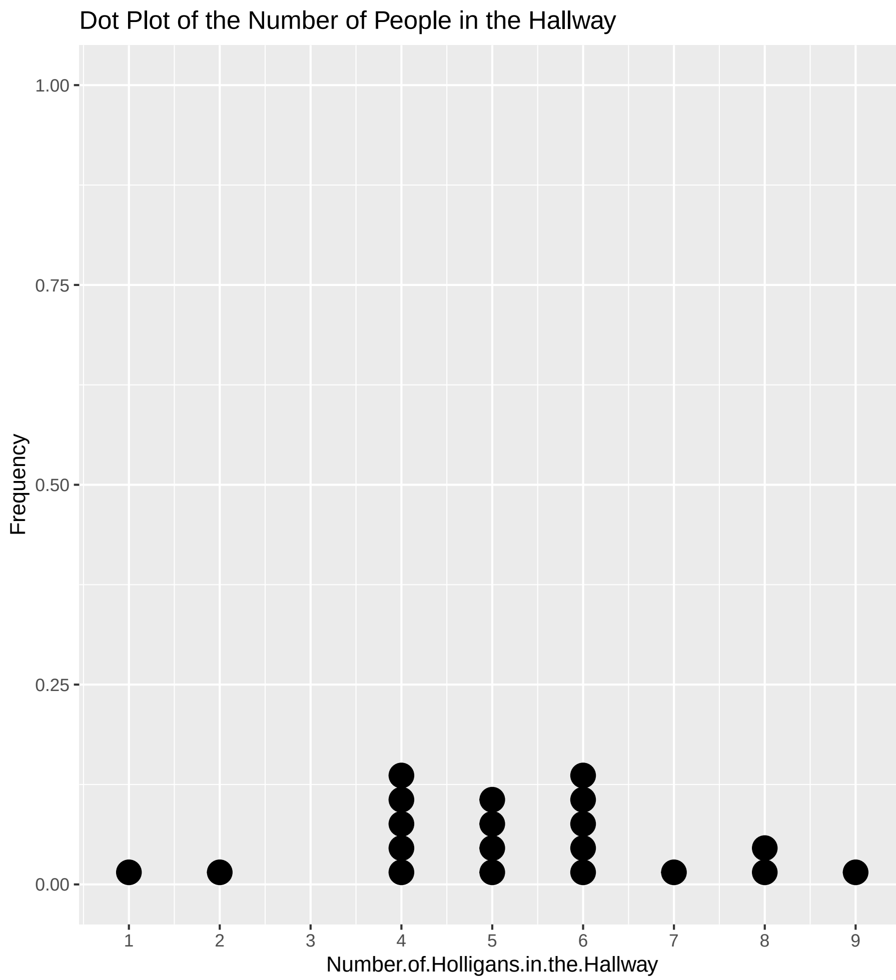
Dot plots are for discrete quantitative variables only, and they are only useful in situations when you have a small range of number so that you can actually see how the data distribution varies across values.

Dot plots are simple, you draw a number line and then plot points above the number for each of the number that you see in the data.

Take this data for example:

Now count up each value to figure out how many dots you need at each value on the number line then plot your graph

```
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
```



3.2 Stemplots

Using this data as an example, a stem plot looks like this:

A stem plot looks like this:

```
## 1 | 2: represents 12
##  leaf unit: 1
##             n: 20
##     0 | 1
##     1 | 8
##     2 |
##     3 | 36689
##     4 | 4
```

```
##      5 | 2339
##      6 | 0228
##      7 |
##      8 | 19
##      9 | 35
```

In a stem plot, you need to determine a common "stem" of all the numbers that you're plotting. So if you have integer numbers from 10 to 200, your stems will be everything from the tens and so on, so you'll have stems from 1-20. Once you take the stems, you just write the "leaves" next to the stem that they belong.

Note that you also have to add a key to show what a stem + leaf means. The stem and leaves give no information on the decimals in the data, so as you see above, you need to give an example like (as shown in the example stemplot):

Key: $1|2 = 12$

Here's another example (sorted for convenience)

```
## 1 | 2: represents 1.2
##  leaf unit: 0.1
##             n: 50
##    3 | 3
##    3 | 579999
##    4 | 00334
##    4 | 556677788899
##    5 | 00112233444
##    5 | 667888999
##    6 | 234
##    6 | 5
##    7 | 12
```

3.3 Boxplots

Also known as a box-and-whisker plot

Boxplots are primarily made of the **five number summary** of the data. The five number summary is made up of the:

- Minimum (min)
- First Quartile (Q_1)
- Median
- Third Quartile (Q_3)
- Maximum (max)

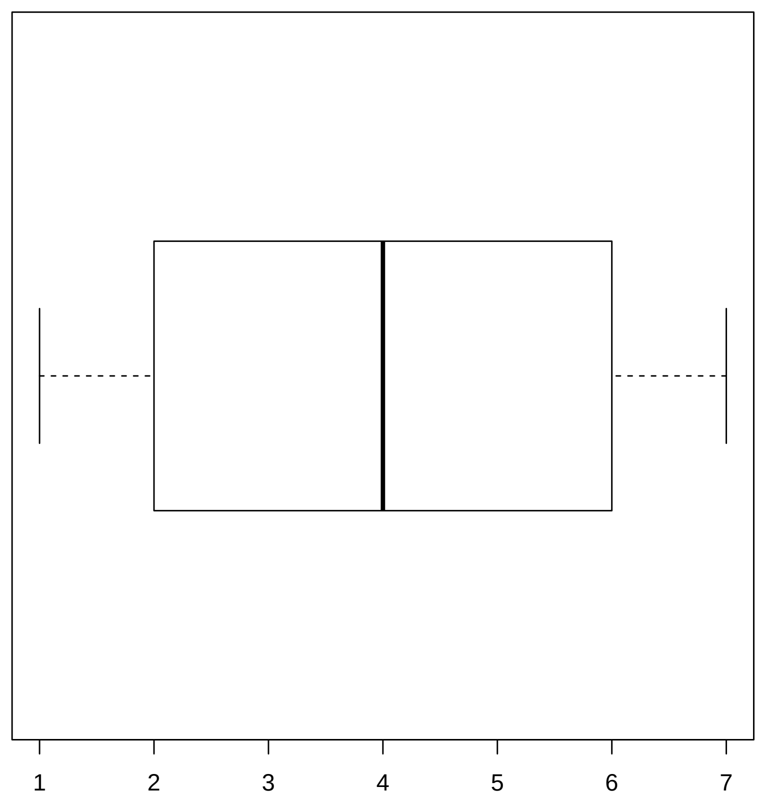
To make a simple boxplot, you use the first quartile, median, and third quartile to make the "box" and then use the minimum and maximum to make the "whiskers."

For this simple list of numbers:

Our five number summary is:

```
##    Min. 1st Qu.  Median 3rd Qu.    Max. 
##       1       2       4       6       7
```

As detailed above, our box plot then looks like:



The last detail is that we can calculate outliers using the 1.5 IQR rule and show them on the boxplot. For either direction (left or right), if we see outliers in that direction, we only extend the whisker to the smallest and/or largest point that is not an outlier. Then we plot any outliers as individual points.

Look at this example data:

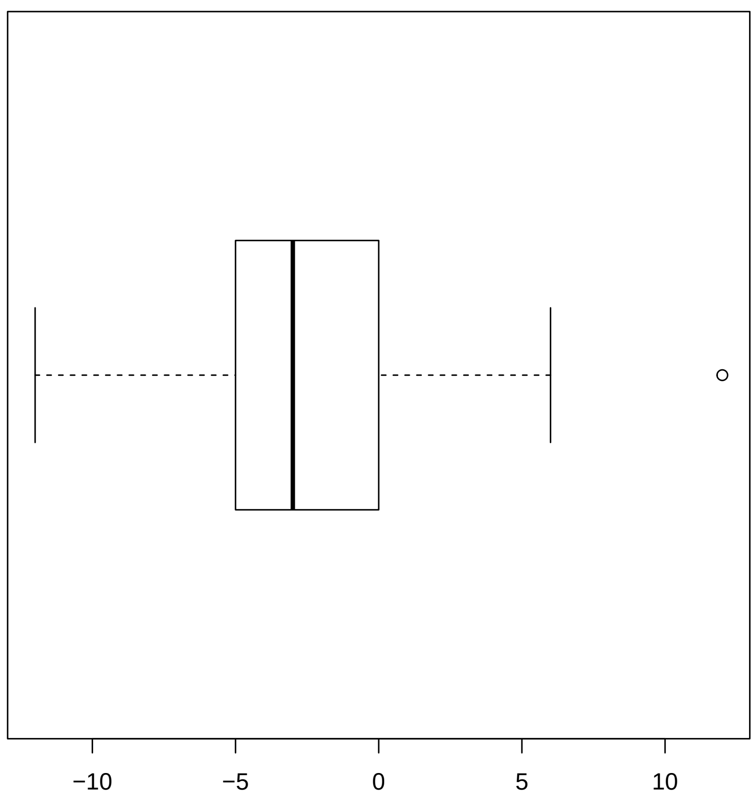
Five number summary:

```
##    Min. 1st Qu.  Median 3rd Qu.    Max.
##     -12      -5      -3       0      12
```

Our numbers calculated by the 1.5 IQR rule are:

```
## [1] -12.5   7.5
```

So our 12 is an outlier. which means we draw our right whisker to 6 and plot the 12 individually on the number line. Like so:



3.4 Histograms

A histogram is similar to a bar plot, except that histograms are made for quantitative data and *bars are continuous* in the sense that there is no gap between bars. To make a histogram, select an appropriate equal intervals that make it so that you don't have too many bars and that you don't have too few bars. Your goal with histograms, as with many other visualizations, is to be able to see the shape and characteristics of the distribution in question. If you have too many bars or too few bars, you won't be able to see much important information (especially think of situations when you have many data points with very precise decimal measurements).

1. Decide on your intervals (e.g. by 5's, by 10's, by 100's)

2. Within your intervals, count up the number of observations that belong in that "bin". When you do so, count up observations so that you count the left end inclusive and the right end inclusive. So if you did intervals of 5, you would do something like counting up points $0 \leq x < 5$, $5 \leq x < 10$, and so on.

3. Plot your bars.

Example:

Consider this example data set:

Our data has this set of summary statistics:

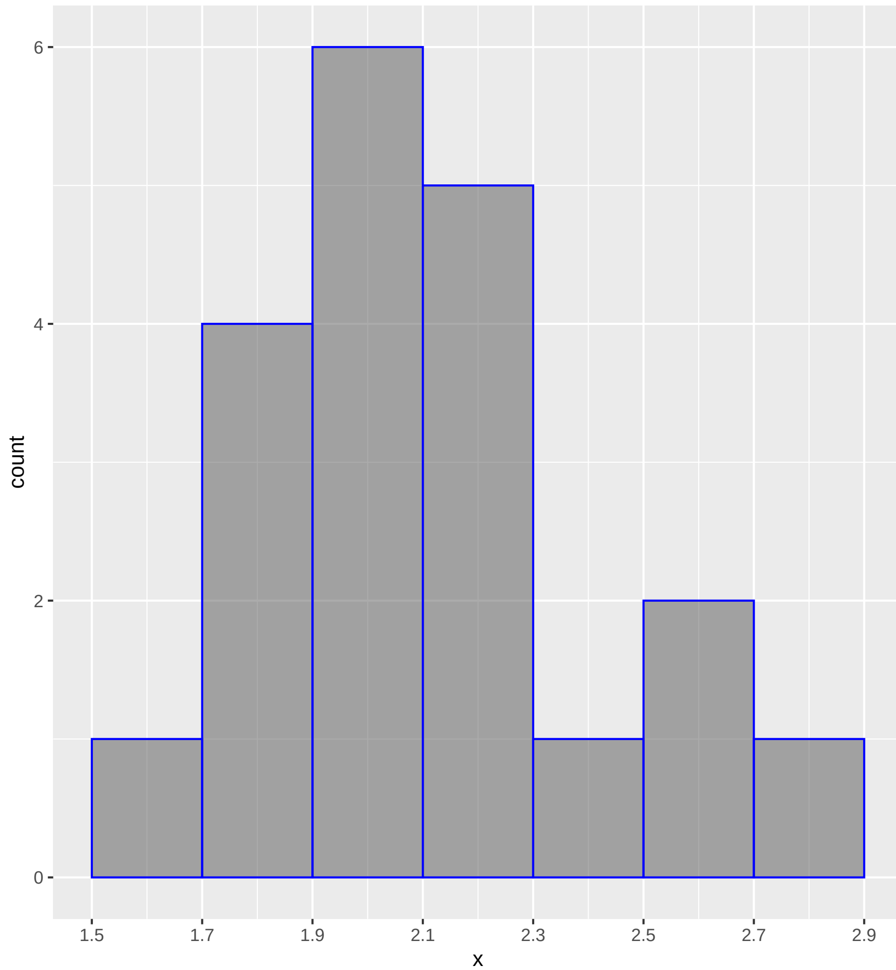
```
##        x
##  Min.   :1.522
##  1st Qu.:1.912
##  Median :2.022
##  Mean   :2.110
##  3rd Qu.:2.224
##  Max.   :2.704
```

With this knowledge, let's make our 7 "bins", so let's do these by every 0.2, starting at 1.5 to 2.9. This will be something that you build by intuition.

Now, count up our values:

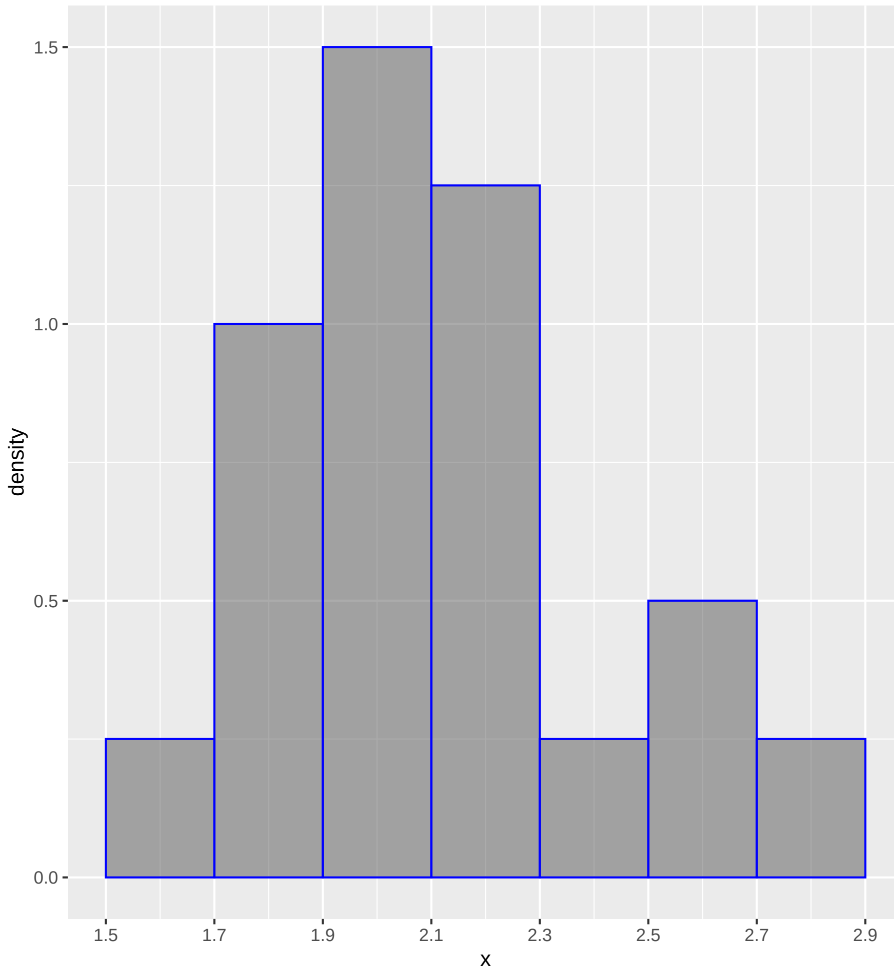
```
## [1.5,1.7) [1.7,1.9) [1.9,2.1) [2.1,2.3) [2.3,2.5) [2.5,2.7) [2.7,2.9]
##         1         4         6         5         1         2         1
```

Now, we just put it together. For each bin, we have a bar and the bars' heights correspond to the number of individuals in each bin.



Again, just like bar graphs, we can instead do the relative frequencies (this is what you'll see most of the time!!!)

```
## [1.5,1.7) [1.7,1.9) [1.9,2.1) [2.1,2.3) [2.3,2.5) [2.5,2.7) [2.7,2.9]
##      0.05      0.20      0.30      0.25      0.05      0.10      0.05
```



When you have a histogram like this, keep in mind that the bars always add up to 1 (or 100%).

Chapter 4

Sampling

When we have a question about a certain population (an entire group of individuals). Ideally we would ask them all (take a **census**). But contacting every member of a population often isn't very practical: it would take too much time and cost too much money. Instead, we put the question to a **sample**, or subset of individuals of the population from which we actually collect data, chosen to represent the entire population.

When you want to identify the population, ask yourself, what does the question want to know about? What group of people does the question/problem address?

When identifying the sample, ask yourself, what group does the work done actually address?

4.1 Bias

When we collect data, there is the possibility of the data becoming systematically pushed towards a specific outcome. For example, if we want to learn about the GPA average in the school and take a sample of students only from a class, it's quite possible that the sample is not representative of the school. We will probably result in a GPA average that is higher or lower than the actual GPA average in the school. There are several ways that this can happen. The main way that we learn are:

4.1.1 Response Bias

4.1.2 Non-response Bias

4.1.3 Voluntary Bias

Chapter 5

Probability

A process has a valid probability model if and only if:

- Each outcome has a positive probability.
- The sum of all outcomes' probabilities is equal to 1.

Chapter 6

Random Variables

A **random variable** is a variable whose value is a numerical outcome of a random phenomenon.

Random variables can be discrete or continuous. A **discrete random variable** X has a countable, or smaller, number of possible outcomes, while a **continuous random variable** X can take on an infinite number (theoretically) of different values.

The **probability distribution** of a random variable X gives us all possible values of X, and their corresponding probabilities. Probability distributions are typically given as tables, histograms (with probability on the y-axis, instead of frequency), or density curves (like the Normal curve).

6.1 Discrete Random Variables

A discrete random variable describes a process that only has specific, predefined outcomes. For example, you can be finding the probability of people having blue, brown, and black eyes. Or you can be finding the probability of people earning salaries in the ranges of <\$30,000, \$30,000 - \$50,000, \$50,000 - \$ 70,000, \$70,000 - \$ 100,000, >\$100,000.

When we have a discrete random variable X whose probability distribution is

Value:	x_1	x_2	x_3	$\cdots$
Probability:	p_1	p_2	p_3	$\cdots$

we know the following about the mean and standard deviation of X:

$$\mu_X = E(X) = \sum x_i P(x_i)$$

$$\sigma_X = \sqrt{\sum \left(x_i - \mu_X\right)^2 P\left(x_i\right)}$$

Keep in mind that the standard deviation formula here (and else where) is equivalent to the formula of population standard deviation when we divide by n instead of multiplying by $P(x_i)$.

$$\begin{aligned}
\sigma_X &= \sqrt{\frac{\sum(x_i - \mu_X)^2}{n}} && (6.1)\\
&= \sqrt{E((X - \bar{x})^2)} && (6.2)\\
&= \sqrt{\sum \left(x_i - \mu_X\right)^2 P\left(x_i\right)} && (6.3)\\
& && (6.4)
\end{aligned}$$

In other words, remember that we define standard deviation as the root mean square of the squared differences from the mean. Since we are taking the mean of the squared differences from the mean in both formulas (I use the definition for mean from step 1 to 2 and the definition of mean for discrete variables from step 2 to 3), the two formulas are equivalent.

The variance of a random variable X is:

$$Var(X) = \sigma_X^2$$

6.1.1 Binomial Random Variables

Binomial random variables have parameters n and p, and can be written $B(n, p)$. Remember, Normal random variables have parameters and and can be written $N(\mu, \sigma)$.

The pdf of a Binomial Random Variable (i.e. the binomial formula) is:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$
$$\text{where } k = 0, 1, 2, 3, \cdots, n$$

To apply this formula in a graphing calculator: `2nd -> vars (distr) -> binompdf`

Usage: `binompdf(n, p[,x])`

The cdf of a Binomial Random Variable is:

$$P(X \leq k) = \sum_{i=0}^{n} \binom{n}{i} p^i (1 - p)^{n-i}$$

In graphing calculator: `2nd -> vars (distr) -> binomcdf`

Usage: `binomcdf(n, p[,x])`

The mean and standard deviation of a binomial random variable is given by:

$$\mu_X = np$$
$$\sigma_X = \sqrt{npq}$$

, where $q = 1 - p$.

6.1.1.1 Binomial setting

We can identify a binomial setting when we know:

1. **B**inary? The possible outcomes of each trial can be classified as "success" or "failure" (in our case rolling a 7 or not).
2. **I**ndependent? Trials must be independent; that is, knowing the result of one trial must not tells us anything about the result of another trial.
3. **N**umber? The number of trials n of the chance process must be fixed in advance. (this was 5, then 100).
4. **S**ame? There is the same probability p of success on each trial (1/6).

6.1.1.2 10% Condition

The second condition is often not perfectly met, as in the case of an SRS from some population. Imagine choosing 10 students from a class of 15 females and 15 males—as we choose people, the remaining population changes, which changes the probability that the next person chosen will be male or female.

When we lack complete independence, we can see the consequence of this is negligible as long as our sample is small relative to the population from which we are sampling. If we were choosing our 10 people from a school of 3,300 students, the change in probability from person to person would be small enough to ignore.

The general rule is that the sample needs to be less than $\frac{1}{10}$, or 10%, of the population. We refer to this as the 10% condition.

$$n \leq (.10)N$$

6.1.1.3 Normal Approximation to the Binomial Distribution

Remember that as n gets large, a binomial random variable X can take on more and more different values, and it can become tedious to continue to treat X as a discrete random variable. As n get larger, we can treat X as a continuous random variable, more specifically:

As n gets larger, the binomial distribution gets closer to a normal distribution. However, before we use a normal distribution to approximate a binomial distribution, we have to check the following condition:

6.1.1.3.1 Large Counts condition If $np \geq 10$ and $n(1-p) \geq 10$, then we can use a Normal distribution to model a binomial distribution. In other words, if the expected number of successes and failures (respectively) is greater than or equal to 10.

6.1.1.3.2 Doing a normal approximation First verify all the conditions for a Binomial setting and the Large Counts Condition. Since we know that we have a binomial setting, we then know the distribution that we want to use is $Normal(np, \sqrt{npq})$ proceed with the calculations according to this distribution.

6.1.2 Geometric Random Variables

If $X \sim G(p)$, in other words, if X has a geometric distribution with parameter p, the pdf of a geometric random variable is:

$$P(X = x) = (1-p)^{x-1}p$$
$$\text{where } x = 1, 2, 3, \cdots$$

In graphing calculator: `2nd -> vars (distr) -> geometpdf`

Usage: `geometpdf(p, x)`

The cdf of a Geometric Variable is:

$$P(X \leq x) = \sum_{i=1}^{x}(1-p)^{i-1}p$$

In graphing calculator: `2nd -> vars (distr) -> geometcdf`

Usage: `geometcdf(p, x)`

The mean and standard deviation of a geometric random variable is given by:

$$\mu_X = \frac{1}{p}$$
$$\sigma_X = \frac{\sqrt{q}}{p}$$
$$\text{, where } q = 1-p.$$

6.1.2.1 The Geometric Setting

A geometric setting is very similar to a binomial setting, except that **n, the number of trials is not fixed**.

A geometric setting is defined as a series of observations where these 4 conditions are met:

1. **B**inary? The Possible outcomes of each trial can be classified as "success" or "failure"
2. **I**ndependent? Trials must be independent, that is, knowing the result of one trial must not have any effect on the result of any other trial.
3. **T**rials? The goal is to count the number of trials until the first success occurs.
4. **S**uccess? On each trial, the probability p of success must be the same.

6.2 Operations with Random Variables

6.2.1 Constants

When we add a constant a and/or multiply by a constant b to a random variable X, we perform a linear transformation of the form

$$a + bX$$

The mean of the transformed variable is:

$$\mu_{a+bX} = a + \mu_{bX} = a + b\mu_X$$

The standard deviation of the transformed variable is:

$$\sigma_{a+bX} = \sigma_{bX} = b\sigma_X$$

6.2.2 Random Variables

In general, we can describe the mean and standard deviation of the sum or difference of independent random variables with these formulas:

$$\mu_{X\pm Y} = \mu_X \pm \mu_Y$$

If the random variables X and Y are independent, then

$$\sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y$$

Chapter 7

Hello bookdown

All chapters start with a first-level heading followed by your chapter title, like the line above. There should be only one first-level heading (`#`) per .Rmd file.

7.1 A section

All chapter sections start with a second-level (`##`) or higher heading followed by your section title, like the sections above and below here. You can have as many as you want within a chapter.

An unnumbered section

Chapters and sections are numbered by default. To un-number a heading, add a `{.unnumbered}` or the shorter `{-}` at the end of the heading, like in this section.

Chapter 8

Cross-references

Cross-references make it easier for your readers to find and link to elements in your book.

8.1 Chapters and sub-chapters

There are two steps to cross-reference any heading:

1. Label the heading: `# Hello world {#nice-label}`.
    - Leave the label off if you like the automated heading generated based on your heading title: for example, `# Hello world` = `# Hello world {#hello-world}`.
    - To label an un-numbered heading, use: `# Hello world {-#nice-label}` or `{# Hello world .unnumbered}`.
2. Next, reference the labeled heading anywhere in the text using `\@ref(nice-label)`; for example, please see Chapter 8.
    - If you prefer text as the link instead of a numbered reference use: any text you want can go here.

8.2 Captioned figures and tables

Figures and tables *with captions* can also be cross-referenced from elsewhere in your book using `\@ref(fig:chunk-label)` and `\@ref(tab:chunk-label)`, respectively.

See Figure 8.1.

Don't miss Table 8.1.

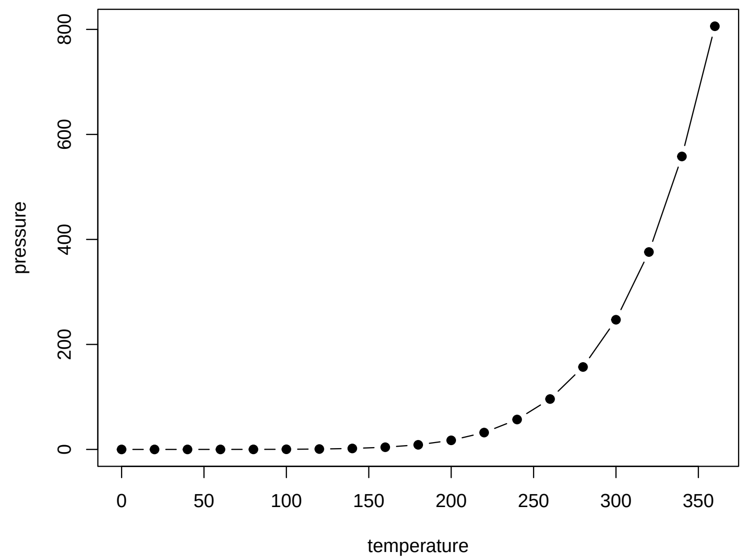


Figure 8.1: Here is a nice figure!

Table 8.1: Here is a nice table!

temperature	pressure
0	0.0002
20	0.0012
40	0.0060
60	0.0300
80	0.0900
100	0.2700
120	0.7500
140	1.8500
160	4.2000
180	8.8000

Chapter 9

Parts

You can add parts to organize one or more book chapters together. Parts can be inserted at the top of an .Rmd file, before the first-level chapter heading in that same file.

Add a numbered part: `# (PART) Act one {-}` (followed by `# A chapter`)

Add an unnumbered part: `# (PART\*) Act one {-}` (followed by `# A chapter`)

Add an appendix as a special kind of un-numbered part: `# (APPENDIX) Other stuff {-}` (followed by `# A chapter`). Chapters in an appendix are prepended with letters instead of numbers.

Chapter 10

Footnotes and citations

10.1 Footnotes

Footnotes are put inside the square brackets after a caret `^[]`. Like this one [1].

10.2 Citations

Reference items in your bibliography file(s) using `@key`.

For example, we are using the **bookdown** package [Xie, 2023] (check out the last code chunk in index.Rmd to see how this citation key was added) in this sample book, which was built on top of R Markdown and **knitr** [Xie, 2015] (this citation was added manually in an external file book.bib). Note that the `.bib` files need to be listed in the index.Rmd with the YAML `bibliography` key.

The RStudio Visual Markdown Editor can also make it easier to insert citations: https://rstudio.github.io/visual-markdown-editing/#/citations

[1] This is a footnote.

Chapter 11

Blocks

11.1 Equations

Here is an equation.

$$f\left(k\right) = \binom{n}{k} p^k \left(1 - p\right)^{n-k} \tag{11.1}$$

You may refer to using `\@ref(eq:binom)`, like see Equation (11.1).

11.2 Theorems and proofs

Labeled theorems can be referenced in text using `\@ref(thm:tri)`, for example, check out this smart theorem 11.1.

Theorem 11.1. *For a right triangle, if c denotes the* length *of the hypotenuse and a and b denote the lengths of the* ***other*** *two sides, we have*

$$a^2 + b^2 = c^2$$

Read more here https://bookdown.org/yihui/bookdown/markdown-extensions-by-bookdown.html.

11.3 Callout blocks

The R Markdown Cookbook provides more help on how to use custom blocks to design your own callouts: https://bookdown.org/yihui/rmarkdown-cookbook/custom-blocks.html

Chapter 12

Sharing your book

12.1 Publishing

HTML books can be published online, see: https://bookdown.org/yihui/bookdown/publishing.html

12.2 404 pages

By default, users will be directed to a 404 page if they try to access a webpage that cannot be found. If you'd like to customize your 404 page instead of using the default, you may add either a `_404.Rmd` or `_404.md` file to your project root and use code and/or Markdown syntax.

12.3 Metadata for sharing

Bookdown HTML books will provide HTML metadata for social sharing on platforms like Twitter, Facebook, and LinkedIn, using information you provide in the `index.Rmd` YAML. To setup, set the `url` for your book and the path to your `cover-image` file. Your book's `title` and `description` are also used.

This `gitbook` uses the same social sharing data across all chapters in your book-all links shared will look the same.

Specify your book's source repository on GitHub using the `edit` key under the configuration options in the `_output.yml` file, which allows users to suggest an edit by linking to a chapter's source file.

Read more about the features of this output format here:

https://pkgs.rstudio.com/bookdown/reference/gitbook.html

Or use:

Bibliography

Yihui Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition, 2015. URL http://yihui.org/knitr/. ISBN 978-1498716963.

Yihui Xie. *bookdown: Authoring Books and Technical Documents with R Markdown*, 2023. URL https://CRAN.R-project.org/package=bookdown. R package version 0.32.