3 Descriptive Statistics, Models, and Distributions

3.1. Models

This book teaches you how to construct statistical models. A model is a simplified representation of a system. For example, the map of a city represents a city in a simplified fashion. A map providing as much detail as the original city would not only be impossible to construct, it would also be pointless. Humans build models, such as maps and statistical models, to make their lives simpler.

Imagine having conducted a reading time experiment that involves measurements from 200 participants. If you wanted to report the outcome of your experiment to an audience, you wouldn't want to talk through each and every data point. Instead, you report a summary, such as 'The 200 participants read sentences with an average speed of 2.4 seconds', thus saving your audience valuable time and mental energy. This chapter focuses on such summaries of numerical information, specifically, the mean, the median, quantiles, the range, and the standard deviation. The mean is a summary of a distribution. What exactly is a distribution?

3.2. Distributions

Imagine throwing a single die 20 times in a row. For a regular die, there are six possible outcomes. For any one throw, each face is just as likely to occur as any other. Let's perform 20 throws and note down how frequently each face occurs. The face '1' comes up 2 times, '2' might come up 5 times, and so on. The result of tallying all counts is a 'frequency distribution', which associates each possible outcome with a particular frequency value. The corresponding histogram is shown in Figure 3.1a (for an explanation of histograms, see Chapter 1.12).

The distribution in Figure 3.1a is an empirically observed distribution because it is based on a set of 20 actual throws of a die ('observations'). Figure 3.1b on the other hand is a theoretical distribution, one that isn't attested in any specific dataset. Notice how the *y*-axis in Figure 3.1b represents probability rather than frequency. Probabilities range from 0 to 1, with 0 indicating that a given event never occurs and 1 indicating that a given event always occurs. The theoretical probability distribution in Figure 3.1b answers the question: how probable is each outcome? In this case, all outcomes are equally probable, namely, 1 divided by 6, which is about 0.17. In other

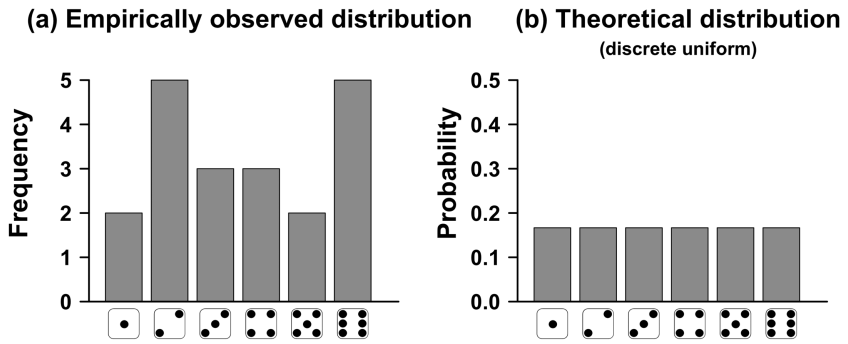

Figure 3.1. (a) An empirically observed distribution based on 20 throws of a die; (b) A theoretical distribution displaying the expected probabilities for an infinite number of throws

words, each face is expected to occur one sixth of the time, although any particular set of throws may deviate from this theoretical expectation.

Commonly used theoretical distributions have names. The particular distribution shown in Figure 3.1b is the 'discrete uniform distribution'. It is a 'uniform' distribution because the probability is uniformly spread across all possible outcomes. It is furthermore a 'discrete' distribution because there are only six particular outcomes and no in-betweens.

Applied statistics involves both empirical distributions and theoretical distributions. The theoretical distributions are tools that help modeling empirically observed data. Most models constructed in applied statistics *assume* that the data has been generated by a process following a certain distribution. In order to model various types of data, you have to learn about various theoretical distributions. This chapter introduces the normal distribution. Later chapters introduce the Bernoulli distribution (Chapter 12) and the Poisson distribution (Chapter 13).

3.3. The Normal Distribution

One of the most common distributions in statistics is the 'normal distribution', also known as the 'bell curve' due to its characteristic bell shape. A more technical name for this distribution is the Gaussian distribution, after the mathematician Carl Friedrich Gauss. The normal distribution is a distribution for continuous data, centered symmetrically around the mean with the bulk of data lying close to the mean.

Figure 3.2a shows three distributions of actual data that are approximately normally distributed. To make this example more concrete, you can imagine that these are language test scores from three different classrooms. Each of the three distributions has a different mean. Classroom A has a mean of 10 (this class performed badly overall), B has a mean of 40, and C has a mean of 80 (this class performed very well). In such a scenario, you can think of the mean as specifying the 'location' of

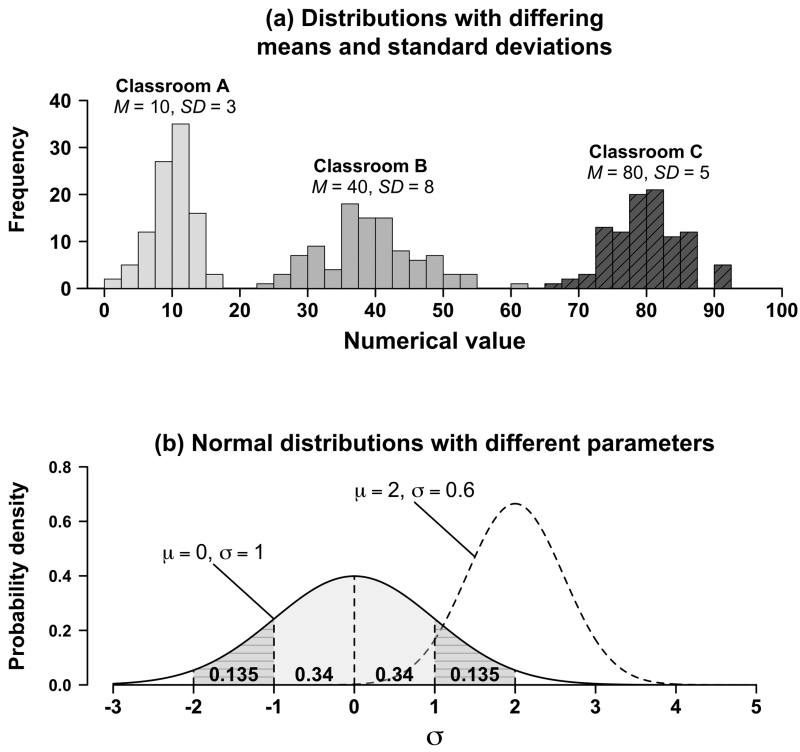

Figure 3.2. (a) Three distributions for groups of students with 100 students each; the data is random data that was generated based on an underlying normal distribution with the specified means and standard deviations. (b) Two normal distributions with different parameters; the parameter μ ('mu') specifies the location of the distribution on the number line; σ ('sigma') specifies the spread; the highlighted areas under the curve show the 68% and 95% intervals

a distribution on the *x*-axis. That is, the mean tells you how large or small a set of numbers is overall.[1]

The distributions in Figure 3.2a also differ in terms of 'spread'. In particular, the distribution for classroom B is wider than the other two distributions. This means that there are more students farther away from the mean in this classroom. Students in this classroom are more different from each other than the students in the other

1 The notation $\bar{x}$ ('x bar') is often used to represent the mean of a set of numbers *x*. The formula for the mean is $\bar{x} = \frac{\sum x}{N}$. Formulas are summary formats for computational procedures. The $\sum x$ in the numerator stands for the procedure that sums up all the numbers. The sum is then divided by *N*, which represents how many numbers there are. For example, imagine collecting three response durations in a psycholinguistic experiment, specifically 300*ms*, 200*ms*, and 400*ms*.The sum of these numbers is 900*ms*, with *N* = *3*, which yields $\frac{900ms}{3} = 300ms$.

classrooms. The standard deviation (*SD*) is used to summarize this spread. Although the actual formula has some quirks,[2] it is OK for the most part to think of the standard deviation as representing the average distance from the mean. Visually, larger standard deviations correspond to flatter histograms that fan out farther away from the mean.

While the mean and standard deviation can be computed for any data, these two numbers have special meaning in the context of the normal distribution: they are the distribution's 'parameters'. A parameter is a property of a distribution. Changing the parameter changes the look of the distribution. Changing the mean moves the normal distribution along the number line, and changing the standard deviation stretches or squeezes the distribution.

Figure 3.2b shows the normal distribution for two different sets of parameters. The *y*-axis label says 'probability density', which means that the height of the graph indicates how probable certain values are. The area under each bell curve adds up to 1.0. You may also have noticed that I tacitly switched notation. Before, I used Roman letters to represent the mean (*M* or $\bar{x}$) and the standard deviation (*SD* or *s*). In Figure 3.2b, I used the Greek letters μ ('mu') and σ ('sigma') instead. It is conventional to use Greek letters when talking about the parameters of theoretical distributions. When talking of empirically observed data, it is conventional to use Roman letters instead. This is not just an arbitrary convention—you will later see that this notation is quite important.

Figure 3.2b also shows an important property of the normal distribution: the areas highlighted with '0.34' add up to a probability of 0.34, and so on. Together, the two middle parts add up to $p = 0.68$. If you were to randomly draw numbers from this distribution, 68% of the time, you would end up with a number that is between −1 and +1 standard deviations. If you add the next two areas (striped) to this, the probability mass under the curve adds up to $p = 0.95$. Thus, if you were to draw random data points from this distribution, 95% of the time you would end up with a number that is between −2 and +2 standard deviations.

The 68%–95% 'rule' allows you to draw a mental picture of a distribution from the mean and standard deviation alone, granted the distribution is approximately normally distributed. Let us apply the rule to a dataset from Warriner, Kuperman, and Brysbaert (2013), who asked native English speakers to rate the positivity or negativity of words on a scale from 1 to 9 ('emotional valence'). In their paper, the authors report that the mean of the ratings was 5.06, with a standard deviation of 1.27. Assuming normality, you can expect 68% of the data to lie between 3.79 and 6.33 (5.06 − 1.27, 5.06 + 1.27). You can furthermore expect 95% of the data to lie between 2.52 and 7.6. To calculate

2 The formula for the standard deviation requires that you calculate the mean first, as the standard deviation measures the spread from the mean. The formula works as follows: you calculate each data point's difference from the mean, square that value, and subsequently sum up the squared differences. This 'sum of squares' is then divided by *N* minus *1* (so *99*, if you have *100* data points). Subsequently, you take the square root of this number, which yields the standard deviation. There are reasons for why the square root has to be used and why one has to divide by $N - 1$ (rather than by *N*) that I will not go into here. The abbreviated formula for the standard deviation of a set of numbers, *x*, is: $SD = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$. When you see a formula like this, it is important to try to understand it piece by piece, and to first focus on those symbols that you recognize. For example, you may observe in this formula that there's an $\bar{x}$, which represents the mean, and that this number is subtracted from each *x*-value. Mathematicians use formulas not to make things more difficult, but to make things *easier* for them. A formula encapsulates a lot of information in a neat way.

this, you have to double the standard deviation 1.27, which yields 2.54. Then you add and subtract this number from the mean (5.06 − 2.54, 5.06 + 2.54).

Finally, it is worth noting that, because the mean and standard deviation work so nicely together, it's generally a good idea to report both. In particular, means without a measure of spread are not very informative.

3.4. Thinking of the Mean as a Model

I invite you to think of the mean as a model of a dataset. For one, this highlights the compressive nature of the mean (given that models are simplified representations). Second, it highlights that the mean is a representation of something, namely, a distribution.

Moreover, thinking of the mean as a model highlights that the mean can be used to make predictions. For example, Warriner et al. (2013) rated 'only' 14,000 English words, even though there are many more words in the English language. The word *moribund* is one of the words that has not been rated. In the absence of any information, can we predict its emotional valence value? Our best guess for this word's value is the mean of the current sample, 5.06. In this sense, the mean allows making predictions for novel words.

Introductory statistics courses often distinguish between 'descriptive statistics' and 'inferential statistics'. Whereas descriptive statistics is understood to involve things like computing summary statistics and making plots, inferential statistics is generally seen as those statistics that allow us to make 'inferences' about populations of interest, such as the population of all English speakers, or the 'population' of all English words. However, the distinction between descriptive and inferential statistics is not as clear-cut. In particular, any description of a dataset can be used to make inferences. Moreover, all inferential statistics are based on descriptive statistics.

In fact, in some sense, you have already performed some form of inferential statistics in this chapter. The Warriner et al. (2013) dataset can be treated as a *sample* of words that is taken from the *population* of all English words. In applied statistics, we almost always deal with samples as the population is generally not available to us. For example, it may be infeasible to test all English speakers in a psycholinguistic experiment and hence we have to resort to a small subset of English speakers to estimate characteristics of the population.

Samples are used to *estimate* population parameters.[3] The distinction between sample estimates and population parameters is enshrined in mathematical notation. As mentioned above, parameters are conventionally represented with Greek letters; sample estimates, with Roman letters. Then, one can say that the sample mean $\bar{x}$ estimates the population parameter μ. Similarly, the sample standard deviation s estimates the population parameter σ. Other texts may use the caret symbol for this distinction, in which case $\hat{\mu}$ estimates μ, and $\hat{\sigma}$ estimates σ.

From now on, whenever you see means and standard deviations in published papers, ask yourself questions such as the following. What is this mean estimating? What is the relevant population of interest? In later chapters, you will quantify the degree of uncertainty with which sample estimates reflect population parameters.

3 This book is focused on a branch of statistics that is called 'parametric statistics'. Just so you've heard about it: there is a whole other branch of statistics that deals with 'non-parametric statistics', which as the name suggests, does not estimate parameters.

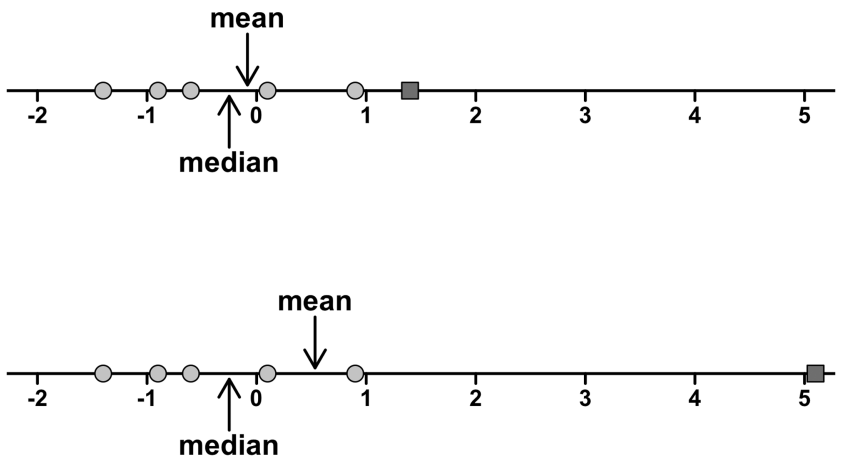

Figure 3.3. Six data points on the number line. The data point represented by the square shifts between the two datasets, pulling the mean upwards. The median stays put because the fact that 50% of the data is on either side hasn't changed by altering the value.

3.5. Other Summary Statistics: Median and Range

The mean and the standard deviation are just two ways of compressing data. Another summary statistic is the median, which is the halfway point of the data (50% of the data are above the median; 50% of the data are below). In contrast, the mean is the 'balance point', which you can think of as a pair of scales: an extreme value shifts the mean, just as a heavy object tips a pair of scales. This is exemplified in Figure 3.3.[4]

The fact that the median *doesn't* move when the position of the square is changed in Figure 3.3 can be seen as an advantage, as well as a disadvantage. The mean incorporates more information because it cares about the actual values of the data points. The median only cares about its position in an ordered sequence.

The median is sometimes reported because it is more robust to variations in extreme values than the mean. For example, most people have relatively low incomes, but some people (such as Bill Gates) have incredibly high incomes. Such extremely rich people skew the mean upwards, but they don't shift the median up as much.

Another summary statistic for the spread is the range, which is the distance between the smallest and the largest data point. For the Warriner et al. (2013) dataset, subtracting the minimum (1.26) from the maximum (8.53) yields the range 7.27. As a general measure of spread, the range is not as useful because it exclusively relies on the two most extreme numbers, ignoring all others. If one of these numbers happens

4 For an uneven number of data points, there is a true middle number. However, for an even number of data points, there are two numbers in the middle. What is the halfway point in this case? The median is defined to be the mean of the two middle numbers, e.g., for the numbers 1, 2, 3, and 4, the median is 2.5.

to change, the range will change as well. However, many times it is useful to know specifically what the smallest and the largest number in a dataset are.

3.6. Boxplots and the Interquartile Range

Now that you have a better understanding of the median, let's return to the boxplot, which was briefly introduced but not explained in Chapter 2. Box-and-whisker plots are very common in many areas of science, and you will find them in many linguistic papers as well. To understand the meaning of a boxplot, you need to learn about the 'interquartile range'.

Figure 3.4 shows the distribution of emotional valence scores for the 14,000 words from the Warriner et al. (2013) rating study. Recall that the sample mean of this distribution is $M = 5.06$. The median is 5.2 and is shown as the thick black line in the middle of the box of the boxplot. The extent of the box covers 50% of the data, that is, 25% of the data above and below the median. The ends of the box have specific names: they are the first, second, and third 'quartile'. You are probably more familiar with percentiles. For example, if someone received a test score that is in the 80th percentile, 80% of the test scores are below that person's score. Q1 is the first quartile, which is the 25th percentile. The next quartile is Q2, the median. Finally, 75% of the data fall below Q3, the third quartile.

For the Warriner et al. (2013) dataset, Q1 is the number 4.25 (25% of all data points fall below this value), and Q3 is 5.95 (75% fall below this value). The difference between Q1 and Q3 yields the 'interquartile range', in this case, 5.95−4.25 = 1.7. This number corresponds to the length of the box seen in Figure 3.4.

What's the meaning of the whiskers extending from the box? Let's focus on the right whisker, the one that is extending from Q3 towards larger emotional valence scores. This whisker ends at the largest number that falls within a distance of 1.5 times the

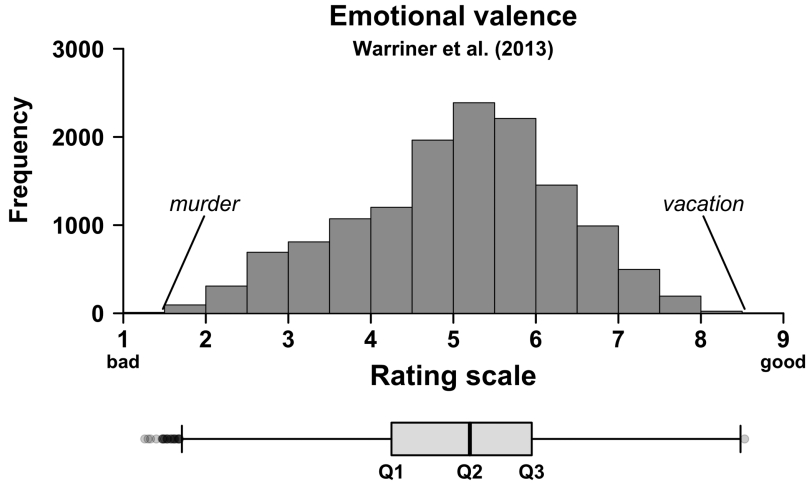

Figure 3.4. A histogram of the emotional valence rating data

interquartile range from Q3. You can think of standing at Q3 and swinging a lasso as wide as 1.5 times the interquartile range. The largest number that you catch—the score for the word *happiness* in this case—is the extent of the whisker. The logic is the same for the lower whisker, except that you are looking for the smallest number that falls within a distance of 1.5 * *IQR*.

The data points that fall outside of the whiskers are indicated by dots. For example, the single dot to the right of the right whisker is the word *vacation*. Words that fall outside the range of the whiskers are often called outliers, but I prefer the term 'extreme value', since 'outlier' suggests that something is qualitatively different from the other data points, which is often used to justify exclusions.[5] Using the term 'extreme value' implies that the same underlying process has generated the extremity.

If 'maximum of $Q3 + 1.5 * IQR$' and 'minimum of $Q1 - 1.5 * IQR$' seems like a horribly non-intuitive way of defining the whiskers to you, perhaps it's best to avoid boxplots. And, if you do decide to use a boxplot, don't hesitate to re-state the definition of the whiskers in the figure captions—it's good to give people reminders.[6] Describing the meaning of the individual plot components in the figure captions should be done anyway. For example, sometimes researchers use the range of the data (minimum and maximum) for the whiskers of a boxplot, which you would need to know in order to interpret the plot correctly. In general, when writing up statistical results, it's good if you describe each plot in as much detail as possible.

3.7. Summary Statistics in R

Let's put our understanding of distributions and summary statistics into practice. First, create 50 uniformly distributed numbers with the `runif()` function. The name of this function stands for 'random uniform'. Since this is a random number generation function, your numbers will be different from the ones shown in this book.

```
# Generate 50 random uniformly distributed numbers:

x <- runif(50)

# Check:

x
```

```
 [1] 0.77436849 0.19722419 0.97801384 0.20132735
 [5] 0.36124443 0.74261194 0.97872844 0.49811371
 [9] 0.01331584 0.25994613 0.77589308 0.01637905
[13] 0.09574478 0.14216354 0.21112624 0.81125644
[17] 0.03654720 0.89163741 0.48323641 0.46666453
```

5 Data should never be excluded unless there are justifiable reasons for doing so.

6 I suspect that many people in the language sciences may not be able to state the definition of the whiskers off the top of their heads. If you want to make new friends at an academic poster session, next time you spot a boxplot, ask the poster presenter to define the whiskers.

```
[21] 0.98422408 0.60134555 0.03834435 0.14149569
[25] 0.80638553 0.26668568 0.04270205 0.61217452
[29] 0.55334840 0.85350077 0.46977854 0.39761656
[33] 0.80463673 0.50889739 0.63491535 0.49425172
[37] 0.28013090 0.90871035 0.78411616 0.55899702
[41] 0.24443749 0.53097066 0.11839594 0.98338343
[45] 0.89775284 0.73857376 0.37731070 0.60616883
[49] 0.51219426 0.98924666
```

Notice that by default, the `runif()` function generates continuous random numbers within the interval 0 to 1. You can override this by specifying the optional arguments `min` and `max`.

```
x <- runif(50, min = 2, max = 6)

head(x)
```

```
[1] 2.276534 2.338483 2.519782 4.984528 2.155517
[6] 4.742542
```

Plot a histogram of these numbers using the `hist()` function. A possible result is shown in Figure 3.5 (left plot). Remember that your plot will be different, and that is OK.

```
hist(x, col = 'steelblue')
```

Next, generate some random normally distributed data using the `rnorm()` function and draw a histogram of it.

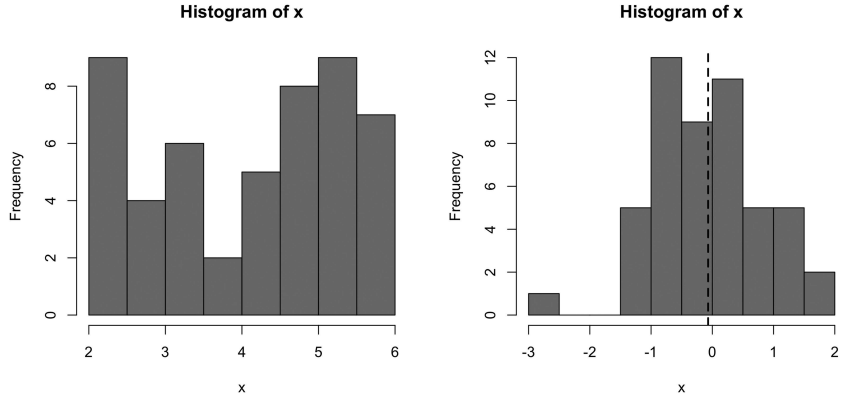

Figure 3.5. Left: random data drawn from aa uniform distribution; right: random data drawn from a normal distribution; the dashed line indicates the mean

```
x <- rnorm(50)

hist(x, col = 'steelblue')

abline(v = mean(x), lty = 2, lwd = 2)
```

This code also plots a vertical line at the mean. The line type is indicated to be dashed (`lty = 2`), and the line width is indicated to be 2 (`lwd = 2`). Figure 3.5 (right plot) is an example distribution.

The `rnorm()` function generates data with a mean of 0 by default. It also has a default for the standard deviation, which is 1 . You can override these defaults by specifying the optional arguments `mean` and `sd`.

```
x <- rnorm(50, mean = 5, sd = 2)
```

Check whether this has worked by computing the mean and the standard deviation using the corresponding `mean()` and `sd()` functions (remember: your values will be different).

```
mean(x)
```

```
[1] 4.853214
```

```
sd(x)
```

```
[1] 2.262328
```

Notice that these values are close to what was specified in the `rnorm()` command (`mean = 5, sd = 2`).

The `quantile()` function is useful for computing percentiles. If you run `quantile()` on the vector of random numbers without supplying any additional arguments, you retrieve the minimum (0th percentile) and the maximum (100th percentile), as well as Q1 (the first 'quartile', the 25th percentile), the median (Q2, the 50th percentile) and Q3 (the 75th percentile).

```
quantile(x)
```

```
        0%       25%       50%       75%      100%
-1.297574 3.100322 4.633111 6.363569 10.157849
```

You can use the `quantile()` function to assess the 68%-95% rule. The 68% interval corresponds to the 16th and 84th percentiles.

```
quantile(x, 0.16)
```

```
     16%
2.749623
```

```
quantile(x, 0.84)
```

```
     84%
7.080644
```

If the 68% rule of thumb works, the resulting numbers should be fairly close to the interval covered by $M - SD$ and $M + SD$.

```
mean(x) - sd(x)
```

```
[1] 2.590886
```

```
mean(x) + sd(x)
```

```
[1] 7.115541
```

And, indeed, the numbers are fairly similar to the percentiles. Let's do the same for the 95% interval, which corresponds to the interval between the 2.5th and the 97.5th percentiles. For this small example dataset, the 95% rule is a little off.

```
# 2.5th percentile:

quantile(x, 0.025)
```

```
    2.5%
1.004807
```

```
# Should correspond to M - 2 * SD:

mean(x) - 2 * sd(x)
```

```
[1] 0.3285584
```

```
# 97.5th percentile:

quantile(x, 0.975)
```

```
   97.5%
8.758132
```

```
# Should correspond to M + 2 * SD:

mean(x) + 2 * sd(x)
```

```
[1] 9.377869
```

I highly recommend using random number generation functions to develop an intuition for how approximate Gaussian data looks like. In some circumstances, a histogram

may look quite non-Gaussian even though an underlying Gaussian distribution was used to generate the data. To get a 'feel' for drawing random samples from the normal distribution, execute the following command repeatedly.

```
hist(rnorm(n = 20))  # execute repeatedly
```

You will notice that with only 20 data points it is often quite difficult to see that the data was drawn from an underlying normal distribution. If you change `n` to a very large number, the histograms should be much more Gaussian in shape. The take-home message here is that it is often difficult to tell what the appropriate reference distribution is for very small sample sizes.

3.8. Exploring the Emotional Valence Ratings

In this section, you will analyze the above-mentioned emotional valence ratings from Warriner et al. (2013). Let's load the `tidyverse` package and the data into your current R session (you will have to make sure that your working directory is set appropriately; see Chapter 1).

```
# Load packages and data:

library(tidyverse)

war <- read_csv(' warriner_2013_emotional_valence.csv')
```

As was emphasized again and again in Chapters 1 and 2, whenever you load data into R, you should spend considerable time familiarizing yourself with its structure.[7]

```
war
```

```
# A tibble: 13,915 x 2
   Word         Val
   <chr>      <dbl>
 1 aardvark    6.26
 2 abalone     5.3
 3 abandon     2.84
 4 abandonment 2.63
 5 abbey       5.85
 6 abdomen     5.43
 7 abdominal   4.48
 8 abduct      2.42
 9 abduction   2.05
10 abide       5.52
# ... with 13,905 more rows
```

7 It's worth remembering the principle 'garbage in, garbage out': if there's some issue with the data that you are unaware of, any stats computed may be worthless.

The tibble has two columns, pairing a `Word` with an emotional valence rating, `Val`. Compute the range to get a feel for this measure:

```
range(war$Val)
```

```
[1] 1.26 8.53
```

The emotional valence scores range from 1.26 to 8.53. To find the corresponding words, use `filter()`. Remember that this function *filters* rows based on a logical condition (see Chapter 2).

```
filter(war, Val == min(Val) | Val == max(Val))
```

```
# A tibble: 2 x 2
  Word       Val
  <chr>    <dbl>
1 pedophile 1.26
2 vacation  8.53
```

The above command uses the logical function 'or' (represented by the vertical bar '`|`') to retrieve all the rows that satisfy either of the two logical statements. This command can be translated into the following English sentence: 'Filter those rows from the `war` tibble for which valence is equal to the minimum or the maximum (both are OK).' The following command achieves the same result in a more compressed fashion (see Chapter 2.6 for an explanation of `%in%`).

```
filter(war, Val %in% range(Val))
```

```
# A tibble: 2 x 2
  Word        Val
  <chr>     <dbl>
1 pedophile  1.26
2 vacation   8.53
```

Let's have a look at the most positive and the most negative words in the dataset by using `arrange()`.

```
arrange(war, Val)  # ascending order
```

```
# A tibble: 13,915 x 2
   Word         Val
   <chr>      <dbl>
 1 pedophile   1.26
 2 rapist      1.30
 3 AIDS        1.33
 4 torture     1.40
 5 leukemia    1.47
 6 molester    1.48
 7 murder      1.48
```

```
 8 racism      1.48
 9 chemo       1.50
10 homicide    1.50
# ... with 13,905 more rows
```

```
arrange(war, desc(Val))  # descending order
```

```
# A tibble: 13,915 x 2
   Word        Val
   <chr>     <dbl>
 1 vacation   8.53
 2 happiness  8.48
 3 happy      8.47
 4 christmas  8.37
 5 enjoyment  8.37
 6 fun        8.37
 7 fantastic  8.36
 8 lovable    8.26
 9 free       8.25
10 hug        8.23
# ... with 13,905 more rows
```

Thanks to surveying lots of different data points, you now have a firm grasp of the nature of this data. Let's compute the mean and standard deviation.

```
mean(war$Val)
```

```
[1] 5.063847
```

```
sd(war$Val)
```

```
[1] 1.274892
```

The mean is 5.06; the standard deviation is 1.27. You expect 68% of the data to follow into the following interval:

```
mean(war$Val) - sd(war$Val)
```

```
[1] 3.788955
```

```
mean(war$Val) + sd(war$Val)
```

```
[1] 6.338738
```

Verify whether this is actually the case using the `quantile()` function. The fact that the resulting numbers are close to $M - SD$ and $M + SD$ shows that the rule was pretty accurate in this case.

```
quantile(war$Val, c(0.16, 0.84))
```

```
   16% 84%
3.67 6.32
```

Finally, let's have a look at the median, which is very similar to the mean in this case.

```
median(war$Val)
```

```
[1] 5.2
```

```
quantile(war$Val, 0.5)
```

```
50%
5.2
```

3.9. Chapter Conclusions

Everything in statistics is grounded in the notion of a distribution, and in (parametric) statistical modeling our goal is to make models of distributions. The mean is a great summary of a distribution, especially if the distribution approximates normality. In the applied R exercise, you then generated some random data and computed summary statistics. Being able to generate random data is a very important skill that will be nurtured throughout this book, alongside working with real data. Finally, you computed summary statistics for the Warriner et al. (2013) emotional valence ratings.

Everything up to this point has dealt with 'univariate' distributions. That is, you always only considered one set of numbers at a time. The next chapter will progress to bivariate data structures, focusing on the relationship between two sets of data.

3.10. Exercises

3.10.1. Exercise 1: Plotting a Histogram of the Emotional Valence Ratings

With the Warriner et al. (2013) data, create a `ggplot2` histogram and plot the mean as a vertical line into the plot using `geom_vline()` and the `xintercept` aesthetic (see Chapter 2). Can you additionally add vertical dashed lines to indicate where 68% and 95% of the data lie? (Ignore any warning messages about binwidth that may arise).

3.10.2. Exercise 2: Plotting Density Graphs

In the plot you created in the last exercise, exchange `geom_histogram()` with `geom_density()`, which produces a kernel density graph. This is a plot that won't be covered in this book, but by looking at it you may be able to figure out that it is essentially a smoothed version of a histogram. There are many other geoms to explore. Check out the vast ecosystem of online tutorials for different types of `ggplot2` functions.

Additional exercise: set the `fill` argument of `geom_density()` to a different color (such as 'peachpuff'). This is not an aesthetic mapping, because it doesn't draw from the data.

3.10.3. Exercise 3: Using the 68%-95% to Interpret Research Papers

Imagine reading a research paper about a grammaticality rating study. It is noted that the mean acceptability rating for a particular grammatical construction is 5.25 with a standard deviation of 0.4. Assuming normality, what is the interval within which you expect 68% of the data to lie? What about 95% of the data? Do you think the assumption of approximate normality is reasonable in this case?