# 2 Probabilities, likelihood, and inference

In the previous chapter we introduced some experimental data and talked about variables and experiments. In this chapter we’re going to talk about probabilities and explain how they can be used to make inferences about our data and research questions. Before beginning this chapter we should note that it’s normal if a some of the topics to be discussed don’t make sense the first time you read this chapter. It will make more sense once you start to actually build models and it becomes less hypothetical and more practical. In addition, as noted in the preface, it is a mistake to think that you can read a chapter in a statistics textbook once and move on having fully absorbed the content. If many of the topics in this chapter are new to you, you should probably: 1) Read this chapter, 2) wait a few days and read it again, and then 3) wait a few more days (or weeks) and read it again. It may also be useful to return to this chapter even once you are working through the following chapters. In our experience, you may find that you *see* information for the first time that was there all along, but that finally makes sense given your increased experience with the subject.

## 2.1 Data and research questions

We’re going to think about a hypothetical value “the average apparent height of adults males in the United States (US)”. To put it another way, we might wonder “how tall do men from the US ‘sound’?”. The apparent heights of men from the US cannot be known **a priori**. In other words, you don’t know how tall a random man from the US will ‘sound’ until you actually observe the judgment. For this reason, “the average apparent height of men from the US” is a random variable. We’re going to think about how we can use our experimental data to try to answer the following two questions:

Q1) How tall does the average man from the US sound?

Q2) Can we set limits on credible average apparent heights based on the data we collected?

These two questions can be though of as relating the the central location and the spread of the data, respectively. Answering the first question tells you what values your variable tends to take on, while answering the second question tells you how much variation you can expect around the most typical values. Scientific research is often focused on questions such as (1) regarding the central location, the average value, of some variable. For example someone might ask “how tall do adult male speakers from the US sound?” and you can say, for example, “I have some data that suggests 174 cm is a reasonable estimate”. However, reliable inference requires answering question (2) as well, and determining what range of values are believable for a certain variable.

Think of the average height of the people in a large city. You can go out and sample 100 individual people, and each one of those samples is an observation from a random variable. You can find the mean of your sample, arriving at a single estimate of the population mean. Now imagine that 50 people went out in the same city and each sampled 100 random people. There is no chance that every single of those 50 people would find identical means across all of their samples. Instead, there will be a distribution of sample means, in the same way there is a distribution of the original data used to calculate the means. Another way to look at this is that there is some degree of **uncertainty** involved when answering any research question.

As a result of this uncertainty, it can be difficult to rule out alternative possible answers to our research questions. For example, if 174 cm is a good estimate, what about 173.99 cm? What about 173 cm? 172 cm? Where do estimates of average height stop being ‘good’? Without being able to say what is *not* a good estimate, it is not quite as useful to be able to say what *is*. A related issue arises with respect to the interpretation of average values. Imagine that you read about a miracle diet that was guaranteed to make you lose one gram of weight a day. You know that is not very impressive. How? Because you understand that a difference of one gram is not large relative to the variation that exists in the weight of a human body on a daily basis. You could gain the weight back (and more) by drinking a teaspoon of water. Without knowing how much human weights tend to vary between and within people, it’s impossible to know whether a reduction of one gram constitutes a meaningful change in the mass of a human. In contrast, a diet that causes one gram of weight loss in the average hamster may actually be of interest to hamster owners, as this is a relatively large value relative to natural variation in hamster weight.

So, we see that imposing limits on credible ranges for our average values can be as important as finding the average values themselves. Further, in order to properly contextualize values and effects, we need to have some idea about the underlying variation in the measurements. Clearly, we need some principled way to ‘guess’ reasonable ranges based on our sample of observations, in addition to just talking about average values. In this chapter we will discuss how statistics provides us with a framework to answer both questions above using only our sample of values.

In order to discuss the apparent heights of adult males, we need to subset these observations from our data. Below we load the book package (bmmb) and select rows from our experimental where a speaker was judged to be an adult male speaker (i.e. a man). Recall that our experiment contained an acoustic manipulation such that speech resonances were changed to make speakers sound bigger (see section 1.3.2). For now, we’re going to focus only for the unmodified productions, the ‘natural’ speech produced by the men in our sample.

We can have a look at some of the quantiles (see section ?? to get an idea of what range of values this variable tends to have. We see that the minimum and maximum values are 139.7 and 192.3 cm, that 174.5 cm is the median, and that half of the observed height judgments for adult males fell between 169.2 and 179.1 cm.

quantile (mens\_height)

## 0% 25% 50% 75% 100%   
## 139.7 169.2 174.5 179.1 192.3

Obviously, an inspection of the distribution of our observed height judgments only gives us direct information about the judgments we *did* observe. To make inferences about the probable characteristics of the height judgments we did *not* observe, or to talk about height judgments for adult males from the US more generally, we rely on methods of statistical inference, as will be described below.

## 2.2 Empirical Probabilities

The **sample space** of a variable is the set of all possible outcomes/values that a variable can take. Slassic examples are a coin flip, which can take on the values ‘heads’ or ‘tails’, or the roll of a die which can take on the values one through six. In other cases the sample space may have an infinite or practically infinite number of members. For example, since time is continuous there are an infinite number of durations an event may have, given adequate precision in measuring time. If we think of the human population, or the population of fish in the sea, these are theoretically finite but practically infinite. It would be extremely difficult to fully sample either of these populations, and impossible to do so before they changed substantially (i.e. before some members have died and others have been born).

The **probability** of an event/outcome is the number of times an outcome is expected to occur, compared to all the other possible outcomes that can occur (i.e. the other outcomes in the sample space). By convention, the probability of each event is assigned a value between 0 and 1 and the total probability of all of the possible outcomes in the sample space is equal to one. As a result of this convention, you know that a probability of 0.5 means something is expected to occur half the time (i.e. on 50% of trials), and a probability of 0.25 indicates that something should happen 25% of the time, about one in every four trials. **Empirical probabilities** are the probabilities of different outcomes in a sample of data. For example, we can flip a coin 100 times and observe 65 heads. This means that the empirical probability of observing heads in our data is 0.65 (65% of trials).

Suppose we want to know the probability of being an adult male in our sample who is identified as being taller than 180 cm. To calculate the empirical probability of this occurring in our data we need to find 1) The number of times it occurred, and 2) the total number of observations for the variable. We can do this easily using the logical operators and variables discussed in Chapter 1. Below, we find the total number of outcomes that satisfy our restrictions (being under 180 cm), and divide this by the total number of observations being considered.

# the evaluation in the parenthesis will return 1 if true, 0 if false  
# number of observations the fall below threshold  
sum (mens\_height > 180)   
## [1] 136  
  
# divided by total number of events  
sum (mens\_height > 180) / length (mens\_height)   
## [1] 0.2014815  
  
# a shortcut to calculate probability, mean = sum/length  
mean (mens\_height > 180)  
## [1] 0.2014815

The top value is the frequency of the occurrence. This is not so useful because this number can mean very different things given different sample sizes (e.g., 136/675, 136/675000). In contrast, the middle and bottom values have been divided by the total number of observations. As a result, these now represent the probability of occurrence in a way that is independent of the total number of observations.

### 2.2.1 Conditional and marginal probabilities

Figure 2.1 presents boxplots of the overall distribution of height judgments in our data (left), and of the distribution of height responses provided by each listener individually. We can see that height judgments range from about 140 to 200 cm, with most responses falling between 170 and 180 cm. Notice that our overall boxplot does not give us any information about the ranges used by different listeners, nor even the fact that the data was contributed by different listeners. This overall distribution of height responses is the **marginal distribution** of height judgments. The marginal distribution of a variable is the overall distribution, *across* all values of all other variables. The marginal boxplot on the left compresses all of the listener-specific boxplots on the right into one single box. It’s common to denote marginal probabilities using notation like this , meaning we might refer to the marginal probabilities of height responses like this .

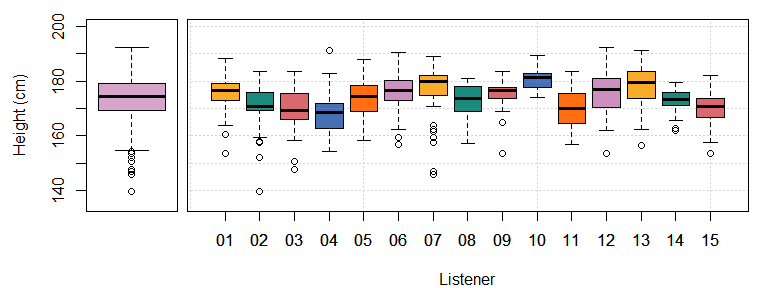


Figure 2.1: (left) Boxplot showing all height judgements for adult male speakers in our experiment. (right) Individual boxplots for each listener’s responses

A look at figure 2.1 reveals that the probability of observing a height response of over 180 cm can vary substantially on the listener that provided it (e.g. compare listener 10 vs listener 15). Recall that height is a quantitative variable and listener is a categorical predictor (a factor) with 15 levels, one for each listener (see section 1.4.3). We can talk about how height judgments vary across the levels of our listener factor (i.e. for different listeners) by considering the **conditional probability** of height given listener. A conditional probability is the probability of an outcome given that some other outcome has occurred. For example, rather than ask “what is the probability of observing an apparent height over 180 cm?”, we can ask “what is the probability of observing an apparent height over 180 cm *given* that we are observing data from listener 10?”. Conditional probabilities basically reduce the sample space by including only the subset of events that satisfy the given condition.

Conditional probabilities are often denoted like this , which in this case would look like . For example, the first box in the right panel of figure 2.1 is the distribution of , and the second box is the distribution of . Below we divide our vector of height judgments (mens\_height) into those contributed by listener 10 and listener 15. We then find the probability of observing a height judgment over 180 cm conditional on listener, and see that these can differ quite a bit from each other.

# create subsets based on listener  
L10 = mens\_height[men$L==10]  
L15 = mens\_height[men$L==15]  
  
# find the conditional probability of height>180 for each listener  
mean (L10 > 180)  
## [1] 0.6444444  
mean (L15 > 180)  
## [1] 0.02222222

In the boxplots in figure 2.1, we see that the distributions of heights vary substantially as a function of the value of L, our listener variable. Contrast this with the boxplots seen in figure 2.2, which shows the distribution of stimulus durations conditional on listener. Since every individual recording had a fixed duration and all listeners heard the same sounds, we know that the distribution of stimulus durations is identical across all levels of the listener variable. As a result, we can see that all of the conditional distributions of duration given listener look just like each other, and just like the marginal probability. This tells us that duration and listener are *statistically independent*. When two variables are statistically independent, the distribution of one variable is not affected by the values of the other. As a result, the conditional distribution of one variable given the other will be the same as its marginal distribution, as seen in figure 2.1. This can be stated for the general case as . In our example above, is *not* equal to , and so we conclude that the variables apparent height and listener are not independent. Instead, they are **statistically dependent**, meaning that these variables *do* affect each other in some way, and that knowing the value of one may tell you something about probable values of the other.

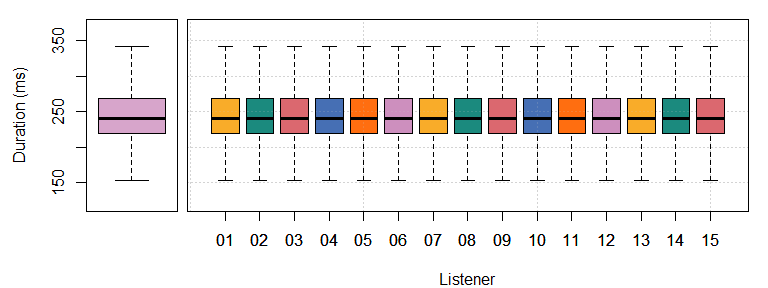


Figure 2.2: (left) Boxplot showing the fundamental frequency (f0) of adult male speakers in our experiment. (right) Individual boxplots for the stimuli presented to each listener.

### 2.2.2 Joint probabilities

**Joint probabilities** reflect the probabilities of two or more things occurring together. We can refer to the joint probability of and using the notation or . Here are some important things to know about joint probabilities:

1. The formula for calculating the joint probability of two outcomes and is given by . In plain English this means that the probability of and is equal to the conditional probability of given , multiplied by the marginal probability of .
2. Recall from section @ref{c2-conditional} that when and are independent . In other words when and are independent the conditional probability of given equals the marginal (unconditional) probability of . As a result of this, when and are independent . Thus, when and are independent their joint probability can be found by simply multiplying their individual marginal probabilities.
3. The probability of and is equal to the probability of and . As a result of this , and as a result of that . This relation will become useful later.

We can demonstrate the above properties using the empirical probabilities in our data. Consider the joint probability of observing a response in our data that was contributed by listener ten (L=='10'), and also being longer than 250 ms in duration. We can find this by joining two logical variables using the & (and) symbol as shown below.

# TRUE is the listener is 01  
L10 = men$L=='10'  
# TRUE if the Resonance is 'a'ctual  
dur\_250 = men$dur > 250  
  
# The probability of A and B  
mean (L10 & dur\_250)  
## [1] 0.02814815  
mean (men$L=='10' & men$dur > 250)  
## [1] 0.02814815

So, we see that the probability of observing this event is 0.28, indicating that we expect this in about 3% of trials. Below, we see that this same joint probability can be calculated based on .

# Marginal probability of observing listener 10 (i.e. P(L=10))  
p\_L10 = mean (men$L=='10')  
  
# Subset containing only listener 10   
L10 = men[men$L == '10',]  
# Conditional probability of dur>250 given listener 10 (i.e., P(dur>250 | L=10))  
p\_dur\_250\_given\_L10 = mean (L10$dur > 250)  
  
# Joint probability = P(dur>250 | L=10) \* P(L=10)  
p\_dur\_250\_given\_L10 \* p\_L10  
## [1] 0.02814815

Because of the experimental design, we know that every listener heard the same vowel durations (see Section 1.3 for a review of the experimental design). This means that duration is independent of listener: Knowing who the listener is in no way changes the fact that every listener is expected to heard the same number of trials with durations longer than 250 ms. As a result of this independence, we could also have calculated the joint probability of the above events by simply multiplying their marginal probabilities, as seen below.

mean (p\_L10) \* mean (dur\_250)  
## [1] 0.02814815

This short cut will not work for variables that are not independent. For example, let’s consider the probability of observing a height judgments of greater than 180 cm provided by listener 10. Below we calculate this the *wrong* way by multiplying the marginal probabilities, substantially underestimating the probability of the event. The problem is that this method of calculation does not take into account that listener 10 was among the most likely listeners to report heights greater than 180 cm.

# TRUE is the listener is 10  
L10 = men$L=='10'  
# TRUE if the height is over 180  
under\_180 = men$height > 180  
  
# Empirical probability of the observation  
mean(men$L=='10' & men$height > 180)  
## [1] 0.04296296  
  
# Wrong: multiplying marginal probabilities  
mean(L10) \* mean(under\_180)  
## [1] 0.0134321

Below we calculate the joint probability correctly by using the conditional probability.

# Marginal probability of observing listener 01  
p\_L10 = mean (men$L=='10')  
  
# Subset containing only listener 01 (i.e., given listener 01)  
L10 = men[men$L == '10',]  
# Conditional probability of a height < 160 given listener 01  
p\_over\_180\_given\_L10 = mean (L10$height > 180)  
  
# Joint probability  
p\_over\_180\_given\_L10 \* p\_L10  
## [1] 0.04296296

This highlights a very important point: The calculation of joint probabilities is much (much) simpler when you can assume that your observations are independent. For example, consider the comparison of the joint probability of four events, one dependent and one independent:

$$$$

This means that we not only need to calculate many conditional probabilities, but most of these feature the conditioning of one variables on several other variables (i.e., ). Although this may not matter when calculating the joint probability of a handful of observations, calculating the joint probability of hundreds or thousands of correlated variables can become difficult if not impossible in practice.

## 2.3 Probability distributions

A probability distribution is a function that assigns probabilities to all the different outcomes in your sample space. We will discuss exactly what this means with reference to histograms. In the left panel of figure 2.3 we see a histogram of counts. This histogram shows the number of times a range of values was observed along the y axis. The heights of the bars reflect the frequencies of different sort of observations and so can tell you about the values you expect, and don’t expect, for the variable. This sort of representation makes it difficult to compare distributions across samples of different sizes. For example, if the number of observations were tripled, so would the heights of all the bars in the histogram.

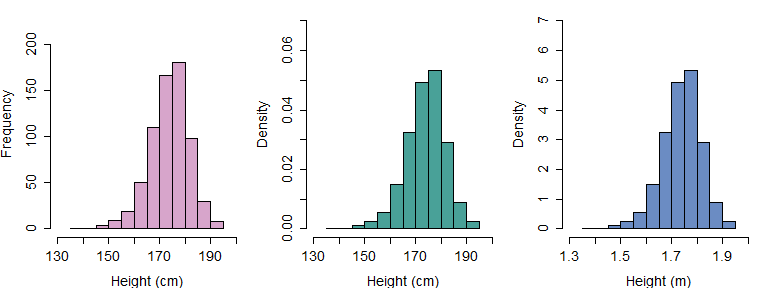


Figure 2.3: (left) A histogram of adult male height judgments showing counts in each bin. (middle) The same data from the left panel, this time showing the density of the distribution. (right) The same data from the middle plot, this time with heights expressed in meters.

Histograms can also be used to show the **probability densities** associated with different values. Imagine a circle, like a Venn diagram, that contains all possible values of your variable (i.e. the sample space). This circle has an area of one since it contains all possible values in the population. Imagine we took this circle and pressed it down against the number line, squishing it. We then spread out this circle along the x axis so that its shape reflected the relative frequencies of different values present in the population. For example, if some outcomes were five times more probable than others, the shape should be five times thicker there, and so on. If we managed to do this, the thickness (or **density**) of this shape would exactly correspond to the probability density of the variable.

In the middle in figure 2.3 we see a histogram that shows the *density* of the distribution of height judgments, i.e. the thickness of the distribution at different locations along the number line. The values of the density are constrained by the fact that the area under the curve must equal one, since these reflect probabilities. In the case of our histograms this means that the total area of all the bars (rectangles) is constrained to equal one. As a result of this, when variables have wide ranges densities tend to be very small (as in the middle panel in Figure 2.3).

In the right panel of figure 2.3 we present heights in meters rather than centimeters. This has the effect of substantially increasing the values of our density (one-hundred fold) but does not otherwise affect the shape of the curve. The relationship between variable ranges and density values means that densities can’t really be interpreted in an absolute sense. A density value of 0.06 means very different things in the histograms in Figure 2.3. Instead, density values need to be interpreted relative to the other values along the same curve. A higher density in a certain location tells us values in that vicinity are more probable than values in locations with lower densities, and differences in densities reflect differences in the relative probability of different values or outcomes.

Probability distributions sometimes have shapes that can be represented using mathematical functions. These functions have a limited number of **parameters** that determine their exact shape. Think of parameters like ways that things can be ‘set’ differently from each other. For example, a radio has three parameters: tuner frequency, band (AM/FM), and volume. A toaster may have only one, a single knob determining the degree of toasting required. The more parameters something has, the more complicated it is (an airplane has thousands). For example, consider the code below, which defines a slope and intercept and draws a line based on these parameters. You can change the values of the intercept and the slope and draw many kinds of lines. However, there no way to change the characteristics of a line other than by changing its slope and intercept: A line is entirely defined by its slope and intercept.

intercept = 3  
slope = 1  
x = seq (-10,10,.1)  
y = intercept + x\*slope  
  
plot (x,y, type='l',lwd=3,col=4)

Sometimes the same sorts of probability density shapes pop up over and over, and these shapes are often well-defined mathematically. **Parametric probability distributions** are those density shapes that can be understood in terms of curves that vary in terms of a limited number of parameters. Just like with lines, the characteristics of a parametric probability distribution are entirely defined by the values of its parameters.  
## The normal distribution {#c2-normal}

The distribution of many random variables (at least approximately) follows what’s known as the **normal** or **Gaussian** distribution. This means that if you take a sample of a random variable and arrange observations into bins, they resulting histogram will tend to have the familiar, bell-shaped curve common to normally-distributed data (seen in the histograms in figure 2.3). The normal distribution has the following important characteristics:

1. The distribution is symmetrical - i.e., producing a higher or lower than average f0 is equally likely.
2. The probability of observing a given value decreases as you get further from the mean (i.e., *average*) value.
3. It’s easy to work with, very well understood, and naturally arises in many domains.

Normal distributions have two parameters, meaning they vary from each other in only two ways. The parameters are:

1. A mean parameter () which determines the location of the distribution along the x axis. When the mean changes, the whole shape of the distribution ‘slides’ along the number line. The mean is the 50% halfway point of the ‘mass’ of the distribution. If the distribution were an physical object, its mean would be its center of gravity and you would balance the distribution on your fingertip at this point.
2. A variance () or standard deviation () that determines its *spread* along the x axis. When the standard deviation changes the distribution is stretched wide or made very narrow, but stays in place. Since every distribution has an area under the curve equal to one (i.e., they all have the same ‘volume’), distributions with small variances must necessarily be very dense.

In principle, a given probability distribution can be thought of as having fixed parameters. For example, we might imagine that the average apparent height of adult male speakers from the US is exactly 175.91254… cm. However, in most situations we can’t be certain of what the ‘true’ parameters of a distribution are. Instead, we must be satisfied with estimating the values of parameters based on our samples. These are **statistics**, estimates of our population parameters based on our sample. **Statistical inference** consists of using statistics (based on our sample) to make inferences about the characteristics of the overall population (i.e. the ‘true’ parameters). In the case of the normal distribution we are interested in two statistics: The sample mean and the sample standard variance.

### 2.3.1 The sample mean

The sample mean is our ‘best’ guess for the population mean. We’ll be more specific about what this means later, but for now we can consider that if you don’t know the population parameter for a given distribution, the sample mean will provide a reliable estimate. The formula for the sample mean is given in equation (2.1). Initially, reading these mathematical formulas may seem daunting. However, learning to read these is just a skill that is developed with practice. In addition, you will begin to see the same ‘chunks’ or structures come up in formulas over and over, and reading these becomes much easier once you start to recognize the meaning of these repetitive structures intuitively.

Equation (2.1) says that the sample mean of () is equal to the sum of all of the elements of the vector , divided by where is equal to the length of the vector. We use the little hat symbol () to indicate that this is an estimate of the mean, and to distinguish it from the population mean which goes hatless (). The summation () symbol represents the repetitive adding of whatever is the the right of the symbol to some total. The summation beings at the number below the and performs one operation for every integer value of between the starting point and the end point (indicated above the ). The counter variable, in this case , is also often used to index values of a vector (or other structure) that is being summed (as seen below).

This behavior analogous to the behavior of a for loop in R. Below we define the variables N (equal to the length of the vector) and initialize a variable to receive the summation (mean\_height). The for loop then proceeds to increment a count variable (i) from one to N along the integers. For each iteration it adds the value of the vector, divided by N, to the summation variable mean\_height. As we can see below, this results in a value identical to that returned by the mean function in R.

# initialize values  
N = length (mens\_height)  
mean\_height = 0  
  
# summation equivalent to equation 2.1 above  
for (i in 1:N) mean\_height = mean\_height + mens\_height[i]/N  
  
# replicates values of the mean function  
mean (mens\_height)  
## [1] 173.7877  
mean\_height  
## [1] 173.7877

Here are some useful things to know about sample means, in no particular order:

1. The mean of a set of observations is affected by addition and multiplication. This means that adding to a set of observations increases its mean by , and multiplying observations by results in an change in the mean by the same factor.
2. The mean of the sum of two sets of variables (of the same length) an is equal to . In other words, the average of the sum is just the sum of the averages.
3. The sum of sample’s deviations from the sample mean equals zero (seen in (2.2)). This means that the sum of the distances between positive and negative differences from the sample mean exactly balance out. To some extent this makes sense since the mean is the ‘center of gravity’ of a distribution. It is worth noting that this does not apply to deviations form the *population* mean since this is not specifically fit to the characteristics of the sample.

### 2.3.2 The sample variance (or standard deviation)

The formula to calculate the sample variance is seen in (2.3). Note that it is quite similar to the structure of (2.1) and clearly involves the averaging of a value. In fact, if we were to replace with the two equations would be identical. The value being averaged () consists of finding a difference () followed by a squaring operation. So, we see that what’s being averaged is squared deviations from the sample mean. This is what the variance is: The expected value of squared deviations around the mean of the variable.

Below we see that we can use a for loop to recreate Equation (2.3). However, we don’t manage to exactly recreate the output of the var (variance) function included in R. The reason for this is that the formula above is our best guess for the *sample* variance, but not for the *population* variance. This is because Equation (2.3) uses the sample mean to calculate variation.

# initialize variable  
variance\_height = 0  
  
# equivalent to 2.3 above  
for (i in 1:N) variance\_height = variance\_height + (mens\_height[i]-mean\_height)^2/N  
  
# this time the values don't match  
var (mens\_height)  
## [1] 60.27316  
variance\_height  
## [1] 60.18386

Recall above that we said that the sample mean is our ‘best’ estimate of the population mean given a sample. A more formal way to state this is that the sample mean is the value which minimizes the sample variance. In other words, if we choose any value of to calculate the sample variance other than the sample mean, the variance will necessarily be larger. However, we know that our sample mean is just an estimate of the population mean and will never be exactly equal to it. As a result of this, the true variance *must* be greater than the sample variance when calculated using the sample mean. We can put it like this: , the sum of squares around the sample mean will always be less than or equal to the sum of squares around the population mean. For reasons that we won’t get into (but which aren’t too complicated), this expected difference may be offset by dividing the squared deviations by rather than as in (2.4).

We can update our R code to reflect this change, and see that this now matches the calculation of the variance carried out by R.

# initialize variable  
variance\_height = 0  
  
# equivalent to 2.4 above  
for (i in 1:N) variance\_height = variance\_height + (mens\_height[i]-mean\_height)^2/(N-1)  
  
# this time the values do match  
var (mens\_height)  
## [1] 60.27316  
variance\_height  
## [1] 60.27316

The sample standard deviation () is simply the square root of the sample variance, as in (2.5)

Here are some useful things to know about variances, in no particular order:

1. Variances are always positive, and can only be zero for variables that do not actually take on different values (i.e., constants).
2. The variance of a set of observations is not affected by addition. So, adding or subtracting some arbitrary value from a data set will not affect the variances in that data.
3. Multiplication *does* affect the variances of a set of values. Multiplying numbers by results in a change of the variances equal to . So, if we took our heights and multiplied them by 10 to express them in milliliters, we would expect the value of to increase by a value of . Since standard deviations are the square roots of variances, this implies that multiplying data by results in an increase in the standard variation of the data by a factor of .
4. The variance of the sum of variables depends on whether they are independent or not. When variables are independent, the sum of their variances is simply equal to for variables x and y. However, when variables are not independent, the variance can be much greater or smaller than this based on the nature of the relationship between the variables. As a result, just as with the calculation of joint probabilities, we need to take into account whether variables are independent or not when we considering the variance of the sum of variables.

### 2.3.3 The normal density

The parameters of a probability distribution are used to draw its shape, which can be used to make inferences about likely values. Think back to high school math and the function defining the shape of a parabola . This function draws a shape based on the settings of its parameters and . The parameter determines the width of the parabola (and whether it points up or down), while the vertex of the parabola will have x and y axis coordinates of and respectively. In the same way, the formula defining the density of the normal distribution draws a shape given the settings of its and parameters. The formula for the probability density function of the normal distribution is seen in (2.5). The function returns a density value for the probability distribution as a function of the value of , and the values of its parameters.

The equation in (2.5) features **exponentiation** (), that is raising the base (Euler’s number) to some power as in . In (2.5), the value being exponentiated is . The **logarithm** (log) is the inverse function to exponentiation, it basically *erases* or *undoes* exponentiation. We can apply a logarithmic transformation to both sides of (2.5), resulting in the **log density** seen in equation (2.7). Before explaining equation (2.7), we will discuss some basic properties of logarithms that are useful to understand probabilities and probability distributions, as these often involve exponentiation and logarithms. The first line in (2.6) shows the basic behavior of logarithms. The next four lines pertain to the values expected, or undefined, for values of . The next two lines highlight the fact that exponentiation of numbers is equivalent to the multiplication of their logarithms. The final two lines highlight the fact that multiplication of two numbers is equivalent to the addition of the logarithms of the numbers.

Armed with knowledge of the behavior of logarithms, we can see that compared to equation (2.5), we have removed the function around the rightmost term , and added the function around all terms that were *not* previously exponentiated. In addition, the multiplication of the two terms in the right hand side of the equation has turned into addition, or subtraction in this case because we are adding a negative term.

We can use the properties of logarithms to turn into and move this term to the rightmost end of the equation as in (2.8). At this point the function is that of a parabola in vertex form, , where , , and . So, we can see that the normal distribution is just an exponentiated parabola that is scaled by so that the area under the curve is equal to one. The parabola has its vertex at , and opens downwards since the term is negative. When a parabola is in vertex form, the relationship between parabola width and the value of is inverted, its width decreases as the value of increases. However, because , this means that in the case of normal distributions the parabola width increases as grows larger. As a result, a larger standard deviation leads to wider parabolas and wider probability densities.

Equation (2.8) shows how the and parameters work to make observations further from the means less probable. First, we know that negative logarithmic values will fall between 0 and 1, with more negative values being closer to zero (i.e. less probable). As observations () are further from the mean, the value of will be greater so that values further from the mean will be generally less probable. However, whether a deviation is big or small is relative, and so this distance is scaled with respect to the average expected squared deviation from the mean (i.e. the variance ). Variation of 1 cm in body length means different things for an earthworm as opposed to an anaconda. As a result, large values of can be offset by large values of when determining the probability of an outcome.

The nice thing about working in R is that you don’t need to take our word for any of this, you can see it for yourself. In fact, we strongly encourage you to ‘test’ the things we say or check to see if your intuitions about things using toy examples in R. You can use the code below to confirm that the ‘shape’ of the normal distribution is just an exponentiated parabola:

par (mfrow = c(1,2), mar = c(4,4,1,1))  
# draw an parabola  
curve (-x^2, xlim = c(-3,3), lwd=2)  
# same thing but exponentiated  
curve (exp (-x^2), xlim = c(-3,3), lwd=2)

The code below will generate densities for normal distributions based on the equations in the text above:

# you can change the ranges, mean and standard deviation (sigma) to any value  
x = seq (-4,4,.01)  
mu = 0  
sigma = 1  
eq25 = 1/(sigma\*sqrt(2\*pi)) \* exp (-(1/(2\*sigma^2))\*(x-mu)^2)  
eq27 = log(1/(sigma\*sqrt(2\*pi))) + -(1/(2\*sigma^2))\*(x-mu)^2  
eq28 = -(1/(2\*sigma^2))\*(x-mu)^2 - log((sigma\*sqrt(2\*pi)))

And these plots can be used to confirm the relations outlined above. For example, the output of equations (2.7)) and (2.8)) should equal, and this can be confirmed below.

par (mfrow = c(1,2), mar = c(4,4,1,1))  
plot (x, eq25, cex=1.2, col=skyblue,pch=1)  
lines (x, exp(eq27), col = coral, lwd = 2)  
plot (x, eq27, cex=1.2, col=skyblue)  
lines (x, eq28, col = coral, lwd = 2)

### 2.3.4 The standard normal distribution

The **standard normal distribution** is a normal distribution with a mean of zero and a standard deviation of one. Variables drawn from a standard normal distribution are often represented by the symbol (sometimes called a **z score**). Any normally distributed variable can be turned into a standard normal variable by an operation known as **standardization**, which consists of **centering** and then **scaling** the variable as in (2.9). To center a variable we subtract the mean from the value of each observation, making the new mean equal to zero. By dividing our observations by the standard deviation, we scale these values so that the new standard deviation is equal to one (since anything divided by itself is equal to one).

Equation (2.10) re-arranges the terms in (2.9) to isolate on the left-hand side. From (2.10) we can see that any normally-distributed variable can be thought of as a standard normal that has been multiplied by a standard deviation and then had a mean added to this product.

Normally-distributed data is often discussed in terms of ‘standard deviations from the mean’. This is because stating things in terms of standard deviations from the mean effectively standardizes a variable, making all variables seem standard normal. For example, if someone says “my test score was two standard deviations above the mean” what do you know about their score? You don’t know what the mean is, nor what the standard deviation is. All you know is that their score is two distance units (standard deviations) above the mean, so their test score can be thought of as a standard normal variable like . This is despite the obvious fact that the true average test score was not zero and the true average standard deviation was not one. This is an extremely useful property because it means that we can discuss the probability of any given event from any given distribution in consistent terms. For example, we can say that an observation four standard deviations from the mean is very unusual in *any* normal distribution. This means that if you have an observation equal to 140, the mean is 174, and the standard deviation is 8, you know that this observation is very improbable. That is because it is 4.25 (34/8) standard deviations from the mean of 174. Figure (2.4) presents our height data again, but this time compares the data to its centered and scaled versions.

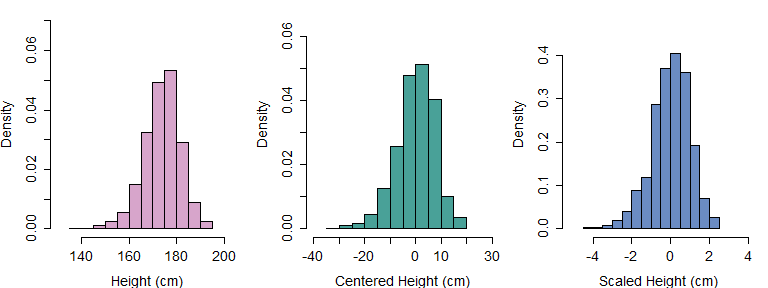


Figure 2.4: (left) A histogram of adult male height judgments in our data. (middle) The same data from the left panel, this time the data has been centered around the mean. (right) The same data from the middle plot, this time the data has been scaled according to the standard deviation.

## 2.4 Models and inference

Models are simplified representations that help us understand things. For example, we may want to understand the movement of balls on a billiards table, perhaps to create a video game about playing pool. To do this we may assume the balls are spherical and that their mass is evenly distributed among its volume. Neither of these things are exactly true but assuming this helps us keep our model simple and manageable. It also helps us build our model in terms of things, like regular spheres, whose properties and behaviors are well understood and are easy to work with. To build a real *exact* model we would need to include friction from the felt on the table, the effect of wind resistance, the gravitational effect of the moon, and a large number of other factors. As a result, a perfect or exact model is not really possible for most things (and maybe for anything). And yet, the simulation of realistic behavior of billiards balls is easy and can be done with great accuracy, suggesting that a simplified model can still be useful to understand the behavior of the more-complicated phenomenon it is meant to represent.

In general, it’s impossible to know what the ‘true’ data distribution is, so that *perfect* inference is not possible. As a result, scientists often use theoretical probability distributions to make inferences about real-life populations and observations. If our measurements more or less follow the ‘shape’ predicted by the theoretical normal distribution, we may be able to use the characteristics of an appropriate normal distribution to make inferences about our variables. Using a normal distribution to make inferences about your data is like using a mathematical model for spheres to understand the behavior of billiard balls. The billiards balls are *spherical enough* to allow us to make useful predictions based on the simplified model.

In general, it is useful to keep in mind that reality will never exactly conform to our model. This can result in unpredictable errors in our conclusions. In general, the things you don’t know you don’t know are the things that will cause the most problems. If you had known that your model was wrong, you would have fixed it! Further, using models to make inferences about the general properties of data assumes that the things you have not seen are more or less just like those you have. Under those conditions then the conclusions you draw may be reliable. It is important to keep this limitation in mind, because you never know for sure that what you have not seen will conform to your model, and as a result the fit between a model and some set of observations can never definitively *prove* the truth of the relations encoded in the model (this is the problem of induction, discussed in section 1.2).

Parametric distributions can be used to establish **theoretical probabilities**, that is expectations about which events are and are not likely based on the general shape expected for the distribution. Basically, if we expect our distribution of values to have the shape of the normal distribution, we can use the shape of the normal distribution to make inferences about our distribution of values. When we used *empirical probabilities* above, our probabilities were estimated only with respect to the data we observed. However, when we refer to theoretical probability distributions we can also think about the behavior of values we did not observe, or the behavior of the distribution in general. In order to calculate theoretical probabilities you first needed to commit to a model of the data. You may be thinking, what model? It may seem too simple to be a model, but by assuming that our data can be understood as coming from a normal distribution with some given and , we have already created a simple model for our data. This is analogous to committing to a spherical shape for our model of our billiards balls: Saying that you expect your data to be normally distributed commits you to a certain distribution ‘shape’ and to more and less probable parameter values for your variable.

In figure 2.5 we compare the histogram of apparent height judgments to the density of a normal distribution with a mean equal to the sample mean () and a standard deviation equal to the sample standard deviation () of our mens\_height vector. The density was drawn using the dnorm function, which draws a curve representing the shape of a theoretical normal distribution with a given mean and standard deviation. Clearly, there is a good alignment between our random sample of real-world data and the theoretical normal density. This suggests that we could potentially use the *theoretical* shape of the normal distribution to talk about the characteristics of our observed random sample of data. Although the distribution of our sample is unlikely to be perfectly normal, it is *normal enough* to make the comparison worthwhile.

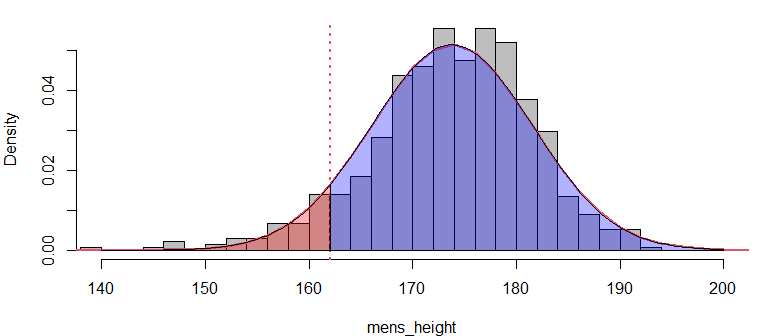


Figure 2.5: The histogram shows the empirical distribution of height judgments for adult male speakers. The shaded area shows the theoretical density of the equivalent normal distribution. The read area corresponds to the theoretical probability of observing a height under 162 cm.

The average female over 20 in the United states is 162.1 cm (CDC). A vertical line has been place at this value in figure 2.5. We might wonder, what is a probability of observing an height judgment for an adult male that is shorter than this average adult female height? Asking this question is equivalent to asking: What is the area under the curve of the density above, to the *left* of the vertical line? Since the *total* area of the density is always equal to one, the area of the red portion below corresponds to a percentage/probability of observing values less than 162.1 cm. One way to answer this question is to calculate the empirical probability of observing an apparent height less than 162.1 cm for male speakers in our data. Another way to do this is by calculating the *theoretical probability* by finding the proportion of values expected to be less 162.1 Hz in the normal distribution that has approximately the same ‘shape’ as our data distribution (i.e. the one seen in figure 2.5).

Below, we use the function pnorm to find the proportion of values that are expected to be less than 162.1 cm. This function takes in a value, a mean, and a standard deviation. It then tells you the proportion of the distribution that is to the *left* of (i.e. less than) a given value. Below, we use parameters estimated from our sample to run the pnorm function, as these are our best guesses of the population parameters. The output of this function is equal to the area of the red section in the density above. As we can see, the theoretical and empirical probabilities are very similar to one another. If we subtract this area from one, we get the area under the curve to the *right* of the vertical line, the blue section of the density above.

# empirical probability of height < 162.1  
mean (mens\_height<162.1)  
## [1] 0.07703704  
  
# Red area of distribution, x < 162.1  
pnorm (162.1, mean (mens\_height), sd(mens\_height))  
## [1] 0.06610382  
  
# Blue area of distribution, x > 162.1  
1 - pnorm (162.1, mean (mens\_height), sd(mens\_height))  
## [1] 0.9338962

Imagine you had 1 pound of clay and you were asked to make a shape *exactly* like the normal density above. This shape should be perfectly flat, i.e., it should have a constant depth (like a coin). If you had this shape made of clay used a knife to remove the part to the left of 162.1 cm (the red subsection) and weighed it, it should weigh 6.6% of a pound (0.066 pounds). The ‘area under the curve’ of this clay sculpture would just correspond to the amount of clay in a certain area, and in this case we know that only 6.6% of the clay should be in that section of the shape. So, the area under the curve, the probability, is just the amount of the *stuff* in the density that falls below/above a certain point, or between two points. The pnorm function allows you to slice and ‘weigh’ the sections of the distribution to tell you how much of it is in any given section.

What our theoretical probabilities tell us is this: *If* height judgments come from a normal distribution, *and* that distribution has a mean and standard deviation that is close to the sample estimates, then we expect (in the long run) that 6.6% of height judgments will be lower than 162.1 cm. What we are expressing here is effectively a conditional probability, we are saying *if* the parameters have certain values, *and* the probability distribution has a certain shape, then we expect certain height judgments to be more or less probable. Of course, if you change any part of that, either the values of the parameters or the probability distribution, then your estimated theoretical probabilities are likely to change. This is an important thing to keep in mind because it means that inference based on theoretical probabilities can change when our assumptions change, and many of these assumptions cannot be ‘proven’ to be true.

## 2.5 Probabilities of events and likelihoods of parameters

We’re going to change from talking about *probabilities* to talking about *likelihoods*. Probability is the odds of observing some data/event/outcome, given some parameter(s). A **likelihood** inverts this, and places odds on different *parameter* values given some observed data. For example, you could say “how probable is it that a random man from the US will be sound shorter than 162.1 cm in height?”. In contrast, you could ask “how likely is it that the average man from the US sounds 162.1 cm tall?”. The likelihood of a parameter represents the joint probability (density) of observing all the data you observed, given a certain parameter value. In other words, the likelihood relates to the probability that a specific probability distribution would produce your first observation, *and* your second observation, *and* your third observation, and so on, for all observations.

The **likelihood function** is a curve showing the relative likelihoods of different parameter values, given a fixed set of observations/data. The likelihood function tells you what parameter values are *credible* given your data. If a value is very unlikely, that means that it is not supported by your data. In other words, unlikely parameter estimates represent conclusions that your data is rejecting as not viable, and hence they are not **credible**. Every parameter for every probability distribution has a likelihood function, given some data. Here, we’re only going to discuss the likelihood of the normal mean parameter, , in detail.

The likelihood of the sample mean reflects the joint probability of observing all of your data, given different values of the mean. An example of how this is calculated is given in Figure 2.6. The left panel shows the likelihood function for based on a random sample of ten height judgments from our data (indicated by the blue points at the bottom of the plot). We can see that the most *likely* values of are centered on the bulk of the observations, and that values become less likely as we deviate from them. The vertical dotted lines indicate three possible mean values that will be highlighted in this discussion.

The likelihood of a parameter value (e.g., = 174 cm in the right panel of Figure 2.6) is equal to the product of the density of each observation in the sample, given the value of the parameter. This sounds like a mouthful but is actually deceptively simple. For example, to calculate the likelihood that , we:

1. Assume that the data is generated by a normal distribution with and equal to your sample standard deviation (7.8 cm).
2. Find the the height of the curve of the probability distribution (the density) over each point (indicated by the vertical lines in the right panel below). This reflects the relative probability of each observation given your parameter value.
3. The joint probability of all of the observations (the likelihood) is the product of all of these densities (heights). This assumes that all of your observations are statistically independent of each other (see section 2.2.2).

So, the value of the likelihood function in the left panel of Figure 2.6 at 174 Hz (rightmost vertical line) is equal to the product of the probabilities of the points in the right panel (i.e., the heights of the lines in the panel). Imagine we did this for a range of values along the axis, recording the likelihood values at each step. If we do this and then plot the product of the densities for each corresponding value the result would be a curve identical to that of the left panel in figure 2.6.

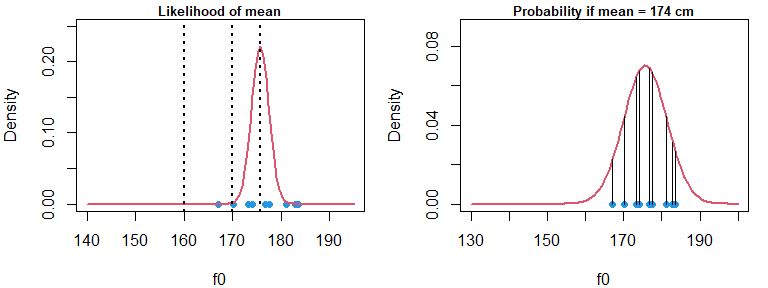


Figure 2.6: (left) The red curve indicates the likelihood of the population mean given the blue points in the figure. The vertical lines indicate three different parameters, or hypotheses, that will be considered. (right) The red curve indicates the probability of the points given an assumed mean of 174 cm. Vertical lines highlight the dentisty over each point, given the assumed mean

The right panel of figure 2.6 shows the probability of points assuming that the population mean is equal to the sample mean (for our tiny sample). We can see in the left panel of figure 2.6 that this is the most likely value for the mean parameter. When we said earlier that the sample mean is the ‘best’ estimate of the population mean, what we really meant was that the sample mean is the **maximum-likelihood** estimate of the population mean. This means that the sample mean provides an estimate of that maximizes the value of the likelihood function given the data. This is related to the fact that, as mentioned in section 2.3.1, the sample mean minimizes the variance of the sample. As a result, if you want to know which mean estimate is most likely given your data, you simply need to calculate the sample mean as in equation (2.1).

In the left panel of figure 2.7 we see that a normal distribution with a of 170 cm is a reasonable fit to the data. However, several observations are very improbable and this relative lack of fit is reflected by the low value of the likelihood function at 170 in Figure 2.6. In the right panel of Figure 2.7 we see that a normal distribution with a mean of 160 is very unlikely to generate this data: Many points are extremely improbable and have densities close to zero. Correspondingly, the value of the likelihood corresponding to in figure 2.6 reflects a very unlikely parameter value.

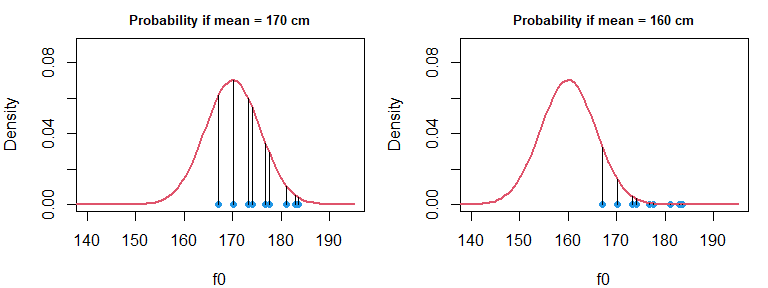


Figure 2.7: (left) The red curve indicates the probability of the points given an assumed mean of 170 cm. Vertical lines highlight the dentisty over each point, given the assumed mean. (right) The same information as in the left plot, but given a mean of 160 cm.

We need to talk about why it makes sense to multiply densities to calculate joint probabilities. Above, we stated that probabilities relate to areas under the curve. A problem we have is that the area under the curve of a single point is always zero. This is because a single point is so thin (it’s width is basically zero) that the area under the curve under the point will equal zero regardless of its density. Suppose that we said “ok let’s agree to use a fixed width, , around our point to calculate our area under the curve”. We can make so very tiny that is is almost as if we were calculating just the area under our density. We can use to represent the height of a density for a given value of . This means that we could approximate the value of the area under the curve at by treating it like a trapezoid and finding . If we wanted to calculate the joint probability of a number of observations, we could multiply a series of these areas, as seen in (2.11).

Since each term in (2.11) contains , this can be factored out as in (2.12).

At this point, we see that the joint probability of a set of values () can be thought of as the the product of the densities over those values () times some arbitrary constant . If we agree to use the same constant for all our calculations, we can ignore it and use the product of the densities to evaluate the relative probabilities of different combinations of observations.

### 2.5.1 Characteristics of likelihoods

We can think about the characteristics of the likelihood of to make some predictions about its expected shape given different numbers of observations and underlying data distributions. The likelihood is the joint probability of your data given different some parameter values. We can calculate the joint probability of two observations and by multiplying their individual densities as in equation (2.13), assuming these observations are independent.

Equation (2.13) defines the likelihood of given the data and and some . We will update the left-hand side to reflect this, replacing a term representing the joint probability of and () with a term representing the likelihood of the mean given the data, . Notice that nothing has changed except our perspective. In equation (2.13) we are treating the parameter as fixed and using it to calculate the joint probability of some data. In equation (2.14) we are treating the data as fixed and using it to calculate the likelihood of values of .

We can log-transform both sides of the equation to turn multiplication into addition, get rid of the exponentiation, and re-arrange the terms as in equation (2.15) (as explained for equation (2.8) above). Because they make working with likelihoods much simpler, statisticians often refer to the logarithms of likelihood functions, referred to as **log-likelihoods**, denoted using the symbol . When you see this, , just think “the logarithm of the likelihood of the mean given the data ”. In equation (2.15) we see that the log-likelihood of these two observations is the sum of two parabolas. Since the sum of parabolas will (almost always) also be a parabola, we know that the log-likelihood of the mean given these, and any number of other observations, will also be a parabola. As a result, we see that the (non-log) likelihood of the mean is an exponentiated parabola and has the same general shape as the normal distribution.

Equation (2.15) is specifically for two variables, however, Because each term is identical except for , the log-likelihood function can be greatly simplified as in equation (2.16). In (2.16) we can see that the log-likelihood of given observations of is equal to the sum of the individual squared deviations from the mean, divided by , with the value subtracted from it. So, we see that just like the log-likelihood is a parabola that has its vertex at . This means that this parabola has its maximum value when is equal to the sample mean. This is why the sample mean is the maximum-likelihood estimate for the population mean.

There are two mechanisms by which the likelihood function may get narrower or wider. The first is the mechanism discussed in section 2.3.3 where a larger value of results in a wider parabola (and wider likelihood), and it also applies here. However, there is another way that likelihood functions can become narrower, and that is by increasing the sample size. To see why this is the case, consider what happens to the relative value of our parameter () as the sample size grows. Imagine that we are calculating the sum of squared deviations about our sample mean. We know that the average squared variation we expect from our mean is equal to the variance. So, let’s replace in (2.16) with in (2.17). Now, instead of adding *N* squared deviations around the mean, , we are multiplying our expected squared deviation by *N*, .

We can also move the over from under the to make the following point simpler, as in (2.18).

Multiplying the numerator of a fraction is exactly equivalent to dividing the denominator of the fraction by the same amount, as in for example . Thus, we see that calculating the likelihood using data points is expected to have about the same effect on parabola width as dividing the data variance by , as shown in (2.19). As a result, increasing numbers of observations reduce the uncertainty in parameter estimates by making the likelihood narrower and narrower.

### 2.5.2 Making inferences using likelihoods

In Section 2.4 we discussed using the normal distribution to make inferences about the probable values of a random variable. When variables are normally distributed we can use the theoretical normal distribution and functions such as pnorm to answer questions about values we expect, and don’t expect, to see. In the same way, we can use likelihood functions to understand probable, and improbable, values of parameters given our data. For example, suppose that you measured the heights of 100 women in a small town (pop. 1500) and found the average height was 160 cm, with a standard deviation of 6 cm. You might accept that the *actual* population average is 161 cm, but may find it difficult to accept that it was actually 180 cm. This is because a true mean of 180 cm is *unlikely* given your observed data: The observations you have are *improbable* given a true mean of 180 cm. The logic is quite simple: a true mean of 180 cm is unlikely to ‘produce’ so many women around 160 cm. You *know* you observed the short women, therefore, you have no reason to believe that the true mean is 180 cm.

Below we calculate the likelihood of different mean parameters given our data. We do this by finding the log-density of each observation and then adding the points together. Log-densities are used because the likelihood is often a number so small that computers have a hard time representing them otherwise. For example, the highest point of the probability densities in figure 2.7 is about 0.07. Let’s pretend it’s 0.1 for the sake of simplicity. Recall that to find joint probabilities we need to multiply the densities above each of the points. This means that the probability of two observations at the mean is , and the probability of observing observations at the mean is equal to . Since we have 675 observations in our mens\_height vector, the probability of observing every one of those at the mean, the most probable outcome, would be equal to , or a decimal point followed by 674 zeros and then a one.

By relying on the logarithms of probabilities instead, we can accurately represent very small numbers more comfortably. This is because adding together the logarithms of two numbers is equivalent to multiplying those numbers, and multiplying a logarithm by another number is equivalent to raising it to that power (see (2.6)). For example, the number can also be expressed like , which equals -1554.245 (i.e. ).This is obviously a much easier number to deal with than one with 675 decimal places.

# make candidates for mean parameter  
mus = seq (172.5,175, .01)  
  
# easy way to make zero vector of same length as above  
log\_likelihood = mus\*0  
  
# add the log-density of all observations. Notice only the mean changes.  
for (i in 1:length(mus))   
 log\_likelihood[i] = sum (dnorm (mens\_height, mus[i], sd(mens\_height), log = TRUE))

In figure 2.8 we plot the log likelihood and the ‘scaled likelihood’. This is just the likelihood that has been scaled so that its peak is equal to one, and the peak of the log-likelihood equals zero. This allows us to actually plot and consider the likelihood function, though the values of the density no longer reflect the actual values of the likelihood function. However, the *relational* characteristics are maintained by this scaling. So, a scaled likelihood value of 0.2 is still five times less likely than a value of one.

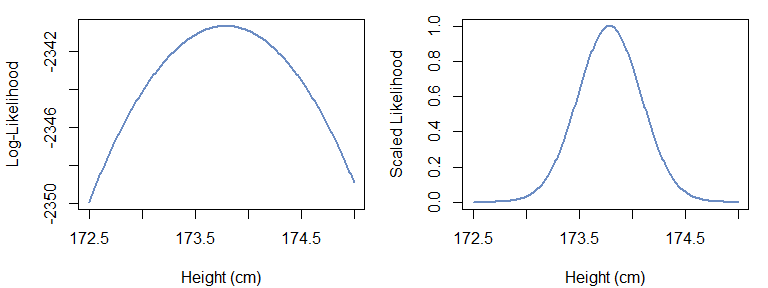


Figure 2.8: (left) Histogram of data. (right) Histogram of centered data, basically the error distribution.

At this point we can provide answers to the questions posed in section 2.1. The questions were:

Q1) How tall does the average man from the US sound?

Q2) Can we set limits on credible average apparent heights based on the data we collected?

Below, we see that the maximum likelihood estimate for the mean corresponds to the sample mean:

# find index number of highest values in log-likelihood  
maximum = which.max(scaled\_log\_likelihood)  
  
# print and compare to sample mean  
mus[maximum]

## [1] 173.79

mean (mens\_height)

## [1] 173.7877

So, we may conclude that the average male speaker is *most likely* to sound about 174 cm tall. We can also conclude informally based on Figure 2.8 that the most likely mean values fall between (approximately) 173 and 174.5 cm. This means that although the sample mean was 174 cm, it is reasonable that the true population mean might actually be 173.5 cm. This is because mean parameters in this range are also reasonably *likely* given our data. Basically, maybe our sample mean is wrong and arose by accident, and 173.5 cm is the true population . This outcome is compatible with our data. However, a value of 172.5 cm is very *unlikely* given our data. Since we think that 172.5 cm is not a plausible mean parameter given our sample, we can rule it out as a plausible value for the parameter of the underlying probability distribution. Using this approach, we can use the information in likelihood functions to rule out implausible values of based on the characteristics of our data.