The foundations of statistics:
A simulation-based approach

Shravan Vasishth[1] and Michael Broe[2]
[1] University of Potsdam, Germany
[2] The Ohio State University, USA

Book draft, version dated July 17, 2009

Preface

Statistics and hypothesis testing are routinely used in areas that are traditionally not mathematically intensive (an example is linguistics). In such fields, when faced with experimental data in any form, many students and researchers tend to rely on commerical packages to carry out statistical data analysis, often without acquiring much understanding of the logic of statistics they rely on. There are two major problems with this. First, the results are often misinterpreted. Second, users are rarely able to flexibly apply techniques relevant to their own research – they use whatever they happened to have learnt from their advisors, and if a slightly new data analysis situation arises, they are unable to use a different method.

A simple solution to the first problem is to teach the foundational ideas of statistical hypothesis testing without using too much mathematics. In order to achieve this, statistics instructors routinely present simulations to students in order to help them intuitively understand things like the Central Limit Theorem. This approach appears to facilitate understanding, but this understanding is fleeting. A deeper and more permanent appreciation of the foundational ideas can be achieved if students run and modify the simulations themselves.

This book addresses the problem of superficial undertanding. It provides a non-mathematical, simulation-based introduction to basic statistical concepts, and encourages the reader to try out the simulations themselves using the code provided. Since the exercises provided in the text almost always require the use of programming constructs previously introduced, the diligent student acquires basic programming ability as a side effect. This helps to build up the confidence necessary for carrying out more sophisticated analyses. The present book can be considered as the background material necessary for more advanced courses in statistics.

The vehicle for simulation is a freely available software package, R (see the CRAN website for further details). This book is written using Sweave (Leisch, 2002) (pronounced S-weave), which means that LaTeX and R code are interwoven into a single source document. This approach to mixing description with code also encourages the user to adopt literate programming from the outset, so that the end product of their own data analyses is a reproducible and readable program.

The style of presentation used in this book is inspired by a short course taught in 2000 by Michael Broe at the Linguistics department of The Ohio State University. The first author (SV) was a student at the time and attended Michael's course; later, SV extended the book in the spirit of the original course (which was prepared using Mathematica).

Since then, SV has used this book to teach linguistics undergraduate and graduate students (thanks to all the participants in these classes for feedback and suggestions for improving the course contents). It appears that the highly motivated reader with little to no programming ability and/or mathematical/statistical training can understand everything presented here, and can move on to using R and statistics productively and sensibly.

The book is designed for self-instruction or as a textbook in a statistics course that involves the use of computers. Many of the examples are from linguistics, but this does not affect the content, which is of general relevance to any scientific discipline.

We do not aspire to teach R per se in this book; if this book is used for self-instruction, the reader is expected to either take the initiative themselves to acquire a basic understanding of R, and if this book is used in a taught course, the first few lectures should be devoted to a simple introduction to R.

After completing this book, the reader will be ready to understand more advanced books like Gelman and Hill's Data analysis using regression and multilevel/hierarchical models, Baayen's *Analyzing Linguistic Data*, and Roger Levy's online lecture notes.

Contents

Chapter 1

Getting started

The main goal of this book is to help you understand the principles behind inferential statistics, and to use and customize statistical tests to your needs. The vehicle for this will be a programming language called R.

1.1 Installation: R, LaTeX, and Emacs

The first thing you need to do is get hold of R. The latest version can be downloaded from the CRAN website. The more common operating systems are catered for; you will have to look at the instructions for your computer's operating system.

After you have installed R on your machine, the second thing you need to do before proceeding any further with this book is to learn a little bit about R. The present book is not intended to be an introduction to R. For short, comprehensive and freely available introductions, look at the Manuals on the R web page, and particularly under the link "Contributed." You should spend a few hours studying the Contributed section of the CRAN archive. In particular you need to know basic things like starting up R, simple arithmetic operations, and quitting R. It is possible to use this book and learn R as you read, but in that case you have to be prepared to look up the online help available with R.

In addition to R, other freely available software provides a set of tools that work together with R to give a very pleasant computing environment. The least that you need to know about is LaTeX, Emacs, and Emacs Speaks Statistics. Other tools that will further enhance your working experience with LaTeX are AucTeX, RefTeX, preview-latex, and python. None of these are required but are highly recommended for typesetting and other sub-tasks necessary for data analysis.

There are many advantages to using R with these tools. For example, R and LaTeX code can be intermixed in emacs using noweb mode. R can output data tables etc. in LaTeX format, allowing you to efficiently integrate your scientific writing with the data analysis. This book was typeset using all of the above tools.

The installation of this working environment differs from one operating system to another. In Linux-like environments, most of these tools are already pre-installed. For Windows you will need to read the manual pages on the R web pages. If this sounds too complicated, note that in order to use the code that comes with this book, you need only to install R.

1.2 Some simple commands in R

We begin with a short session that aims to familiarize you with R and very basic interaction with data.

Let's assume for argument's sake that we have the grades of eleven students in a final examination for a statistics course. Both the instructor and the students are probably interested in finding out at least the maximum and minimum scores. But hidden in these scores is much more information about the students.

Assuming a maximum possible score of 100, let's first start up R and input the scores (which are fictional). Then we ask the following questions using R: (a) what's the maximum score? (b) what's the minimum?

```
> scores <- c(99, 97, 72, 56, 88, 80, 74, 95, 66, 57, 89)

 [1] 99 97 72 56 88 80 74 95 66 57 89

> max(scores)

[1] 99

> min(scores)

[1] 56
```

We could stop here. But there is much more information in this simple dataset, and it tells us a great deal more about the students than the maximum and minimum grades.

The first thing we can ask is: what is the average or mean score? For any collection of numbers, their mean is the sum of the numbers divided by the length of the vector:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.1}$$

The notation $\sum\limits_{i=1}^{n}$ is simply an abbreviation for the list of numbers going from x_1 to x_n.

The mean tells you something interesting about that collection of students: if they had all scored high marks, say in the 90's, the mean would be high, and if not then it would be relatively low. The mean gives you one number that summarizes the data succinctly. We can ask R to compute the mean as follows:

```
> mean(scores)

[1] 79.36364
```

Another such summary number is called the variance. It tells you how far away the individual scores are from the mean score on average, and it's defined as follows:

$$\text{variance} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{1.2}$$

The variance formula gives you a single number that tells you how "spread out" the scores are with respect to the mean. The smaller the spread, the smaller the variance. So let's have R calculate the variance for us:

```
> var(scores)

[1] 241.6545
```

Notice that the number is much larger than the maximum possible score of 100; this is not surprising because we are squaring the differences of each score from the mean when we compute variance. It's natural to ask what the variance is in the same scale as the scores themselves, and to achieve this we can simply take the square root of the variance. That's called the STANDARD DEVIATION, and it's defined like this:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{1.3}$$

Here's how to compute it in R; you can easily verify that it is indeed the square root of the variance:

```
> sd(scores)

[1] 15.54524

> sqrt(var(scores))

[1] 15.54524
```

At this point you are likely to have at least one question about the definition of variance (1.2). *Why do we divide by $n-1$ and not n?* One answer to this question is that the sum of deviations from the mean is always zero, so if we know $n-1$ of the deviations, the last deviation is predictable. The mean is an average of n unrelated numbers and that's why the formula for mean sums up all the numbers and divides by n. But s is an average of $n-1$ unrelated numbers. The unrelated numbers that give us the mean and standard deviation are also called the DEGREES OF FREEDOM.

Let us convince ourselves of the observation above that the sum of the deviations from the mean always equals zero. To see this, let's take a look at the definition of mean, and do some simple rearranging.

1. First, look at the definition of mean:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{1.4}$$

2. Now move over the n to the left-hand side:

$$n\bar{x} = x_1 + x_2 + \cdots + x_n \tag{1.5}$$

3. Now if we subtract $n\bar{x}$ from both sides

$$n\bar{x} - n\bar{x} = x_1 + x_2 + \cdots + x_n - n\bar{x} \tag{1.6}$$

4. we get

$$0 = x_1 - \bar{x} + x_2 - \bar{x} + \cdots + x_n - \bar{x} \tag{1.7}$$

This fact implies that if we know the mean of a collection of numbers, and all but one of the numbers in the collection, the last one is predictable. In equation (1.7) above, we can find the value of ("solve for") any one x_i if we know the values of all the other x's.

Thus, when we calculate variance or standard deviation, we are calculating the average deviation of $n-1$ unknown numbers from the mean, hence it makes sense to divide by $n-1$ and not n as we do with mean. We return to this issue again in Section 5.18.

There are other summary numbers too that can tell us about the center-point of the scores, and their spread. One measure is the MEDIAN. This is the midpoint of a sorted (increasing order) list of a distribution. For example, the list 1 2 3 4 5 has median 3. In the list 1 2 3 4 5 6 the median is the mean of the two center observations. In our running example:

```
> median(scores)

[1] 80
```

The QUARTILES Q_1 and Q_3 are measures of spread about the median. They are the median of the observations below (Q_1) and above (Q_3) the 'grand' median. We can also talk about spread about the median in terms of the INTERQUARTILE RANGE (IQR): $Q_3 - Q_1$. It is fairly common to summarize a collection of numbers in terms of the FIVE-NUMBER SUMMARY: Min Q_1 Median Q_3 Max

The R commands for these are shown below; here you also see that the command `summary` gives you several of the measures of spread and central tendency we have just learnt.

```
> quantile(scores, 0.25)

25%
 69

> IQR(scores)

[1] 23

> fivenum(scores)

[1] 56 69 80 92 99

> summary(scores)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  56.00   69.00   80.00   79.36   92.00   99.00
```

1.3 Graphical summaries

Apart from calculating summary numbers that tell us about the center and spread of a collection of numbers, we can also get a graphical overview of these measures. A very informative plot is called the boxplot: it essentially shows the five number summary. The box in the middle has a line going through it, that's the median. The lower and upper ends of the box are Q1 and Q3 respectively, and the two "whiskers" at either end of the box extend to the minimum and maximum values.

```
> boxplot(scores)
```

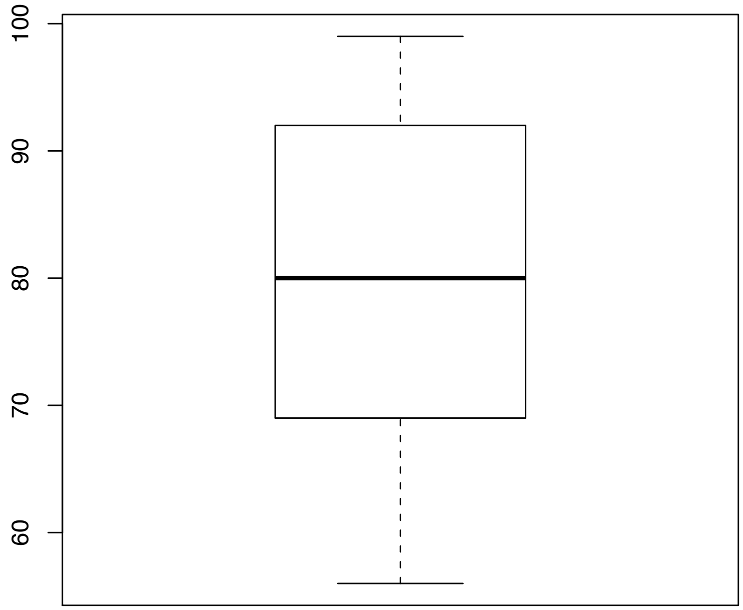


Figure 1.1: A boxplot of the *scores* dataset.

Another very informative graph is called the histogram. What it shows is the number of scores that occur within particular ranges. In our current example, the number of scores in the range 50-60 is 2, 60-70 has 1, and so on. The `hist` function can plot it for us; see Figure 1.2.

1.4 Acquiring basic competence in R

At this point we would recommend working through Baron and Li's excellent tutorial on basic competence in R. The tutorial is available in the Contributed section of the CRAN website.

```
> hist(scores)
```

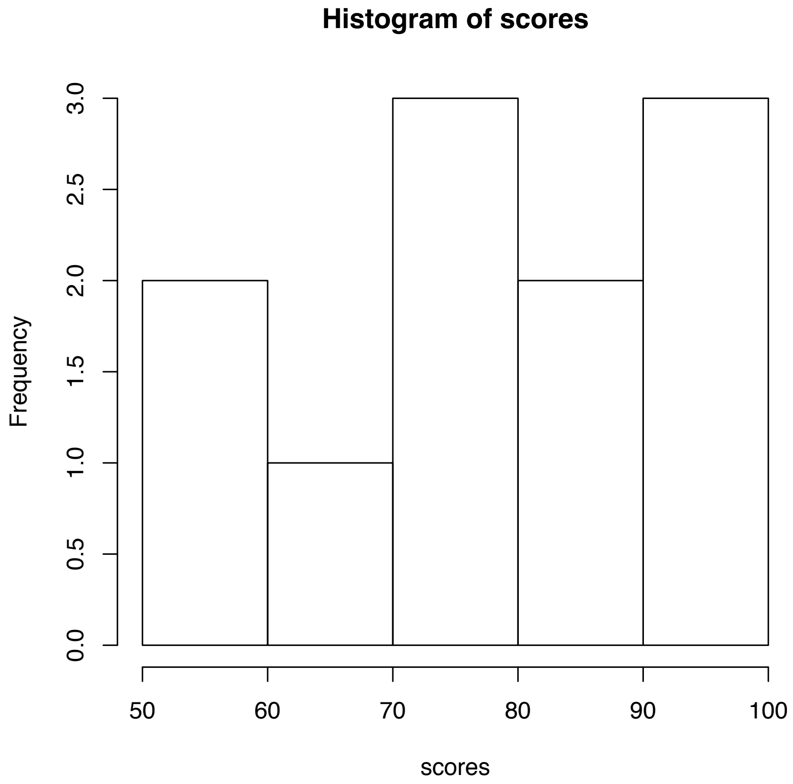


Figure 1.2: A histogram of the *scores* dataset.

1.5 Summary

Collections of scores such as our running example can be described graphically quite comprehensively, and/or with a combination of measures that summarize central tendency and spread: the mean, variance, standard deviation, median, quartiles, etc. In the coming chapters we use these concepts repeatedly as we build up the theory of hypothesis testing from the ground up. But first we have to acquire a very basic understanding of probability theory.

Chapter 2

Randomness and Probability

Suppose that, for some reason, we want to know how many times a second-language learner makes an error in a writing task; to be more specific, let's assume we will only count verb inflection errors. The dependent variable (here, the number of inflection errors) is random in the sense that we don't know in advance exactly what its value will be each time we assign a writing task to our subject. The starting point for us is the question: What's the pattern of variability (assuming there is any) in the dependent variable?

The key idea for inferential statistics is as follows: If we know what a "random" distribution looks like, we can tell random variation from non-random variation. We will start by supposing that the variation observed is random – and then try to prove ourselves wrong. This is called "Making the null hypothesis."

In this chapter and the next, we are going to pursue this key idea in great detail. Our goal here is to look at distribution patterns in random variation (and to learn some R on the side). Before we get to this goal, we need know a little bit about probability theory, so let's look at that first.

2.1 Elementary probability theory

2.1.1 The sum and product rules

We will first go over two very basic facts from probability theory. Amazingly, these are the only two facts we need for the entire book. We are going to present these ideas completely informally. There are very good books that cover more detail; in particular we would recommend Introduction to Probability by Charles M. Grinstead and J. Laurie Snell. The book is available online.

Consider the toss of a fair coin, which has two sides, H(eads) and T(ails). Suppose we toss the coin once. What is the probability of an H, or a T? You might say, 0.5, but why do you say that? You are positing a theoretical value based on your prior expectations or beliefs about that coin. (We leave aside the possibility that the coin lands on its side.) We will represent this prior expectation by saying that $P(H) = P(T) = \frac{1}{2}$.

Now consider what all the logically possible outcomes are: an H or a T. What's the possibility of either one of these happening when we toss a coin? Of course, you'd say, 1; we're hundred percent certain it's going to be an H or a T. We can express this intuition as an equation, as the sum of two mutually exclusive events:

$$P(H) + P(T) = 1 \tag{2.1}$$

There are two things to note here. One is that the two events are *mutually exclusive*; you can't have an H and a T in any one coin toss. The second is that these two events exhaust all the logical possibilities in this example. The important thing to note is that **the probability of any mutually exclusive events occurring is the sum of the probabilities of each of the events**. This is called the SUM RULE.

To understand this idea better, think of a fair six-sided die. The probability of each side s is $\frac{1}{6}$. If you toss the die once, what is the probability of a 1 or a 3? The answer is $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$.

Suppose now that we have not one but *two* fair coins and we toss each one once. What are the logical possibilities now? In other words, what sequences of heads and tails are possible? I think you'll agree that the answer is: HH, HT, TH, HH, and also that *all of these are equiprobable*. In other words: P(HH)=P(HT)=P(TH)=P(TT). There are four possible EVENTS and each is equally likely. This implies that the probability of each of these is $P(HH) = P(HT) = P(TH) = P(TT) = \frac{1}{4}$. If you see this intuitively, you also understand intuitively the concept of PROBABILITY MASS. As the word "mass" suggests, we have redistributed the total "weight" (1) equally over all the logically possible outcomes (there are 4 of them).

Now consider this: the probability of any one coin landing an H is $\frac{1}{2}$, and of landing a T is also $\frac{1}{2}$. Suppose we toss the two coins one after another as discussed above. What is the probability of getting an H with the first coin followed by a T in the second coin? We could look back to the previous paragraph and decide the answer is $\frac{1}{4}$. But probability theory has a rule that gives you a way of calculating the probability of this event:

$$P(H) \times P(T) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \tag{2.2}$$

In this situation, an H in the first coin and an H or T in the second are completely independent events—one event cannot influence the other's outcome. This is the PRODUCT RULE, which says that **when two or more events are independent, the probability of both of them occurring is the product of their individual probabilities**.

And that's all we need to know for this book. At this point, you may want to try solving a simple probability problem: Suppose we toss three coins; what are the probabilities of getting 0, 1, 2, and 3 heads?

2.1.2 Stones and rain: A variant on the coin-toss problem

Having mastered the two facts we need from probability theory, we finally begin our study of randomness and uncertainty, using simulations.

Because the coin example is so tired and over-used, we take a different variant for purposes of our discussion of probability. Suppose we have two identical stones (labeled L, or 0; and R, or 1), and some rain falling on them. We will now create an artificial world in R and observe the raindrops falling on the stones. We can simulate the falling of one raindrop quite easily in R:

```
> rbinom(1, 1, 0.5)

[1] 0
```

The above command says that, assuming that the prior probability of a R-stone hit is 0.5 (a reasonable assumption), sample one drop once. If we want to sample two drops, we say:

```
> rbinom(2, 1, 0.5)

[1] 1 1
```

In the next piece of R code, we will "observe" 40 raindrops and if a raindrop falls on the right stone, we write down a 1, else we write a 0.

```
> size <- 1
> p <- 0.5
> fortydrops <- rbinom(40, size, p)
```

At this point you might be wondering what the function `rbinom` is in the code chunk below. You can either ignore it for now, or type `?rbinom` at the R command prompt to get a summary of what it does. Notice that we store the number of right-stone hits in a variable we will call `fortydrops`; the name can be anything but it's better to use an informative name rather than obscure one like `x`.

Next, we can ask R to tell us: what were the total number of right-stone hits in the 40-drop sequence? And what was the *proportion* of right-stone hits? To do this we just need to calculate the mean number of 1's in the 40-drop sequence:

```
> sum(fortydrops)/40
```

```
[1] 0.45
```

Using R we can ask an even more informative question: instead of just looking at 40 drops *only once*, we do this many times. We observe 40 drops again and again i times, where $i = 15, 25, 50, 100, 500, 1000$; and for each observation (going from 1st to 15th, 1 to 25*th*, and so on), we note the number of Right-stone hits. After i observations, we can record our results in a vector of Right-stone hits; we call this the vector *results* below. If we plot the distribution of Right-stone hits, a remarkable fact becomes apparent: the most frequently occurring value in this list is (about) 20.

The code and the final plot of the distributions is shown in Figure 2.1. Let's expend some energy trying to understand what this code does before we go any further.

1. We are going to plot six different histograms, each corresponding to the six values $i = 15, 25, 50, 100, 500, 1000$. For this purpose, we instruct R to plot a 2 × 3 plot. That's what the command below does:

   ```
   > op <- par(mfrow = c(2, 3), pty = "s")
   ```

2. The next few lines are just fixed values for the simulation and should be self-explanatory.

   ```
   > size <- 1
   > p <- 0.5
   > k <- 40
   > observations <- c(15, 25, 50, 100, 500, 1000)
   > n <- length(observations)
   ```

3. Then two *for*-loops begin. The first *for*-loop sets up things so that each of the six values (15,25,...) is considered. The second *for*-loop ensures that the 40 drops are counted i=1 ... 15 times, then i=1 ... 25 times, and so on. Each time the 40 drops are counted, the total number of Right-stone hits is recorded and stored in a vector called *results*.

   After calculating the number of Right-stone hits for each of the values, the distribution of each set of results is successively plotted. For example, when 1... 15 observations are made, there are 15 values in the *results* vector, each showing the total number of Right-stone hits. When 25 observations are made, there are 25 values in the *results* vector, and so on.

Exercise 1 — Plotting proportions and modifying the plot

```
> op <- par(mfrow = c(2, 3), pty = "s")
> for (j in 1:n) {
+     results <- rep(NA, observations[j])
+     for (i in 1:observations[j]) {
+         results[i] <- sum(rbinom(k, size, p))
+     }
+     title <- paste(c("Num Obs.", observations[j], sep = " "))
+     hist(results, ylab = "Frequency", xlab = "No. of R-stone hits",
+         main = title)
+ }
```

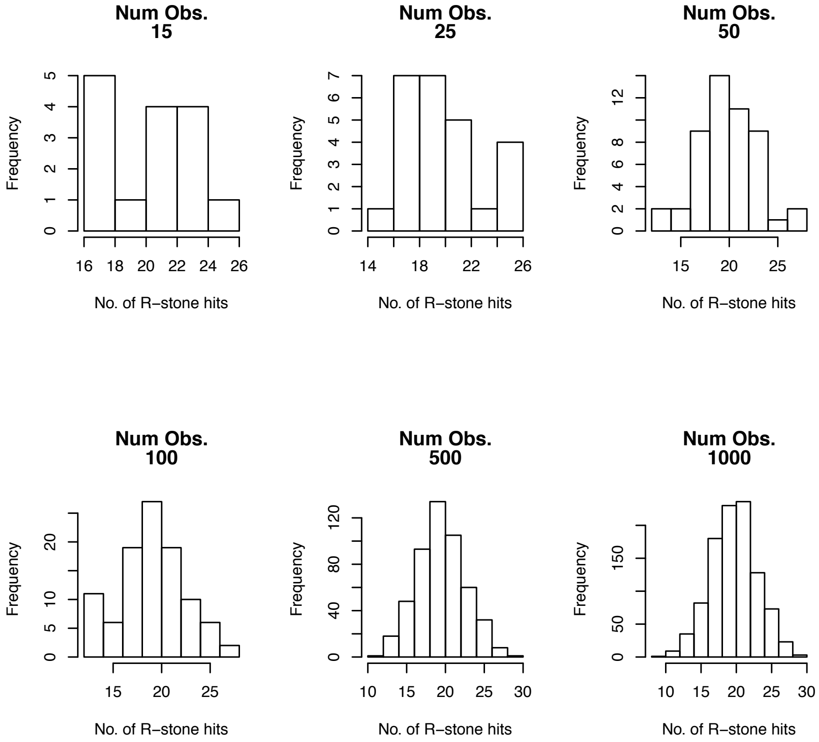


Figure 2.1: The frequency of Right-stone hits as the number of observations increases from 15 to 1000. Note that, as the number of observations increases, the most frequently occurring number of Right-stone hits is in the range of 20–exactly half the total number of drops observed each time.

1. Make a small change in the code above to plot, instead of the absolute number (sum) of right-stone hits, plot the proportion of Right-stone hits.
2. Change the x-axis description in each plot to reflect the fact that we are now looking at the proportion of right-stone hits rather than the absolute number.

The stabilization about a central value that you see in Figure 2.1 is typical of random phenomena. The central value here is 20. A common definition of probability is this theoretical stable final value of frequency. In the stones examples, we assumed that the theoretical probability of a Right-stone hit is 0.5. This decision was based on our beliefs about the situation under discussion (here, our belief was that Right-stone and Left-stone hits have equal probability).

Now, consider what happens when we observe the fall of four drops four times. What is the prior probability of there being 0 . . . 4 Right-stone hits? We can do this computation by filling in a table like this:

Number of R-stone hits	0	1	2	3	4
Probability of R-stone hits	?	?	?	?	?

Table 2.1: A tabulation of the probabilities of 0. . . 4 Right-stone hits when we observe four drops.

Suppose that in one observation or event E_1 we got RRRL (in that order). That is:

$$E_1 = (R \wedge R \wedge R \wedge L) \tag{2.3}$$

What's the probability of this happening? Well, we "know" that P(L)=P(R)=$\frac{1}{2}$, and we know the multiplication rule for independent events.

$$P(E_1) = P(R) \times P(R) \times P(R) \times P(L) = \frac{1}{16} \tag{2.4}$$

But there are four distinct ways to get three Rights, call them E_1, E_2, E_3, E_4: E_1 = RRRL, E_2 = RRLR, E_3 = RLRR, E_4 = LRRR. So we have a complex event E, made up of four mutually exclusive possibilities:

$$E = E_1 \vee E_2 \vee E_3 \vee E_4 \tag{2.5}$$

which means we can use the summation rule:

$$P(E) = P(E_1) + P(E_2) + P(E_3) + P(E_4) = \frac{1}{4} \tag{2.6}$$

You can already see figuring out the answer is going to be a pretty tedious business. Let's think of a better way to work this out. Towards this end, consider a simpler scenario: three fair coins tossed once each. What is the probability of getting 0 . . . 3 heads? Let's visualize this; see Figure 2.2.

Figure 2.2 helps us work out the relevant answers:

- Probability of zero heads: 0.5^3
- Probability of only one head: 3×0.5^3
- Probability of exactly two heads: 3×0.5^3
- Probability of three heads: 0.5^3

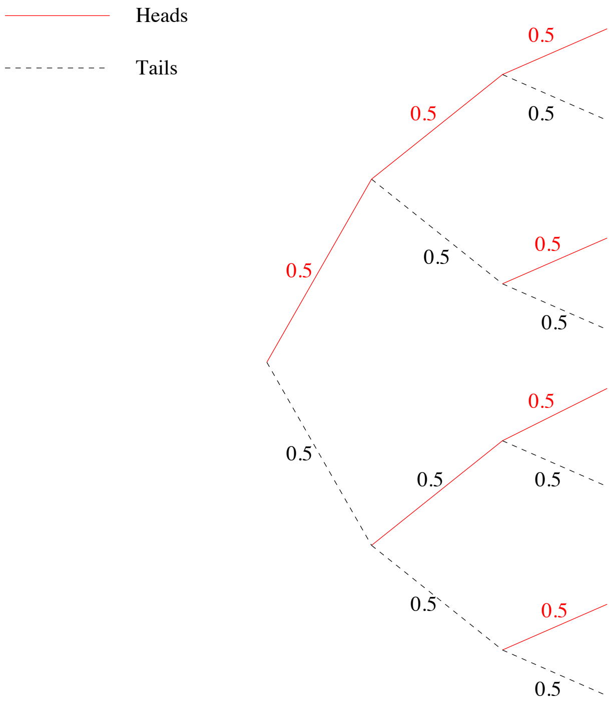


Figure 2.2: The probability space for three fair coins tossed once.

How did we do this calculation? Well, the left side of the tree that represents the probability space is the initial state, when no coin has been tossed. When we toss a coin once, we can get either a heads or a tails, and these mutually exclusive events are represented by the two edges emanating from the left-hand side. Each is an equi-probable event. After each of these possible events, another coin toss will yield a heads or a tails, and so on. So if we go from the left-hand side to the right, following each possible path in the tree, we have all the logical possibilities of heads and tails in this three-coin toss example. If we multiply the probabilities along each path of the tree and then add them up, they will sum to 1. This visualization method generalizes to our four-drop example, which we leave as an exercise for the reader (see below).

2.2 The binomial theorem

The above tree-based procedure that we used to calculate the probabilities yields a generalization: When we have a set of n items and choose k items from the set, all the possible ways to choose these k is $\binom{n}{k} = \frac{n!}{k! \times (n-k)!}$. Any discrete mathematics text (e.g., Rosen, 1994) will give you more details if you are unfamiliar with this and/or want to know more (this book does not require any further study of the binomial theorem).

Using this fact, we can compute the probability of k Right-stone hits when we make n observations, when the prior probability of a Right-stone hit is p:

$$\binom{n}{k} \times p^k (1-p)^{n-k} \tag{2.7}$$

Exercise 2 — A simple probability problem

Try out the above equation "by hand"; compute the probability of 0 ... 3 heads in the coins example. You can use R as a calculator.

The formula above is the binomial theorem, and can be applied when there are only two possible outcomes, the fixed, n observations are mutually independent, the probability p of a "success" is the same for each observation. This brings us to the binomial distribution.

Recall the big generalization about stabilization around the mean value when we sample raindrops: As the number of observations (the number of times we observe 40-drop sequences) increases from 10 to 10,000, the *relative frequency* of R-stone hits settles down to a stable value. The *distribution* of R-stone hits has a stable final shape. Just as we expressed the final *value* in terms of a *theoretical probability*, so we can express the final *shape* in terms of a *theoretical probability distribution* (which we arrived at empirically). The stones example is a perfectly random process; in the "long run" (in the limit, i.e., when the number of observations approaches infinity) the relative frequency of R-stone hits will settle at 0.5. However, at anything **less** than the long run, it's not always true that **exactly** half the observations will be R-stone hits.

2.3 Some terminology

We got to these conclusions by looking at a *limited number of observations*. This is called a **sample**. The number of *possible* R-stone hits (or the possibility of a heads or tails in a coin, or the possibility of getting 1...6 in a die) is called a *random variable*. The quantity computed from a sample (here, the number of R-stone hits) is called a *statistic*. The statistic we have been computing is also called a *sample count*. E.g., the binomial distribution is the sampling distribution of a sample count. A number that describes the population (e.g., mean) is called a *parameter*. An important point is that we usually don't know what this parameter is. Our focus in the coming chapters is going to be on the estimation of one or more of these parameters.

2.4 Back to the stones

Earlier we looked at the sampling distribution of the sample count using a 40-drop sequence. Suppose we plot the result of 100 observations of n drops, where n is (a) 4, (b) 40, and (c) 400, and (d) 4000. And we calculate the mean and standard deviation in each case (Figure 2.3).

As we increase the number of drops observed from 4 to 4000 (and observe these n-drop sequences 100 times), the spread, i.e., the standard deviation, decreases. We can see that visually in Figure 2.3. Or does it? Let's plot the standard deviation by sample size. This time let's look at drops going from 1 to 400 (Figure 2.4).

To our great surprise, we get increasing values for standard deviation as sample size increases. What's going on is that in *absolute* terms standard deviation is increasing, but not if we *relativize* it to the differing sample sizes – they're not comparable as things stand. If we look at proportions rather than absolute sample counts, we'll have normalized the various sample sizes so that they're comparable. So, when we look at, e.g., 40 drops each time, instead of saying each time, "18 Right-stone hits", we say "the proportion of R-stone hits is 18/40." Let's plot by proportion rather than sample count and see what we get. At this juncture you should spend a few minutes trying to modify the above code in order to plot normalized sample counts rather than absolute ones; we should get the same distribution as before (Figure 2.5). Now let's plot the standard deviation of the proportion-based counts (Figure 2.6).

Now everything makes sense: the spread, or equivalently standard deviation, decreases as we increase sample size. This is an important insight that we will come back to.

```
> size <- 1
> p <- 0.5
> drops <- c(4, 40, 400, 4000)
> op <- par(mfrow = c(2, 2), pty = "s")
> for (num.drops in drops) {
+     results <- rep(NA, 100)
+     for (i in 1:100) {
+         results[i] <- sum(rbinom(num.drops, size, p))
+     }
+     maintitle <- paste(num.drops, "drops", sep = " ")
+     hist(results, xlim = range(c(0:num.drops)), xlab = "Number of R-stone hits",
+         main = maintitle)
+ }
```

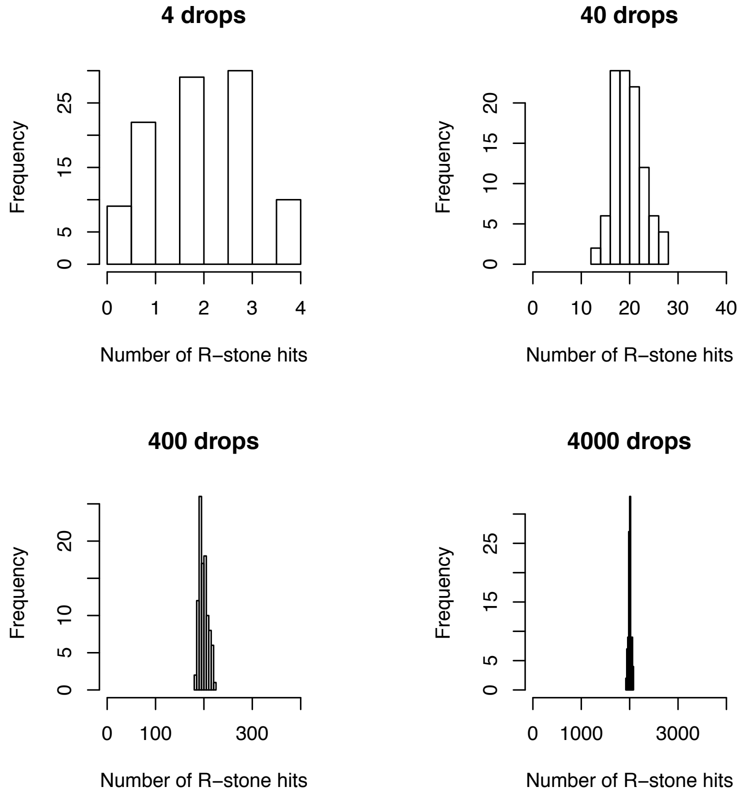


Figure 2.3: Increasing the number of drops observed from 4 to 4000 results in a tighter distribution.

```
> drops <- rep(1:400, 1)
> standard.dev <- rep(NA, 400)
> for (j in 1:400) {
+     results <- rep(NA, 100)
+     for (i in 1:100) {
+         results[i] <- sum(rbinom(drops[j], size, p))
+     }
+     standard.dev[j] <- sd(results)
+ }
> plot(drops, standard.dev, xlim = c(1, 400), xlab = "Number of drops",
+     ylab = "Standard deviation", main = expression("SD " * italic("increases") *
+         " as we increase sample size."))
```

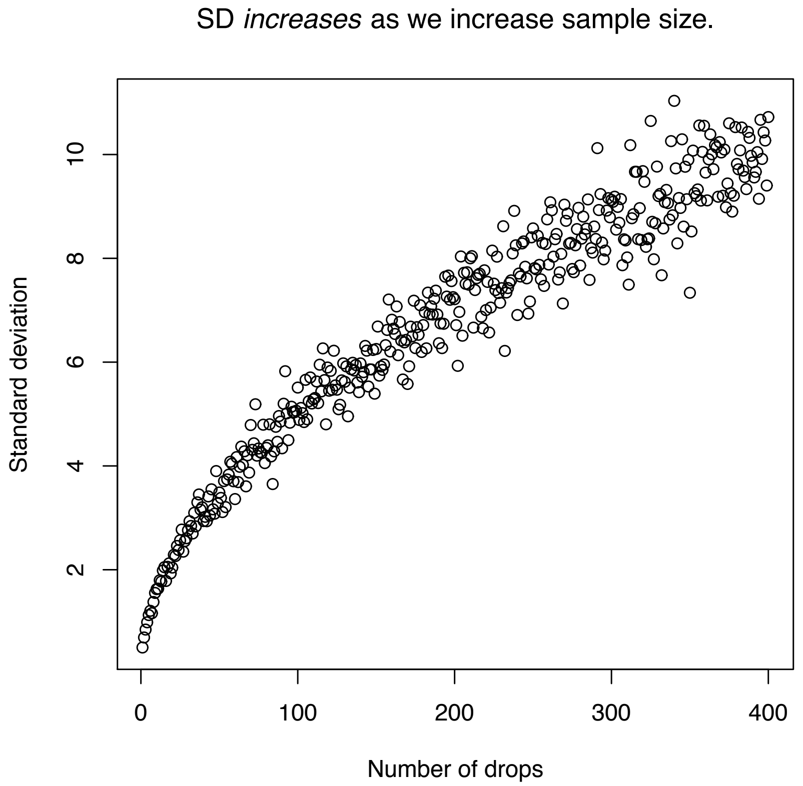


Figure 2.4: Standard deviation seems to increase as we increase sample size, which does not make any sense given the preceding figure showing increasing tighter distributions as sample size increases.

```
> size <- 1
> p <- 0.5
> drops <- c(4, 40, 400, 4000)
> op <- par(mfrow = c(2, 2), pty = "s")
> for (num.drops in drops) {
+     results <- rep(NA, 100)
+     for (i in 1:100) {
+         results[i] <- mean(rbinom(num.drops, size, p))
+     }
+     maintitle <- paste(num.drops, "drops", sep = " ")
+     hist(results, xlim = range(c(0:1)), xlab = "Proportion of R-stone hits",
+         main = maintitle)
+ }
```

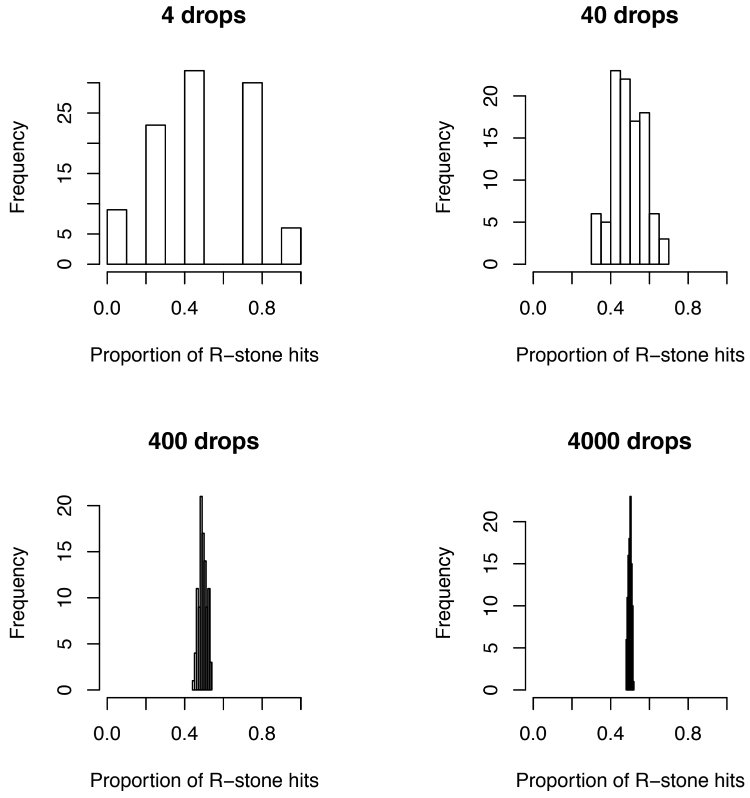


Figure 2.5: Plot of proportions of Right-stone hits as sample size increases.

```
> drops <- rep(1:400, 1)
> standard.dev <- rep(NA, 400)
> for (j in 1:400) {
+     results <- rep(NA, 100)
+     for (i in 1:100) {
+         results[i] <- mean(rbinom(drops[j], size, p))
+     }
+     standard.dev[j] <- sd(results)
+ }
> plot(drops, standard.dev, xlim = c(1, 400), xlab = "Number of drops",
+     ylab = "Standard deviation", main = expression("SD now decreases as we increase sample size."))
```

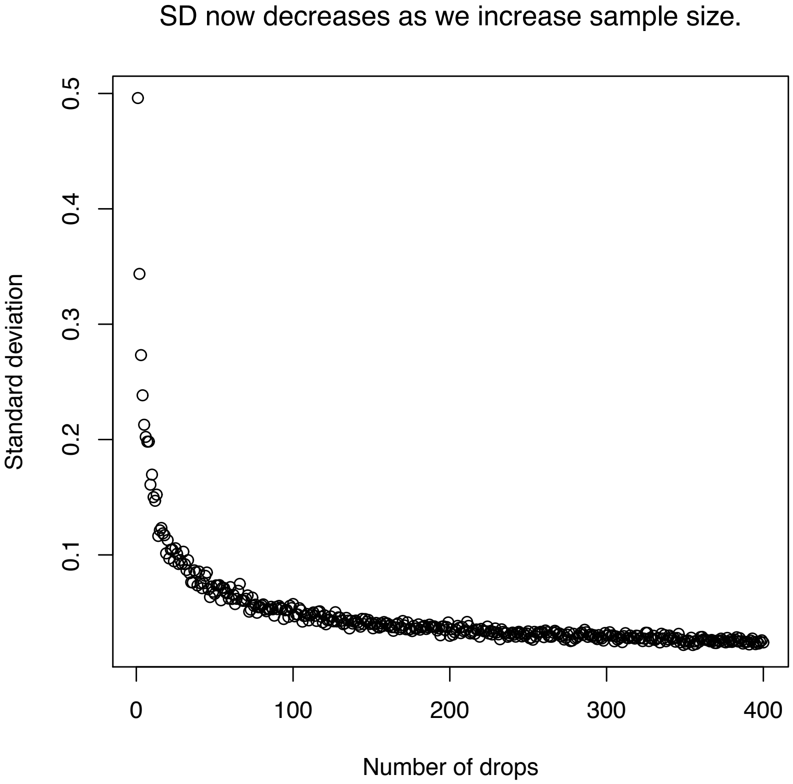


Figure 2.6: When we look at the standard deviation of proportions of Right-stone hits, we see that SD decreases as sample size increases, as expected.

2.4.1 Another insight: mean minimizes variance

Recall that variance is defined as follows:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.8}$$

Standard deviation, s, is a measure of spread about the mean, as we just saw. Recall our earlier observation that the sum of deviations from the mean will always equal zero. A related fact is that the squared deviations from the mean are smaller than from any other number – the mean is a special number in that sense. Together with the standard deviation, the mean is a good summary number for a set of scores.

Let's quickly convince ourselves that the squared deviations from the mean are smaller than from any other number (Figure 2.7).

2.5 Balls in a box: A new scenario

Suppose now that we have 12,000 balls in a big box, we *know* that 8000 (i.e., 2/3) are white, the others red. Suppose we take a random sample of 100 balls from these 12,000. We'd expect to draw about 66 white balls. What's the probability of getting *exactly* 66? We need the binomial theorem:

$$\binom{n}{k} \times p^k (1-p)^{n-k} \tag{2.9}$$

Let's first define a function in R to calculate this quickly:

```
> binomialprobability <- function(n, p, k) {
+     choose(n, k) * p^k * (1 - p)^(n - k)
+ }
```

If we run that function now with n=100, k=66, p=2/3, we find that it's fairly unlikely that we will get *exactly* 66:

```
> binomialprobability(100, 2/3, 66)
```

```
[1] 0.08314174
```

As an aside, note that R actually provides this function under the obscure name of *dbinom*:

```
> dbinom(66, 100, 2/3)
```

```
[1] 0.08314174
```

Now we ask an interesting question: suppose the sample was larger, say 1000. Would the probability of drawing 2/3 white balls from the 1000 balls (666.67 white balls) be higher or lower than the number we got above, 0.083141738816924? Think about this before reading further.

Let's work this out. With a sample size of sixty, what's the probability of two-thirds (40) being white?

```
> binomialprobability(60, 2/3, 40)
```

```
[1] 0.1087251
```

With a sample size of six hundred, what's the probability of two-thirds (400) being white?

```
> size <- 1
> p <- 0.5
> num.drops <- 4000
> results <- rep(NA, 100)
> for (i in 1:100) {
+     results[i] <- sum(rbinom(num.drops, size, p))
+ }
> mean.results <- mean(results)
> n <- floor(mean.results - 1)
> m <- floor(mean.results + 1)
> xvalues <- c(1:n, mean.results, m:4000)
> totaldev <- rep(NA, length(xvalues))
> for (i in xvalues) {
+     vectori <- rep(i, 100)
+     diffs <- results - vectori
+     sqdeviations <- sum(diffs * diffs)
+     totaldev[i] <- sqdeviations
+ }
> plot(xvalues, totaldev, xlab = "Potential minimizers of squared deviation",
+     ylab = "Squared Deviation", main = "Squared deviations from the mean versus other numbers")
> lines(xvalues, totaldev)
```

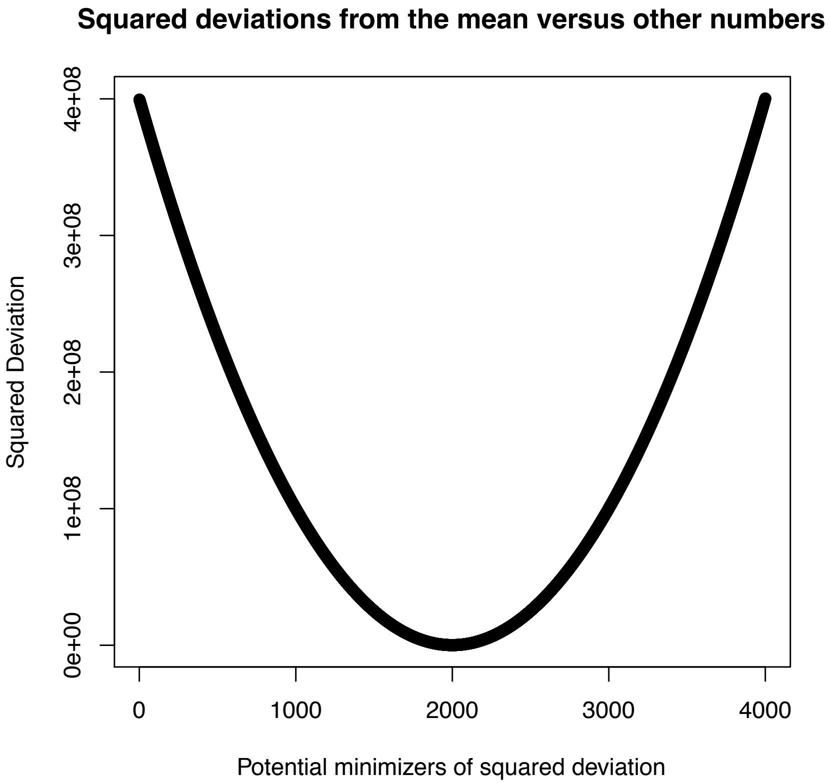


Figure 2.7: The mean minimizes variance: deviations from the mean are smaller than from any other number.

```
> binomialprobability(600, 2/3, 400)
```

```
[1] 0.03453262
```

Thus, as the sample size goes up, the probability of getting exactly the number corresponding to the theoretical probability 2/3 goes *down*. However—and this is a critical point—since the values in the large sample case are usually much closer to the mean than in the small sample case, the probability that the sample of white balls will be *close to* (not *exactly equal to*) the population parameter (here, the number of white balls we expect to draw) will be greater if you take a bigger sample.

To see this, consider an alternative (simpler) scenario where we have 12,000 red and white balls, and exactly half are red (p=0.5). If we take a sample of 40 balls, we can calculate the probability of getting 1 . . . 39, 40 white balls:

```
> numballs <- 40
> p <- 0.5
> probs <- rep(NA, numballs)
> for (k in 1:numballs) {
+     probs[k] <- binomialprobability(numballs, p, k)
+ }
```

Note as an aside that an alternative way to do this is:

```
> probs2 <- rep(NA, 40)
> for (i in 1:40) {
+     probs2[i] <- dbinom(i, 40, 0.5)
+ }
```

The variable `probs` now contains a list of probabilities:

```
> head(probs)
```

```
[1] 3.637979e-11 7.094059e-10 8.985808e-09 8.311872e-08 5.984548e-07
[6] 3.490986e-06
```

Note that the probability of getting exactly 20 white balls is 0.125370687619579. This is kind of low. We could relax our stringent requirement that the sample reflect the *exact* mean of the population, and ask about the probability of a *range* around the sample mean containing the population mean.

What's the probability of getting 19, or 20, or 21 white balls in a sample of 40 (Margin of error 1)? Or of getting 18, 19, 20, 21, or 22 white balls (Margin of error 2)? (Note on R: the parentheses around the commands below is just a way to get R to print out the result; if we didn't have the parentheses R would store the result in the variables *withinone* and *withintwo*).

```
> (withinone <- sum(probs[19:21]))
```

```
[1] 0.364172
```

```
> (withintwo <- sum(probs[18:22]))
```

```
[1] 0.5704095
```

Let's just compute the probabilities for *all* the margins of error.

```
> mean.index <- 20
> intervals <- rep(NA, 19)
> for (i in 1:19) {
+     indices <- seq(mean.index - i, mean.index + i, by = 1)
+     range <- probs[indices]
+     intervals[i] <- sum(range)
+ }
> conf.intervals <- data.frame(margin = rep(1:19), probability = intervals)
> conf.intervals40 <- conf.intervals
> print(head(conf.intervals))
```

```
  margin probability
1      1   0.3641720
2      2   0.5704095
3      3   0.7318127
4      4   0.8461401
5      5   0.9193095
6      6   0.9615227
```

The main point here is that when we relax the margin of error to be plus or minus six around the precise expected mean number of white balls (20), the probability is now approximately 95%. Let's visualize this (Figure 2.8).

The gray line in Figure 2.8 marks the margin of error (about 6), which corresponds to 95% probability. When we take a sample of 40 balls, we can be 95% sure that the true expected mean number of white balls (which we know to be 20 in this case) lies within the range of plus/minus 6 about the mean.

What would happen if the sample size were increased from 40 to 400? Our expected mean number of white balls would now be 200. Now we can compare the situation with 40 versus 400 drops when we allow the margin of error to encompass an area such that the probability of the population mean lying within that margin is 0.95.

Exercise 3 — Confidence intervals for a sample size of 400

Modify the code given above to calculate the 95% confidence interval for a draw of 400 balls, where the probability of drawing a white ball is 0.5. How many margins of error do we need to have to get a 95% confidence interval? We give the solution below, but spend a few minutes working this out before reading further.

For a sample of 40, between 5 and 6 margins of error about the sample mean we can be 95% certain that the population mean lies within this margin—if the sample has a binomial distribution. For a sample of 400, between 19 and 20 margins of error about the sample mean we can be 95% certain that the population mean lies within this margin—if the sample has a binomial distribution. Figure 2.9 shows the result.

```
> numballs <- 400
> p <- 0.5
> probs <- rep(NA, numballs)
> for (k in 1:numballs) {
+     currentk <- binomialprobability(numballs, p, k)
+     probs[k] <- currentk
+ }
> mean.index <- 200
> intervals <- rep(NA, 199)
> plot(conf.intervals$margin, conf.intervals$probability, type = "b",
+     xlab = "Margins", ylab = "Probability", main = "Sample size 40")
> segments(0, 0.95, 5.7, 0.95, col = "gray")
> segments(5.7, 0, 5.7, 0.95, col = "gray")
```

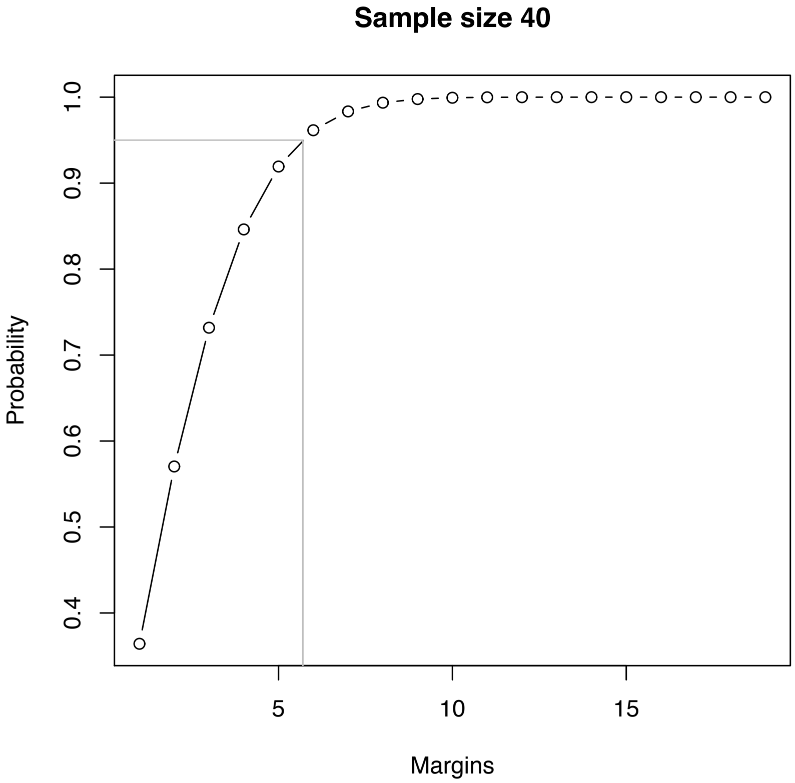


Figure 2.8: The probability of getting 20 plus/minus some n white balls from a random sample of 40, where n is the margin of error we allow (ranging from 1 to 20).

```
> for (i in 1:199) {
+     indices <- seq(mean.index - i, mean.index + i, by = 1)
+     range <- probs[indices]
+     intervals[i] <- sum(range)
+ }
> conf.intervals <- data.frame(margin = rep(1:199), probability = intervals)
> conf.intervals400 <- conf.intervals
> print(head(conf.intervals))

  margin probability
1      1   0.1192112
2      2   0.1973747
3      3   0.2736131
4      4   0.3472354
5      5   0.4176255
6      6   0.4842569
```

In the above examples, we essentially used the same R code twice, with just a few changes. In such situations it makes sense to write a function that can do the same thing, but with different settings (here, different sample sizes). Let's write such a function.

```
> compute_margins <- function(numballs, p) {
+     probs <- rep(NA, numballs)
+     for (k in 1:numballs) {
+         currentk <- binomialprobability(numballs, p, k)
+         probs[k] <- currentk
+     }
+     mean.index <- numballs * p
+     max.margin <- numballs * p - 1
+     intervals <- rep(NA, max.margin)
+     for (i in 1:max.margin) {
+         indices <- seq(mean.index - i, mean.index + i, by = 1)
+         range <- probs[indices]
+         intervals[i] <- sum(range)
+     }
+     conf.intervals <- data.frame(margin = rep(1:max.margin),
+         probability = intervals)
+     return(conf.intervals)
+ }
```

Before reading on, the reader should confirm whether the above function can be used to compute the probabilities for all the margins of error for *any* sample size and *any* probability (not just 0.5).

We just established that when the sample size is 40, we need to have 5 margins of error to obtain a region within which we are 95% certain that the population mean lies. For a sample size of 400, we need 19 margins of error.

Interestingly, we can now, in the same plot, compare the probability distribution of the margins for both samples. However, in order to compare them, we have to normalize the margins so that their range is constant in both cases (currently, in the 40 sample case the margins range from 1 to 19, and in the 400 sample case they range from 1 to 199). This normalization can be done by converting them to proportions; for example, in the 40 sample case, we simply treat the margin plus/minus 1 (19 and 21) as 19/40 and 21/40 respectively; for the 400 sample case, we treat the

```
> plot(conf.intervals$margin, conf.intervals$probability, type = "b",
+     xlab = "Margins", ylab = "Probability", main = "Sample size 40")
> segments(-6, 0.95, 19, 0.95, col = "gray")
> segments(19, 0, 19, 0.95, col = "gray")
```

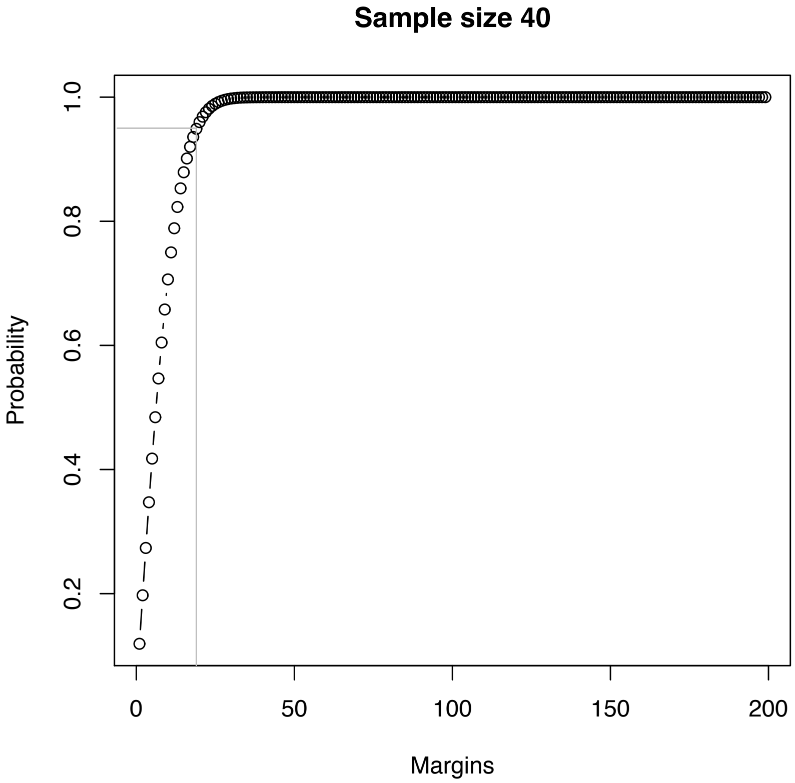


Figure 2.9: The probability of getting 200 plus/minus some n white balls from a random sample of 40, where n is the margin of error we allow (ranging from 1 to 199).

margin plus/minus 1 (199 and 201) as 199/400 and 201/400 respectively. When we do that, we get Figure 2.10. (We first define a function called plotcis to get the confidence intervals plotted).

```
> plotcis <- function(numballs, p, color = "black", margin, maintitle,
+     interval = TRUE) {
+     probs <- rep(NA, numballs + 1)
+     for (k in 0:numballs) {
+         currentk <- binomialprobability(numballs, p, k)
+         probs[k + 1] <- currentk
+     }
+     proportions <- 0:numballs/numballs
+     plot(proportions, probs, type = "l", col = "black", xlab = "Proportions",
+         ylab = "Probability", main = maintitle)
+     if (interval == TRUE) {
+         segments(proportions[(numballs/2 + 1) - margin], -0.5,
+             proportions[(numballs/2 + 1) - margin], 0.06, col = color,
+             lty = 1, lwd = 2)
+         segments(proportions[(numballs/2 + 1) + margin], -0.5,
+             proportions[(numballs/2 + 1) + margin], 0.06, col = color,
+             lty = 1, lwd = 2)
+     }
+ }
```

There are two important insights to take away from Figure 2.10. As sample size increases from 40 to 400, we get proportionally tighter 95% probability regions. The second (which is completely non-obvious at the moment, but will become clear in the coming chapters), is that, regardless of sample size, the 95% probability region corresponds to approximately 2 times the *standard deviation of the distribution of white-ball draws*.

Now, if we had some way to calculate this standard deviation *from a single sample*, we would not have to repeatedly sample from the collection of balls to build the distributions in Figure 2.10. It turns out that there is a way to obtain this information. This is discussed in the next chapter. But before we proceed further, we would like to introduce a distribution that is very similar to the binomial distribution we have seen in this chapter.

2.5.1 Applying the binomial theorem: Some useful R functions

As mentioned above, when we want to compute the probability of getting 0 to 20 right stone hits when we observe 40 raindrops, we can do this using dbinom:

```
> sums <- rep(NA, 21)
> for (i in 0:20) {
+     sums[i + 1] <- dbinom(i, 40, 0.5)
+ }
> sum(sums)

[1] 0.5626853
```

An even easier way to do this in R is:

```
> sum(dbinom(0:20, 40, 0.5))

[1] 0.5626853
```

```
> op <- par(mfrow = c(1, 2), pty = "s")
> plotcis(40, 0.5, margin = 5, maintitle = "Sample size 40")
> plotcis(400, 0.5, margin = 19, maintitle = "Sample size 400")
```

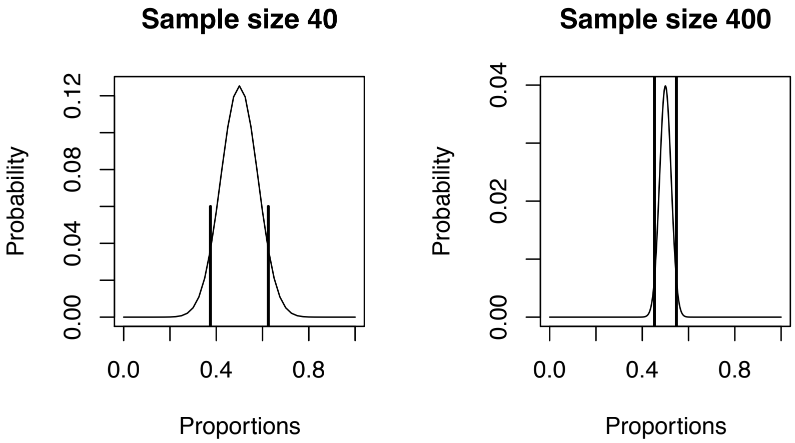


Figure 2.10: The 95% probability ranges in the 40 and 400 sample case with the margins of error normalized.

And yet another way is to say:

```
> pbinom(20, 40, 0.5)
[1] 0.5626853
```

Thus, there is a family of functions for the binomial distribution that we can use to do very useful things:

- rbinom: the **r**andom number generation function
- dbinom: The probability **d**ensity function
- pbinom: The cumulative distribution function (the proportion of values which have a value x or lower)

2.6 The binomial versus the normal distribution

It happens to be the case that the distributions shown in Figure 2.10 are remarkably similar to the distribution defined by this somewhat intimidating-looking function:

$$f(x) = \frac{1}{(\sigma\sqrt{2\pi})} E^{-((x-\mu)^2/(2\sigma^2))} \tag{2.10}$$

Given a range of values for x, and μ, and σ, we could plot this function. Let's plot this function and compare it with the binomial distribution. First we have to define the function:

```
> newfunction <- function(x, mu, sigma) {
+     1/(sqrt(2 * pi) * sigma) * exp(1)^(-((x - mu)^2/(2 * sigma^2)))
+ }
```

The binomial distribution and the normal distribution function have pretty similar shapes. Look at the R help for `dnorm` and `rnorm` for using the normal distribution in R. (Just type ?`dnorm` at the command prompt.)

One important difference between the normal and binomial distributions is that the former refers to continuous dependent variables, whereas the latter refers to a binomial variable.

With binomial distributions we already know how to find the probability that the population mean is within a given margin of error (sum up the probabilities of values around the mean for a given margin of error). This summation procedure is the same as computing the area under the curve so, in the case of the normal distribution, we can do an integration (which is just the equivalent of "summation" of continuous values). Try this:

```
> integrate(dnorm, -1, 1)
0.6826895 with absolute error < 7.6e-15
> integrate(dnorm, -1.96, 1.96)
0.9500042 with absolute error < 1.0e-11
> integrate(dnorm, -2, 2)
0.9544997 with absolute error < 1.8e-11
```

The normal distribution is useful when we are interested in continuous data (not binary responses 1, 0). An example would be reaction time or reading time data.

In the next chapter we are going to use the normal distribution to understand the concept of a sampling distribution of the sample means. This is the key to our main problem: how to estimate the 95% probability region we calculated above with a single sample, as opposed to repeatedly sampling from the population.

```
> plotcis(40, 0.5, 40, margin = 20, maintitle = "Comparing the binomial and normal distributions",
+     interval = FALSE)
> lines(c(1:40)/40, newfunction(c(1:40), 20, 3), col = "black",
+     lty = 2)
```

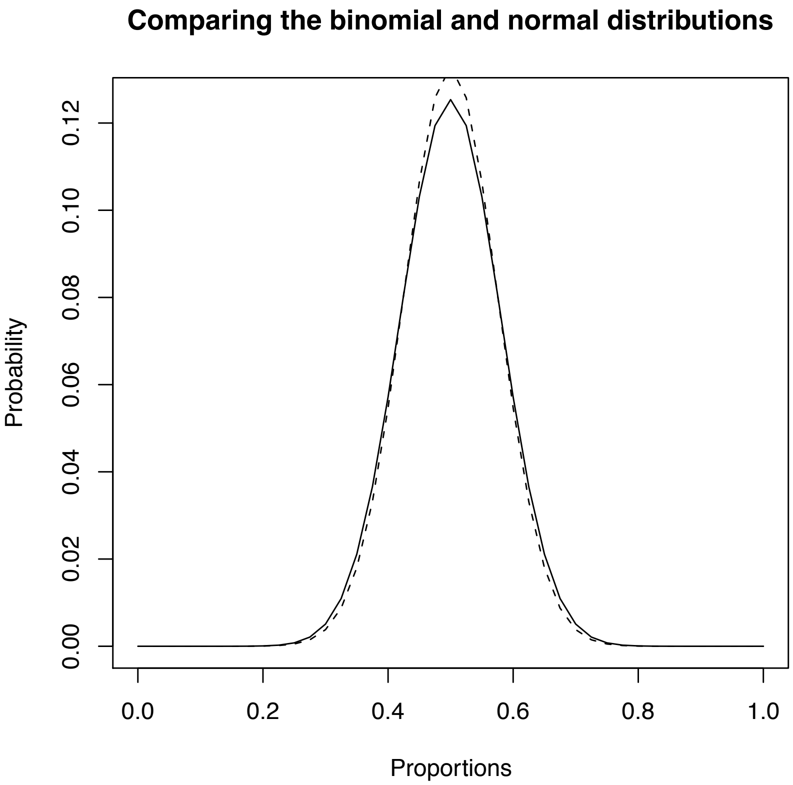


Figure 2.11: Comparing the binomial vs normal distributions.

Chapter 3

The sampling distribution of the sample mean

Suppose that we have a population of people, and that we know the age of each individual; let us assume also that distribution of the ages is approximately normal (i.e., the shape of the age distribution resembles the normal distribution we saw in the preceding chapter). Finally, let us also suppose that we know that mean age of the population is 60 and the population SD is 8.

Now suppose that we repeatedly sample from this population: we take samples of 40 a total of 1000 times, and calculate the mean each time we take a sample. After taking 1000 samples, we have 1000 means; if we plot the distribution of these means, we have the sampling distribution of the sample means.

```
> means <- rep(NA, 1000)
> for (i in c(1:1000)) {
+     sample40 <- rnorm(40, mean = 60, sd = 8)
+     means[i] <- mean(sample40)
+ }

[1] 60.00927

[1] 1.30528
```

If we plot this distribution of means, we find that it is roughly normal. We can characterize this distribution of means visually, as done in Figure 3.1 below, or in terms of the mean and standard deviation of the distribution (i.e., in terms of the means of the means, and the standard deviation of the means). The mean value in the above simulation is 60.01 and the standard deviation of the distribution of means is 1.3053.

Consider now the situation where our sample size is 100. Note that the mean and standard deviation of the population scores is the same as above.

```
> samplesize <- 100
> means <- rep(NA, 1000)
> for (i in c(1:1000)) {
+     sample100 <- rnorm(samplesize, mean = 60, sd = 8)
+     means[i] <- mean(sample100)
+ }
```

```
> hist(means)
```

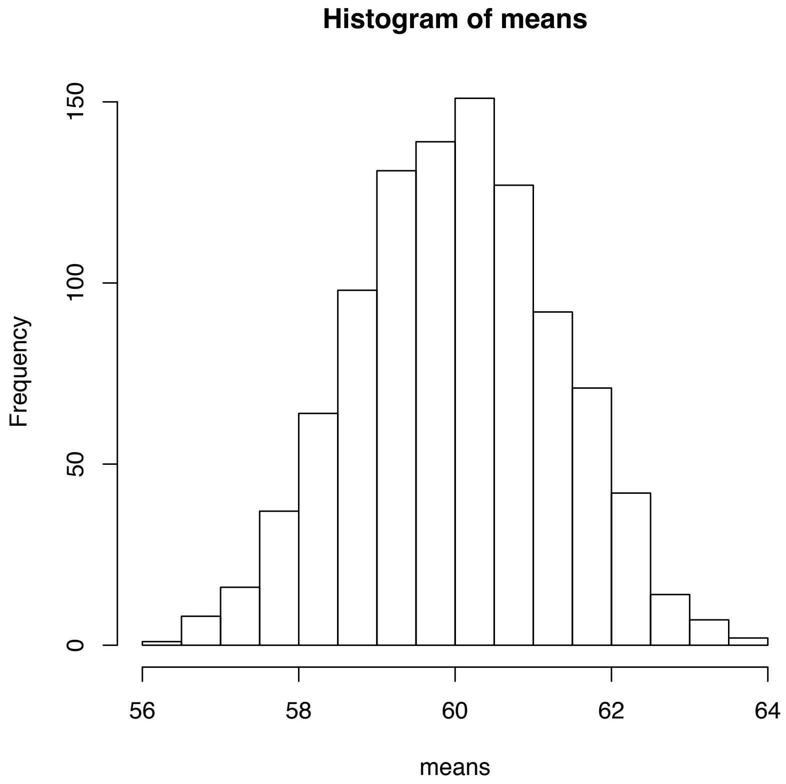


Figure 3.1: The sampling distribution of the sample means with 1000 samples of size 40.

In this simulation run, the mean of the means is 60 and the standard deviation of the distribution of means is 0.8099.

```
[1] 59.99701
```

```
[1] 0.8098542
```

The above simulations show us several things. First, the standard deviation of the distribution of means gets smaller as we increase sample size. When the sample size is 40, the standard deviation is 1.305; when it is 100, the standard deviation is 0.8099. Second, as the sample size is increased, the mean of the means comes closer and closer to the *population* mean. A third point (which is not obvious at the moment) is that there is a lawful relationship between the standard deviation σ of the population and the standard deviation of the distribution of means, which we will call $\sigma_{\bar{x}}$.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{3.1}$$

Here, n is the sample size. It is possible to derive equation 3.1 from first principles. We do this in the appendix.

For now, let's take this equation on trust and use it to compute $\sigma_{\bar{x}}$ by using the population standard deviation (which we know). Let's do this for a sample of size 40 and another of size 100:

```
> 8/sqrt(40)
```

```
[1] 1.264911
```

```
> 8/sqrt(100)
> hist(means)
```

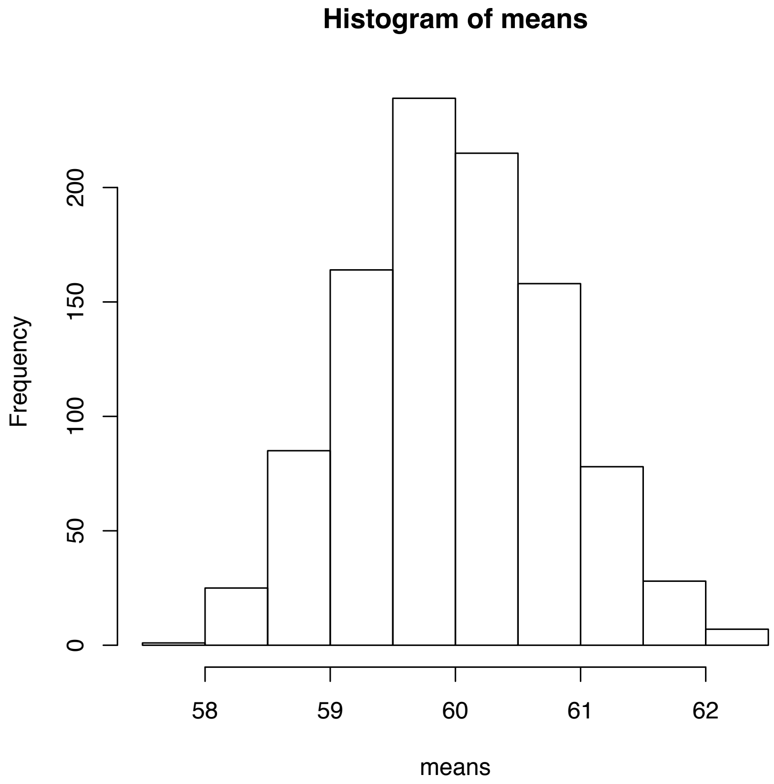


Figure 3.2: The sampling distribution of the sample means with samples of size 100.

```
[1] 0.8
```

The above calculation shows that $\sigma_{\bar{x}}$ gets smaller and smaller as we increase sample size.

3.1 The Central Limit Theorem

We've seen in the previous chapter that the distribution of a sample proportion is normally distributed. Now we see that the sampling distribution of the sample mean is too—and in the above example it was also drawn from a normally distributed population which is quite similar to the binomial distribution (see Section 2.6). It turns out that the sampling distribution of the sample means will be normal even if the population is not normally distributed, as long as the sample size is large enough. This is known as the Central Limit Theorem, and is so important that we will say it twice:

> Provided the sample size is large enough, the sampling distribution of the sample mean will be close to normal *irrespective of what the population's distribution looks like*.

Let's check this with a simulation of a population which we *know* is non-normally distributed. Let us assume that the population distribution is exponential, not normal.

Now let us plot the sampling distribution of the sample mean. We take 1000 samples of size 100 each from this exponentially distributed population. The distribution of the means is again normal!

```
> population <- rexp(1000)
> hist(population)
```

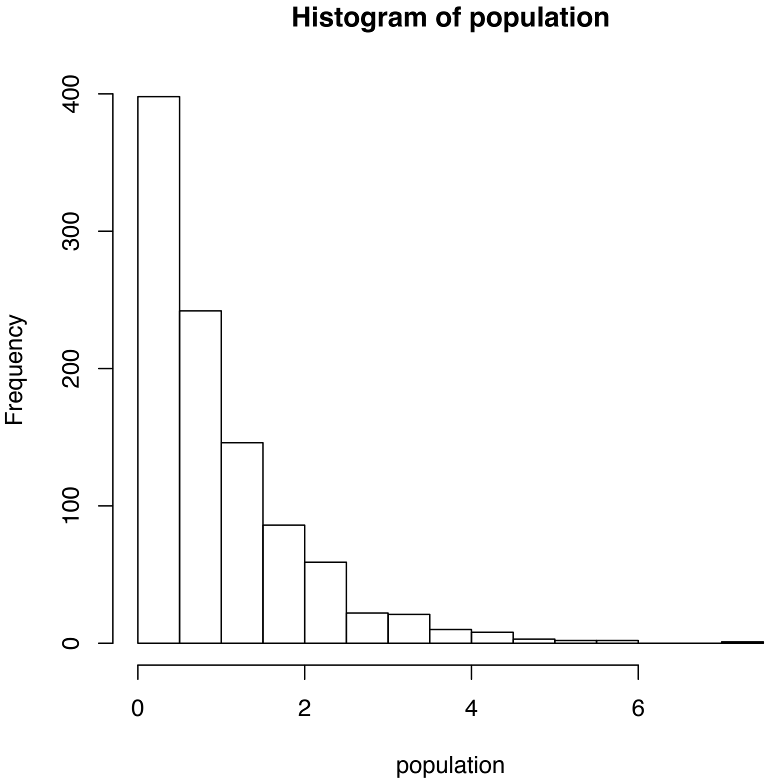


Figure 3.3: A set of exponentially distributed population scores.

```
> means <- rep(NA, 1000)
> for (i in c(1:1000)) {
+     samp100 <- rexp(100)
+     means[i] <- mean(samp100)
+ }
> hist(means)
```

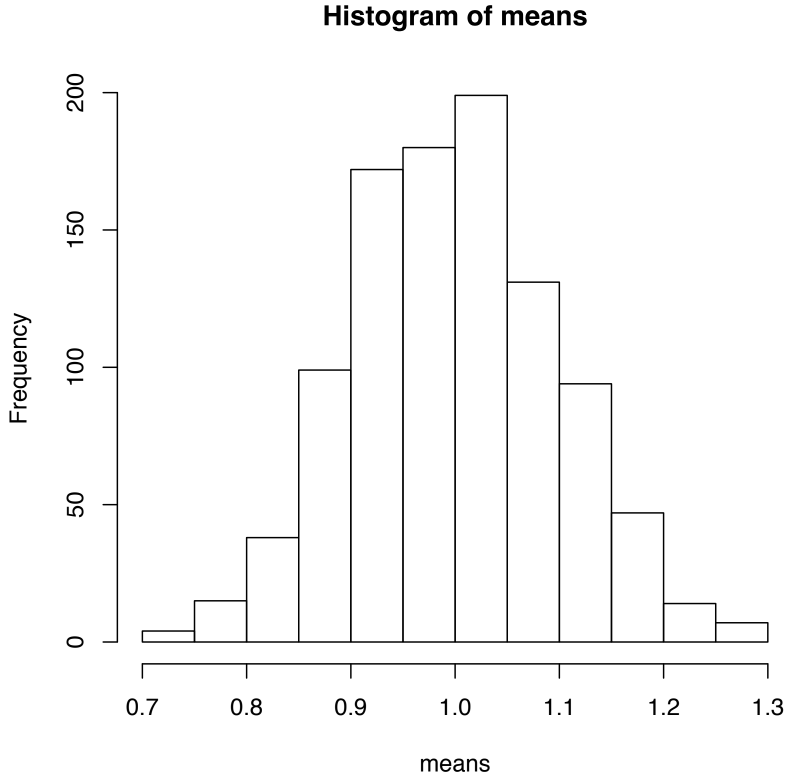


Figure 3.4: The exponential distribution's sampling distribution of the sample means.

To summarize:

- The sampling distributions of various statistics (the sampling distribution of the sample means or sample proportion or sample count) are nearly normal or gaussian. The gaussian distribution implies that a sample statistic that is close to the mean has a higher probability than one that's far away.

- The mean of the sampling distribution of the sample mean is (in the limit) the same as the population mean.

- It follows from the above two facts that the mean of a sample is more likely to be close to the population mean than not.

3.2 The SD of the population and of the sampling distribution of the sample means

We saw earlier that the standard deviation of the sampling distribution *of the sample mean*, $\sigma_{\bar{x}}$ gets smaller as we increase sample size. When the sample has size 40, this standard deviation is 7.302; when it is 100, this standard deviation is 7.8363.

Let's study the relationship between $\sigma_{\bar{x}}$ and σ. Recall that population mean = 60, $\sigma = 8$. The equation below summarizes the relationship; it shouldn't surprise you, since we just saw it above (also see the appendix):

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{3.2}$$

But note also that the tighter the distribution, the greater the probability that the estimate of the mean based on a single sample is close to the population mean. So the $\sigma_{\bar{x}}$ is an indicator of how good our estimate of the population mean is.

3.3 The 95% confidence interval

Let's take a sample of 11 heights from a normally distributed population with known mean height 60 and SD (σ) 4 (inches).

```
> sample11 <- rnorm(11, mean = 60, sd = 4)
```

Let us estimate a population mean from the sample using the sample mean, and compute the $\sigma_{\bar{x}}$. Recall that we know the precise population standard deviation so we can get a precise value for $\sigma_{\bar{x}}$.

```
> (estimated.mean <- mean(sample11))

[1] 58.51356

> popSD <- 4
> sample.size <- length(sample11)
> (sigma.mu <- popSD/sqrt(sample.size))

[1] 1.206045
```

We know from the Central Limit Theorem that the sampling distribution of the sample mean is roughly normal, and we know that our $\sigma_{\bar{x}} = 1.2$. Recall from Chapter 2 that the probability that the population mean is within 2 $\sigma_{\bar{x}}$ of the sample mean is a bit over 0.95. Let's calculate this range:

$$\bar{x} \pm 2 \times \sigma_{\bar{x}} = 59 \pm 2 \times 1.206 \tag{3.3}$$

The .95 probability region is 56.1 and 60.9. The number 0.95 is a probability from the point of view of the sampling distribution, and a confidence level from the point of view of parameter estimation – in the latter case it's conventionally expressed as a percentage and is called the 95% confidence interval.

Suppose now that sample size was four times bigger (44). Let's calculate the sample means, estimated standard deviation of the sampling distribution of the sample means, and from this information, plus the sample size, we get the 95% confidence interval.

```
> sample44 <- rnorm(44, mean = 60, sd = 4)
> estimated.mean <- mean(sample44)
> sample.size <- length(sample44)
> (sigma.mu <- 4/sqrt(sample.size))
```

```
[1] 0.6030227
```

Now we get a much tighter 95% confidence interval:

$$\bar{x} \pm 2 \times \sigma_{\bar{x}} = 59 \pm 2 \times 0.603 \tag{3.4}$$

The interval now is 57.7 and 60.1; it is smaller than the one we got for the smaller sampler size.

3.4 Realistic statistical inference

Until now we have been sampling from a population whose mean and standard deviation we know. However, normally we don't know the population parameters. In other words, although we know that:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{3.5}$$

when we take samples in real life, we almost never know σ. But we can just *estimate* σ using the standard deviation s of the sample. However, now we can only get an *estimate* of $\sigma_{\bar{x}}$. This is called the Standard Error *of the (sample) mean* or *of the statistic*:

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{3.6}$$

Pay careful attention to the distinction between s (an estimate of σ) and $SE_{\bar{x}}$ (as estimate of $\sigma_{\bar{x}}$). In particular, note that the Standard Error is an estimate of the standard deviation of the sampling distribution of the sample mean. In other words, it is a standard deviation, but not of the sample itself—that is called s here. Rather, Standard Error is the standard deviation of the vector of sample means you would get if you were to sample multiple times from a population, as we have been doing.

One question that should arise in your mind is: can we safely assume that s is a reliable estimate of σ? It turns out that the answer is yes. Let's explore this issue next.

3.5 s is an unbiased estimator of σ

Earlier in this chapter we repeatedly sampled from a population of people with mean age 60 and standard deviation 8; then we plotted the distribution of sample means that resulted from the repeated samples. One thing we noticed was that any one sample means was more likely to be close to the population mean (this follows from the normal distribution of the means resulting from the repeated sampling).

Let us repeat this experiment, but this time we plot the distribution of the standard deviations. What we will find is that any one sample's standard deviation s is more likely than not to be close to the population standard deviation σ. This is because the distribution of the standard deviations of the repeated samples also has a normal distribution.

```
> sample.sd <- rep(NA, 1000)
> for (i in c(1:1000)) {
+     sample40 <- rnorm(40, mean = 60, sd = 8)
+     sample.sd[i] <- sd(sample40)
+ }
> hist(sample.sd)
```

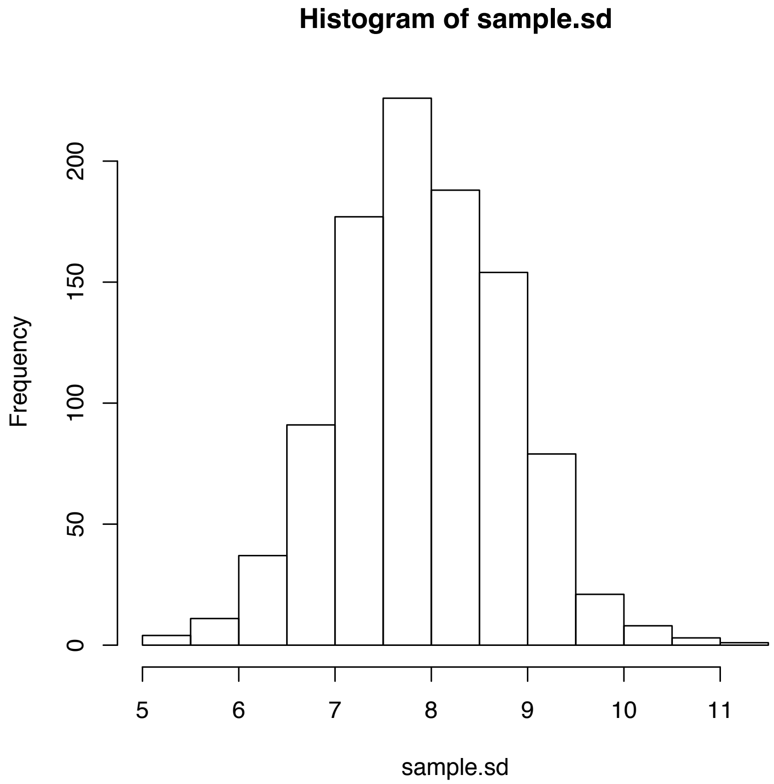


Figure 3.5: The distribution of the standard deviations of the samples, sample size 40. The population is normally distributed

What this tells us is that if we use s as an estimator of σ we're more likely to get close to the right value than not. This is true even if the population is not normally distributed. Let's check this (Figure 3.6).

```
> sample.sd <- rep(NA, 1000)
> for (i in c(1:1000)) {
+     sample40 <- rexp(40)
+     sample.sd[i] <- sd(sample40)
+ }
> hist(sample.sd)
```

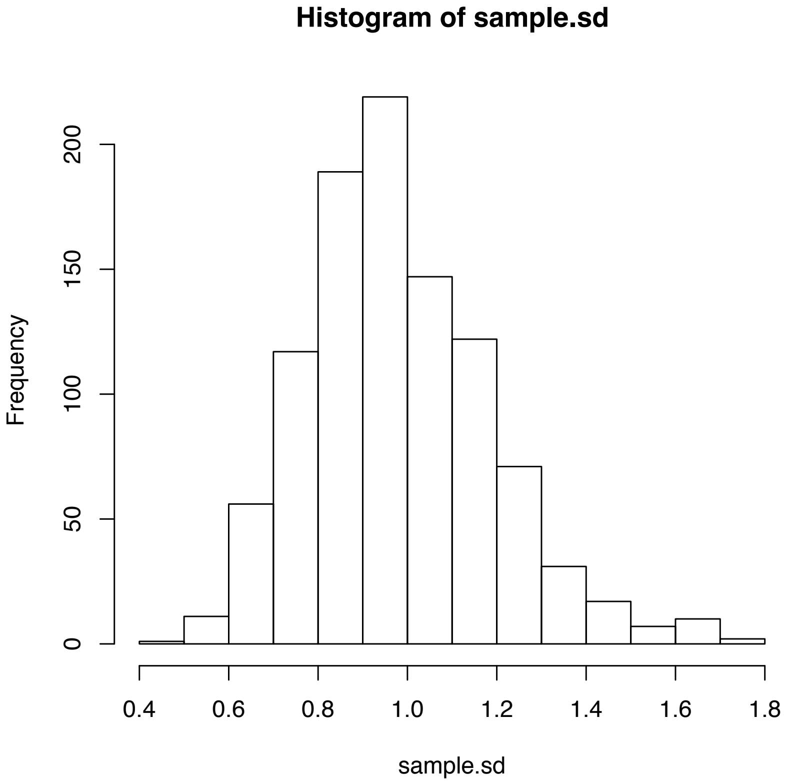


Figure 3.6: The distribution of the standard deviations of the samples, sample size 40. This time the population scores sampled from is exponentially distributed.

We are now at the point that we can safely use the sample standard deviation s as an estimate of the unknown population standard deviation σ, and this allows us to estimate $\sigma_{\bar{x}}$; we call this estimate the Standard Error and write it $SE_{\bar{x}}$.

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{3.7}$$

One problem now is that, especially for smaller sample sizes, the sampling distribution of the sample mean cannot be modeled by the normal distribution defined by $\mathcal{N}(\mu, SE_{\bar{x}})$: the $SE_{\bar{x}}$ is just an estimate for $\sigma_{\bar{x}}$. To model our sample, we need a distribution shape which has greater uncertainty built into it than the normal distribution. This is the motivation for using the t-distribution rather than the normal distribution.

3.6 The t-distribution

This distribution is defined by the degrees of freedom (sample size minus 1); it has more of its probability in its tails (=greater uncertainty), but approximates to the normal distribution with increasing degrees of freedom. The standard deviation also approaches 1; and with infinite degrees of freedom, it *is* the normal distribution. But with about 15 degrees of freedom, it's already very close to normal; see Figure 3.7.

What we have available to us to work with now: We have a new estimate s of the population SD, and a new estimate $SE_{\bar{x}}$ of the SD of the sample:

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{3.8}$$

We have a more spread out distribution than the normal, the t-distribution, and it's defined by the degrees of freedom (roughly, sample size). We are now ready to do some statistical inference.

3.7 The t-test

We know how to estimate the standard deviation σ of the population using the sample standard deviation s:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.9}$$

We also know how to compute the SE of the sample means:

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{3.10}$$

We can now ask: how many SE's do we need to go to the left and right of the sample mean to be 95% sure that the population mean lies in that range? You could look up a table that tells you, for $n-1$ degrees of freedom, how many SE's you need to go around the sample mean to get a 95% CI. Or you could ask R. First we take a sample of size 11 from a population with mean 60 and standard deviation 4.

```
> sample <- rnorm(11, mean = 60, sd = 4)
```

Using this sample, we can ask what the 95% confidence interval is:

```
> print(t.test(sample)$conf.int)

[1] 56.58449 63.61205
attr(,"conf.level")
[1] 0.95
```

Sure enough, if our sample size had been larger, our confidence interval would be narrower:

```
> sample <- rnorm(100, mean = 60, sd = 4)
> print(t.test(sample)$conf.int)

[1] 59.36266 60.87575
attr(,"conf.level")
[1] 0.95
```

```
> range <- seq(-4, 4, 0.01)
> multiplot(2, 3)
> for (i in c(2, 5, 10, 15, 20, 50)) {
+     plot(range, dnorm(range), lty = 1, col = "gray")
+     lines(range, dt(range, df = i), lty = 2)
+     mtext(paste("df=", i), cex = 1.2)
+ }
```

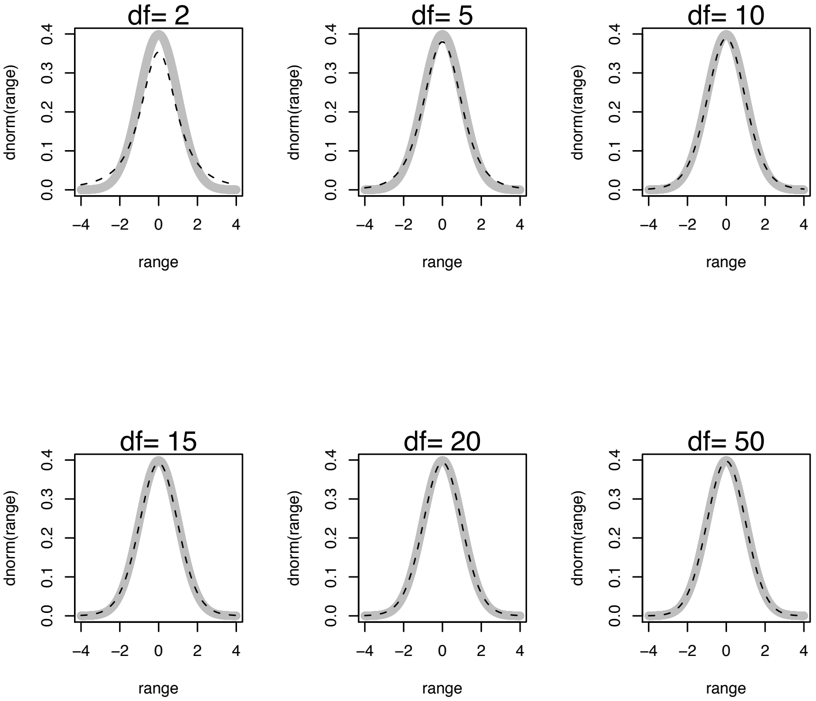


Figure 3.7: A comparison between the normal (solid gray line) and t-distribution (broken black line) for different degrees of freedom.

3.8 Some observations on confidence intervals

There are some subtleties associated with confidence intervals that are often not brought up in elementary discussions, simply because the issues are just too daunting to tackle. However, we will use simulations to unpack some of these subtleties. We hope that the reader will see that the issues are in reality quite simple.

The first critical point to understand is the meaning of the confidence interval. We have been saying up till now that the 95% confidence interval tells you the range within which we are 95% sure that the population mean lies. However, one critical point to notice is that the range defined by the confidence interval will vary with each sample even if the sample size is kept constant. The reason is that the sample mean will vary each time, and the standard deviation will vary too. We can check this fact quite easily.

First we define a function for computing 95% CIs:

```
> se <- function(x) {
+     y <- x[!is.na(x)]
+     sqrt(var(as.vector(y))/length(y))
+ }
> ci <- function(scores) {
+     m <- mean(scores, na.rm = TRUE)
+     stderr <- se(scores)
+     len <- length(scores)
+     upper <- m + qt(0.975, df = len - 1) * stderr
+     lower <- m + qt(0.025, df = len - 1) * stderr
+     return(data.frame(lower = lower, upper = upper))
+ }
```

Next, we simulate 100 samples, computing the confidence interval each time.

```
> lower <- rep(NA, 100)
> upper <- rep(NA, 100)
> for (i in 1:100) {
+     sam <- rnorm(100, mean = 60, sd = 4)
+     lower[i] <- ci(sam)$lower
+     upper[i] <- ci(sam)$upper
+ }
> cis <- cbind(lower, upper)
> head(cis)
```

```
        lower    upper
[1,] 58.67948 60.32930
[2,] 59.93228 61.52773
[3,] 59.11583 60.77723
[4,] 59.49290 61.18646
[5,] 59.29711 60.91636
[6,] 58.82243 60.23183
```

Thus, any one particular confidence interval, based on a single sample, will tell you what the probability region is based on the particular sample mean and standard deviation you happen to get in that one sample. These particular mean and standard deviation values are likely to be close to the population mean and population standard deviation but they are ultimately just estimates of the true parameters (the population mean and standard deviation).

Importantly, because of the gaussian shapes of the distribution of sample means and sample standard deviations (see Figures 3.4 and 3.6), if we repeatedly sample from a population and compute the confidence intervals each time, in approximately 95% of the confidence intervals the population mean will lie within the ranges specified. In the other 5% or so of the cases, the confidence intervals will not contain the population mean.

This is what the confidence interval means: it's a statement about hypothetical repeated samples; more specificially, it's a statement about the probability of the hypothetical confidence intervals (that would be computed from the hypothetical repeated samples) containing the population mean.

Let's check the above statement. We can repeatedly build 95% CIs and determine whether the population mean lies within them. The claim is that it will lie within the CI 95% of the time.

```
> store <- rep(NA, 100)
> for (i in 1:100) {
+     sam <- rnorm(100, mean = 60, sd = 4)
+     if (ci(sam)$lower < 60 & ci(sam)$upper > 60) {
+         store[i] <- TRUE
+     }
+     else {
+         store[i] <- FALSE
+     }
+ }
> summary(store)
```

```
   Mode   FALSE    TRUE    NA's 
logical       7      93       0 
```

So that's true.

Note that when we compute a 95% confidence interval for a particular sample, we have only one interval. Strictly speaking, that particular interval does **not** mean that the probability that the population mean lies within that interval is 0.95. For that statement to be true, it would have to be the case that the population mean is a random variable, like the heads and tails in a coin are random variables, and 1 through 6 in a die are random variables.

The population mean is a single point value that cannot have a multitude of possible values and is therefore not a random variable. If we relax this assumption, that the population mean is a point value, and assume instead that "the" population mean is in reality a range of possible values (each value having different probabilities of being the population mean), then we could say that any one 95% confidence interval represents the range within with the population mean lies with probability 0.95. We recommend reading (Gelman & Hill, 2007) for more detail on this approach.

It's worth repeating the above point about confidence intervals. The meaning of the confidence interval depends crucially on hypothetical repeated samples: the confidence intervals computed in 95% of these repeated samples will contain the population mean. In essence, the confidence interval from a single sample is a random variable just like heads and tails in a coin toss, or the numbers 1 through 6 in a die, are random variables. Just as a fair coin has a 0.5 chance of yielding a heads, and just as a fair die has a 1/6 chance of landing a 1 or 2 etc., a confidence interval in repeated sampling has a 0.95 chance of containing the population mean.

Exercise 4 — Confidence intervals

1. Choose one answer in each:

   95% Confidence intervals describe:

   a. The range of individual scores

   b. Plausible values for the population mean
   c. Plausible values for the sample mean
   d. The range of scores within one standard deviation

3. 95% Confidence interval has a ?% chance of describing the sample mean:
   a. 95%
   b. 100%

4. For the same data, a 90% CI will be wider than a 95% CI.
   a. True
   b. False

3.9 Sample SD s, degrees of freedom, unbiased estimators

Let us reconsider the question: What's special about $n-1$ in the equation for standard deviation? Recall that the sample standard deviation s is just the average distance of the numbers in the list from the mean of the numbers.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.11}$$

We can explore the question of why $n-1$ by asking what would happen if we used n instead. As we see below, if we'd used n, s^2 (which is an estimate of the population variance σ^2) would be smaller. This smaller s^2 turns out to be a poorer estimate than when we use $n-1$. Let's verify this using simulations.

```
> newvar <- function(x) {
+     m <- rep(mean(x), length(x))
+     d <- (x - m)^2
+     return(sum(d)/length(x))
+ }
> correct <- rep(NA, 1000)
> incorrect <- rep(NA, 1000)
> for (i in c(1:1000)) {
+     sample10 <- rnorm(10, mean = 0, sd = 1)
+     correctvar <- var(sample10)
+     incorrectvar <- newvar(sample10)
+     correct[i] <- correctvar
+     incorrect[i] <- incorrectvar
+ }
```

```
> op <- par(mfrow = c(1, 2))
> hist(correct, main = paste("Mean ", round(mean(correct), digits = 2),
+     sep = " "))
> hist(incorrect, main = paste("Mean ", round(mean(incorrect),
+     digits = 2), sep = " "))
```

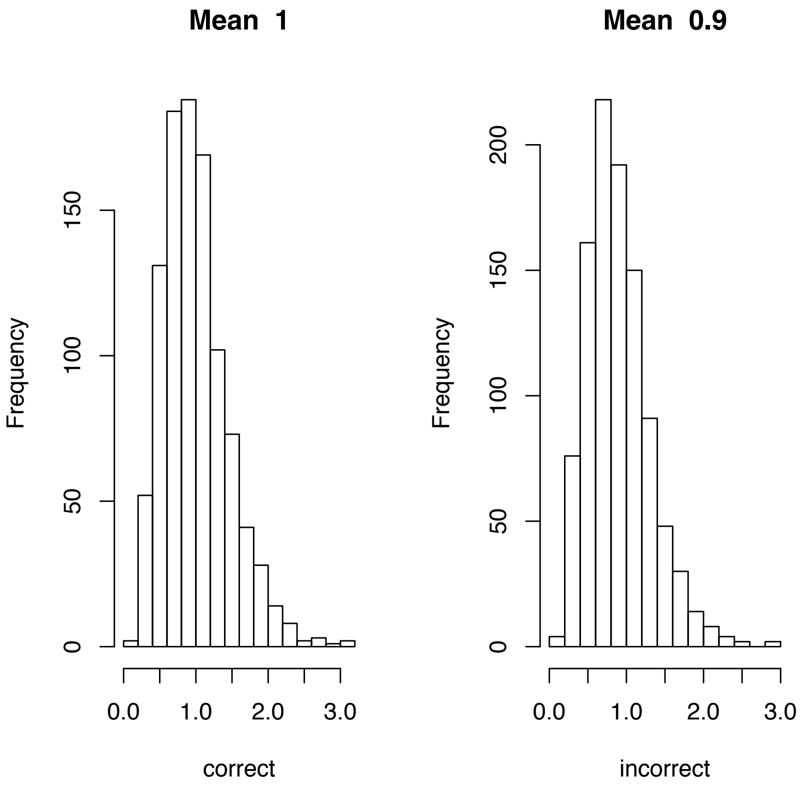


Figure 3.8: The consequence of taking $n - 1$ versus n in the denominator for calculating variance, sample size 10.

One interesting fact is that if the sample size is increased, from 10 to, say, 100, it ceases to matter whether we use n or $n-1$ in the denominator. Let's verify this. When we use a sample size of 100, the mean variance is approximately the same in both approaches to computing the variance.

```
> for (i in c(1:1000)) {
+     sample100 <- rnorm(100, mean = 0, sd = 1)
+     correctvar <- var(sample100)
+     incorrectvar <- newvar(sample100)
+     correct[i] <- correctvar
+     incorrect[i] <- incorrectvar
+ }
> op <- par(mfrow = c(1, 2))
> hist(correct, main = paste("Mean", round(mean(correct), digits = 2),
+     sep = " "))
> hist(incorrect, main = paste("Mean", round(mean(incorrect), digits = 2),
+     sep = " "))
```

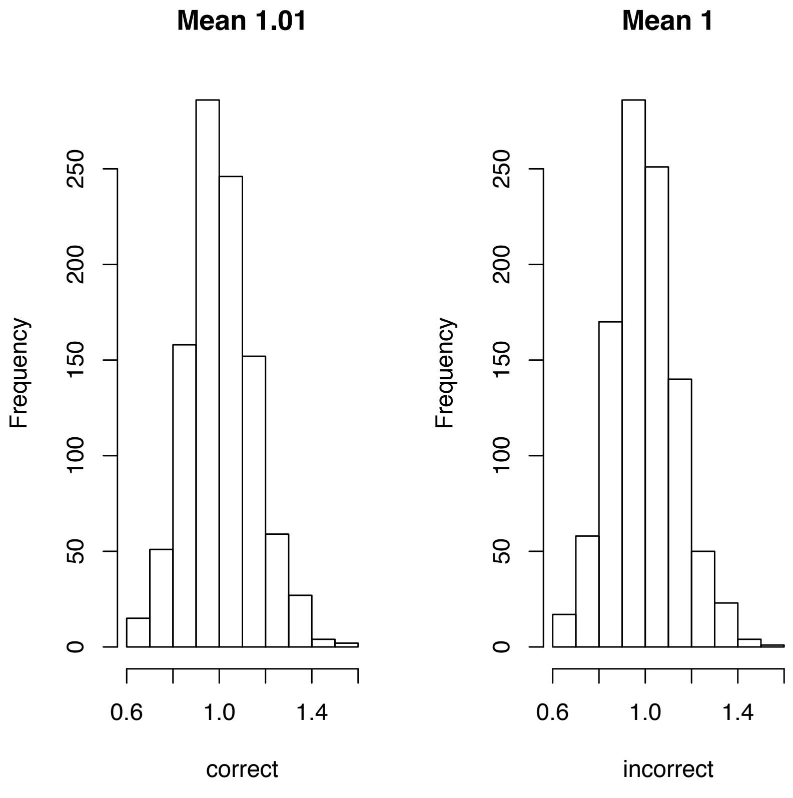


Figure 3.9: The consequence of taking $n-1$ versus n in the denominator for calculating variance, sample size 100.

In summary, using n gives a BIASED ESTIMATE of the true variance. The smaller the sample size, the greater this discrepancy between the unbiased and biased estimator.

3.10 Summary of the sampling process

It is useful at this point to summarize the terminology we have been using, and the logic of sampling. First, take a look at the concepts we have covered so far. We provide a list of the different concepts in a table below for easy reference.

this sample statistic	is an unbiased estimate of
sample mean $\bar{x}$	population mean μ
sample sd s	population sd σ
sample standard error $SE_{\bar{x}}$	population's $\sigma_{\bar{x}}$

where

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{3.12}$$

and

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{3.13}$$

- Statistical inference involves a single sample but assumes that some distribution of sample means exists out there in nature.

- The statistic (e.g., mean) in a random sample is more likely to be closer to the population parameter (the population mean) than not. This follows from the gaussian distribution of the sample means.

- The further away we get from the statistic, the lower the probability of this further-away value being the population parameter. This probability can be calculated precisely using elementary probability theory.

- In the limit the mean of the sampling distribution is equal to the population parameter.

- The standard deviation of the sampling distribution, $\sigma_{\bar{x}}$, is determined by sample size, and tells us how steeply the probability falls off from the center. If $\sigma_{\bar{x}}$ is small, then the fall in probability off the center is steep – random samples are more likely to be very close to the mean, samples are better indicators of the population parameters, and inference is more certain. If $\sigma_{\bar{x}}$ is large, then the fall in probability off the center is gradual – random samples far from the true mean are more likely, samples are not such good indicators of the population parameters, and inference is less certain.

We now turn to the main topic of this book: significance testing.

3.11 Significance tests

Recall the discussion of 95% confidence intervals: The sample gives us a mean $\bar{x}$. We compute $SE_{\bar{x}}$ (an estimate of $\sigma_{\bar{x}}$) using s (an estimate of σ) and sample size n. Then we calculate the range $\bar{x} \pm 2 \times SE_{\bar{x}}$. That's the 95% CI.

We don't know the population mean—if we did, why bother sampling? But suppose we had a HYPOTHESIS about the population mean having a certain value. If we have a HYPOTHESIS about the population mean, then we can measure the distance of our sample mean from the hypothesized population mean, and use the facts of the sampling distribution to determine the probability of occurrence of our actual sample mean, *assuming the hypothesis that the (hypothesized) population mean has a certain value.*

If the probability of the sample mean is high, then the evidence *might be consistent with* the null hypothesis. If the probability is low, this is evidence against the hypothesis. A SIGNIFICANCE TEST is a formal procedure for comparing observed data with a hypothesis whose truth we want to assess. The hypothesis is a statement about the parameters in a population, and the results of a test are expressed in terms of a probability that measures how well the data and the hypothesis agree.

3.12 The null hypothesis

The statement being tested in a significance test is called the NULL HYPOTHESIS, $\mathcal{H}_0$. Perhaps oddly, tests of significance are designed to assess the strength of the evidence *against* $\mathcal{H}_0$.

$\mathcal{H}_0$ is usually "chance". I.e., no effect, or no real difference between the sample mean and the population mean (any differences seen are just due to chance). Let's do some simulation to understand this better.

Suppose our hypothesis, based perhaps on previous research, is that the population mean is 70. Suppose also that we take a sample of 11 from a population whose mean is actually 60, not 70, and sd is 4:

```
> sample <- rnorm(11, mean = 60, sd = 4)
> sample.mean <- mean(sample)
> sigma.mu <- 4/sqrt(11)
```

Figure 3.10 shows what we expect our population distribution to look like if our hypothesis were *in fact* true. This hypothetical distribution is going to be our reference distribution on which we base all our inference.

Given this hypothetical sampling distribution, the probability of the sample mean 60 occurring is low (in fact the probability of some value like 69 occurring is also low, but not as low as 59).

Note that our goal is to make a decision: reject the null hypothesis, or fail to reject the null hypothesis. We can define a threshold to make this decision: Since we know that, on repeated sampling, in 95% of the samples the observed mean would fall within two SEs, any value further out than this range will occur only 5% of the time. So let's set the threshold to 95%: if an observed sample mean is more than about 2 SEs away from the hypothesized mean (in a distribution created using the hypothesized mean), we can confidently reject the hypothesis.

In other words, we want to know: how many SEs away is our sample mean from the hypothesized mean? The distance from the observed value $\bar{x}$ to the hypothesized mean μ_0 is some number z times $\sigma_{\bar{x}}$. We want to know this number z.

$$\bar{x} - \mu_0 = z\sigma_{\bar{x}} \tag{3.14}$$

Solving for z:

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} \tag{3.15}$$

$$= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{3.16}$$

z is called the STANDARDIZED VALUE.

```
> range <- seq(55, 85, 0.01)
> plot(range, dnorm(range, mean = 70, sd = sigma.mu), type = "l",
+     ylab = "", col = "red", main = "The null hypothesis")
```

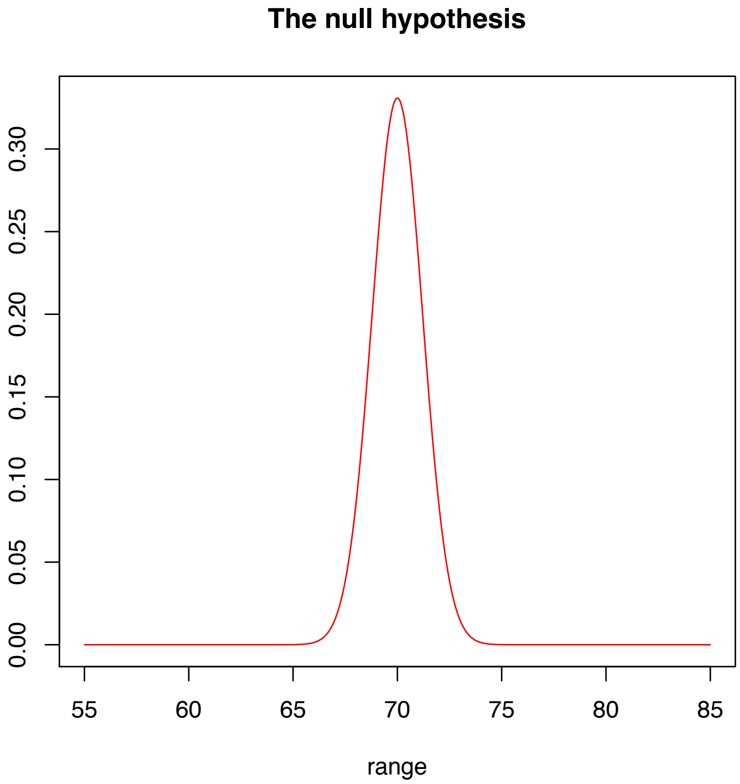


Figure 3.10: A sampling distribution with mean 70 and $\sigma_{\bar{x}} = 1.206$ computed from the sample.

3.13 z-scores

In our current simulation, $\bar{x} = 60, \mu_0 = 70, \sigma = 4, n = 11$.

So we get:

$$z = \frac{60 - 70}{4/\sqrt{11}} \tag{3.17}$$

$$= -8.496664 \tag{3.18}$$

Recall that $\bar{x}$ is a statistic; z is called a TEST STATISTIC. So we now have a way to express how far away the sample mean is, given a μ_0, and given σ.

3.14 P-values

Recall once again that reporting the exact probability of a particular value is not useful; such a value will always be low (see page 22). We can, however, usefully ask how much of the total probability lies beyond the observed value—this tells us how far out from the edge of "plausible" values the sample value is.

The P-VALUE of a test is the probability, computed assuming that $\mathcal{H}_0$ is true, that the test statistic would take a value as extreme as the one observed or more extreme than that actually observed.

How to determine this probability? We've seen how to do this—just sum (integrate) the area under the curve, going from our observed mean of 60 to the left edge of the curve—minus infinity. Note that our null hypothesis $\mathcal{H}_0$ was: the observed mean $\bar{x}$ is equal to the hypothesized mean μ_0. Rejecting the null hypothesis amounts to accepting the alternative hypothesis $\mathcal{H}_a$: $\bar{x} < \mu_0$ or $\mu_0 < \bar{x}$.

This means that as evidence for rejection of $\mathcal{H}_0$ we will use data beyond 2 SEs on **both** sides of the μ. So the above test is called a two-sided significance test. If the p-value is $\leq \alpha$ we say that the data are significant at level α.

3.14.1 The p-value is a conditional probability

Is it true that the smaller the p-value, the lower the probability that thet $\mathcal{H}_0$ is true? Note that the p-value is a conditional probability: it's the probability of obtaining a particular test statistic (like a t-score) as extreme or more extreme than the one observed, conditional on the assumption that the null hypothesis is true. From probability theory we know that the conditional probability of B given A, P(B|A), is: $\frac{P(A\&B)}{P(A)}$. The p-value is telling you nothing about the probability of the null hypothesis being true, so a lower p-value does not necessarily mean that the probability of the null hypothesis being true is lower (just look at the equation for conditional probability).

Exercise 5 — P-values

1. True or False?
   The p-value is the probability of the null hypothesis being true.
2. True or False?
   The p-value is the probability that the result occurred by chance.

3.15 Hypothesis testing: A more realistic scenario

In the above example we knew the $\sigma_{\bar{x}}$, because we knew σ. In the real world we do not know σ, so: instead of σ we use the unbiased estimator s; instead of $\sigma_{\bar{x}}$ we use the unbiased estimator $SE_{\bar{x}}$; switch to a t curve instead of a normal one.

Recall the z-score:

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} \tag{3.19}$$

$$= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{3.20}$$

Following exactly the same logic, we can compute a t-score:

$$t = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} \tag{3.21}$$

$$= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{3.22}$$

The normal distribution we used for the z-score was defined by a hypothesized mean and $\sigma_{\bar{x}}$. The t-distribution is defined by degrees of freedom—we have to find the probability under the "correct curve" (recall that the larger the sample size the tighter the spread of the distribution). Once we have the curve, we can compute the t test statistic as we did the z statistic, the two-sided significance at level 0.05. R does all this for us as follows:

```
> (z <- (sample.mean - 70)/(4/sqrt(11)))

[1] -8.190623

> sample <- rnorm(11, mean = 60, sd = 4)
> t.test(sample, alternative = "two.sided", mu = 70, conf.level = 0.95)

        One Sample t-test

data:  sample
t = -12.2216, df = 10, p-value = 2.459e-07
alternative hypothesis: true mean is not equal to 70
95 percent confidence interval:
 58.21453 61.84916
sample estimates:
mean of x
 60.03184
```

3.16 Comparing two samples

In one sample situations our null hypothesis is:

$$\mathcal{H}_0 : \bar{x} = \mu \tag{3.23}$$

When we compare two samples, we are asking the question: are the two populations of the two samples identical or not? Our goal now is to figure out some way to define our null hypothesis in this situation.

An example of a common scenario in experimental research is the following. Mean voice onset times and standard deviations are available of children and adults. The research question is, are

group	sample size	$\bar{x}$ (VOT)	s
children	10	-3.67	33.89
adults	20	-23.17	50.74

children different from adults in terms of voice onset time? We can re-frame this question as follows: the difference observed in the two sample means a true difference or just a chance event?

Such research problems have the property that (a) the goal is to compare the responses in two groups; (b) each group is considered a sample from a distinct population; (c) the responses in each group are independent of those in the other group; (d) the sample sizes of each group can be different.

The question now is, how can we formulate the null hypothesis?

3.16.1 $\mathcal{H}_0$ in two sample problems

We can say:

$$\mathcal{H}_0 : \mu_1 = \mu_2 \tag{3.24}$$

Alternatively:

$$\mathcal{H}_0 : \mu_1 - \mu_2 = 0 = \delta \tag{3.25}$$

We have effectively created a new population parameter δ:

$$\mathcal{H}_0 : \delta = 0 \tag{3.26}$$

We can define a *new* statistic $d = \bar{x}_1 - \bar{x}_2$ and use that as an estimator of δ, which we've hypothesized to be equal to zero. But to do this we need a sampling distribution of the difference of the sample means.

Let's do some simulation to get an understanding of this. Assume a population with mean (μ_1) 60, sd (σ_1) 4, and another with mean (μ_2) 62, sd (σ_2) 6. So we already know in this case that the null hypothesis is false. But let's take 1000 sets of samples of each population, compute the differences in mean in each set of samples, and plot that distribution of the differences of the sample mean.

The above figure suggests that we can safely take d to be an unbiased estimator of δ. What's the standard deviation of this new sampling distribution? It is clearly dependent on the standard deviation of the two populations in some way:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = f(\sigma_1, \sigma_2) \tag{3.27}$$

The precise relationship is:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{4^2}{11} + \frac{6^2}{15}} = 1.9633 \tag{3.28}$$

Suppose that in a single sample, $\bar{x}_1 - \bar{x}_2 = -5.2$. The null hypothesis $\mu_1 - \mu_2 = 0$. How to proceed? Recall that:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\text{sample mean} - \text{pop. mean}}{\text{sd of sampling distribution}} \tag{3.29}$$

```
> d <- rep(NA, 1000)
> for (i in c(1:1000)) {
+     sample1 <- rnorm(11, mean = 60, sd = 4)
+     sample2 <- rnorm(15, mean = 62, sd = 6)
+     d[i] <- mean(sample1) - mean(sample2)
+ }
> hist(d)
```

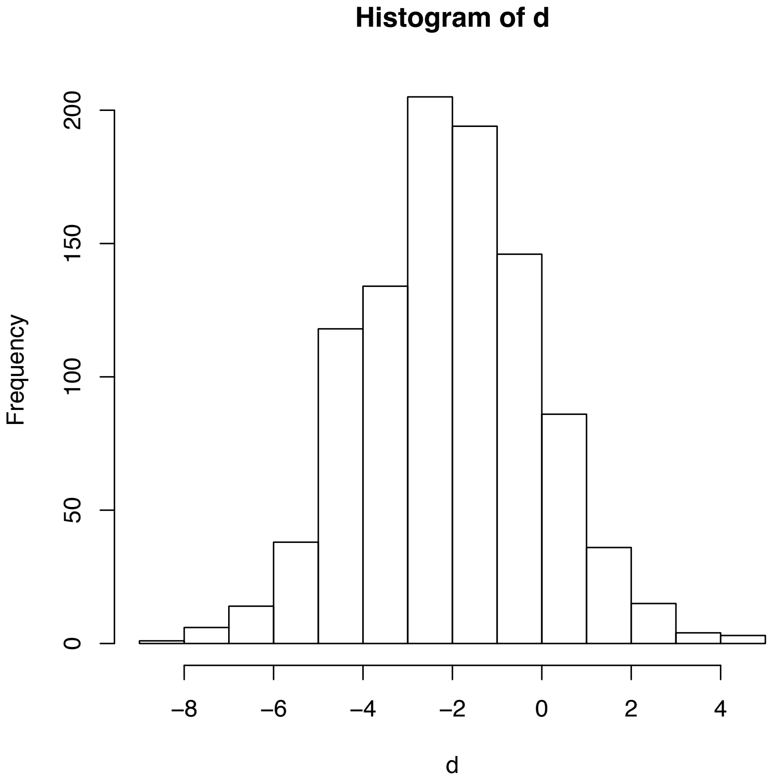


Figure 3.11: The distribution of the difference of sample means of two samples.

It follows that in the two-means case:

$$\begin{aligned} z &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{3.30} \\ &= \frac{-5.2 - 0}{1.9633} \tag{3.31} \\ &= -2.65 \tag{3.32} \end{aligned}$$

Using exactly the same logic as before (and because we don't know the population parameters in realistic settings),

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.33}$$

This is the two-sample t-statistic. One problem we face here is that the degrees of freedom needed for the correct t-curve is not obvious. The t-distribution assumes that only one s replaces a single σ; but we have two of these. If $\sigma_1 = \sigma_2$, we could just take a *weighted average* of the two sample SDs s_1 and s_2. This gives a POOLED ESTIMATOR and in this case the right t curve turns out to have $n_1 - 1 + n_2 - 1$ degrees of freedom.

However, in real life we don't know whether $\sigma_1 = \sigma_2$. In response to this, something called Welch's correction puts in a correction for possibly unequal variances into the t-curve. R does this correction for you if you specify that the variances are to be assumed to be unequal (`var.equal=FALSE`).

```
> sample1 <- rnorm(11, mean = 60, sd = 4)
> sample2 <- rnorm(15, mean = 62, sd = 6)
> t.test(sample1, sample2, mu = 0, alternative = "two.sided", conf.level = 0.95,
+     var.equal = FALSE)

        Welch Two Sample t-test

data:  sample1 and sample2
t = -2.171, df = 23.287, p-value = 0.04036
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.9718137 -0.2197232
sample estimates:
mean of x mean of y
 58.95264  63.54841
```

We now have the core concepts for carrying out statistical inference.

Chapter 4

Power

4.1 Review – z-scores

Let's quickly review what we have worked out so far. The sampling process for a single sample when the population parameters are known is as follows:

We take a sample, and the sample gives us a mean $\bar{x}$. We compute $\sigma_{\bar{x}}$ using σ and sample size n:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{4.1}$$

Then we calculate the range $\bar{x} \pm 2 \times \sigma_{\bar{x}}$. This 95% CI is a range of values within which we're 95% certain that the population mean lies.[1] Given knowledge about the population mean, we can express how far away the sample mean is using a z-score:

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{4.2}$$

If the z-score lies outside the CI, we conclude the population mean is not the hypothesized mean. However, in the real world we don't know σ, so: (a) instead of σ we use the unbiased estimator s; (b) instead of $\sigma_{\bar{x}}$ we use the unbiased estimator $SE_{\bar{x}}$; (c) switch to a t curve instead of a normal one.

Following the same logic as z-scores, we compute a t-score:

$$t = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{4.3}$$

We "look up" the t-curve, and establish whether the observed t-score falls outside the 95% CI or not. In two-sample problems, all that changes is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{4.4}$$

The relevant t-curve now is defined by $n_1 - 1 + n_2 - 1$ degrees of freedom.

4.2 Hypothesis testing revisited

Let's assume we do an experiment and compute the t-value and p-value; and we get a significant difference. The alternative possibility is that we don't get a significant difference: a "null result". What to do in the latter situation? Let's think again about the logic of hypothesis testing.

[1] We'll assume that the population mean is not a point value but rather a range of possible values, i.e., a random variable. See the earlier discussion on confidence intervals and their meaning.

4.3 Type I and Type II errors

We fix some conventions first. Let: A = "Accept the null hypothesis $\mathcal{H}_0$", and ¬A = "Reject the null hypothesis $\mathcal{H}_0$". The decision A or ¬A is based on the sample. In realistic situations, we don't know whether the null hypothesis is true or not. Let $\mathrm{P}(\neg\ \mathrm{A}\mid \mathcal{H}_0)$ = "Probability of rejecting the null hypothesis *assuming that the null hypothesis is in fact true.*"

Let's work through all the logical possibilities (Table 4.1):

Reality:	$\mathcal{H}_0$	$\neg\mathcal{H}_0$
Decision from sample is "reject":	$\mathrm{P}(\neg\mathrm{A}\mid \mathcal{H}_0)=\alpha$	$\mathrm{P}(\neg\mathrm{A}\mid \neg\mathcal{H}_0)= 1-\beta$
Decision from sample is "accept":	$\mathrm{P}(\mathrm{A}\mid \mathcal{H}_0)= 1-\alpha$	$\mathrm{P}(\mathrm{A}\mid \neg\mathcal{H}_0) = \beta$

Table 4.1: The logical possibilities given the two possible realities: null hypothesis true or false.

Reality:	$\mathcal{H}_0$	$\neg\mathcal{H}_0$
Decision from sample:	¬A (Type I)	¬A (Power)
Decision from sample:	A	A (Type II)

Table 4.2: Type I error, Type II error and power.

As shown in Table 4.2, Type I error is $\mathrm{P}(\neg A \mid \mathcal{H}_0) = \alpha$ and is conventionally held at 0.05. Type II error is $\mathrm{P}(A \mid \neg\mathcal{H}_0) = \beta$. Power refers to $\mathrm{P}(\neg A \mid \neg\mathcal{H}_0) = (1-\beta)$.

Let's do some simulation to get a get a better understanding of these various definitions. Consider first the case where the null hypothesis is in fact true.

Recall an example from an earlier part of this book: Assume a population with mean (μ_1) 60, sd (σ_1) 4, and another with mean (μ_2) 62, sd (σ_2) 6. Here we already *know* in this case that the null hypothesis is false. What is the null hypothesis? That the difference of the means is zero. So the distribution that we'd use to do inferencing is shown in Figure 4.1.

Figure 4.1 shows the distribution corresponding to the null hypothesis. The vertical lines are 95% CIs. In this simulation we **know** that there is a difference in population means: and that difference is -2. Now think about the right side of Table 4.2; in our current simulation, the null hypothesis is false with a specific value, -2. What's the distribution corresponding to **this particular situation where the actual population difference is -2**?

Figure 4.2 shows the distribution corresponding to the null hypothesis overlaid with the **actual** distribution, which we **know** is centered around -2. The vertical lines are 95% CIs **assuming the null hypothesis is true**.

Now let's remove from the figure the distribution corresponding to the null hypothesis, Figure 4.3. Some important insights emerge from this figure.

First, making α smaller (widening the CIs) means β becomes bigger (the area under the curve between the CI bars increases), and vice versa. Second, making α smaller means $1-\beta$ (or power) decreases too. Third, making α bigger means $1-\beta$ (or power) increases too.

Now recall this figure from page **??**. We reproduce it here for convenience:

```
> numdrops <- 40
> p <- 0.5
> n <- c(0:numdrops)
> num <- numdrops
> probs <- c()
> for (k in n) {
+     currentk <- binomialprobability(num, p, k)
```

```
> d <- c()
> for (i in c(1:1000)) {
+     sample1 <- rnorm(11, mean = 60, sd = 4)
+     sample2 <- rnorm(15, mean = 62, sd = 6)
+     currentd <- mean(sample1) - mean(sample2)
+     d <- append(d, currentd)
+ }
> xvals <- seq(-6, 6, 0.1)
> plot(xvals, dnorm(xvals, mean = 0, sd = 1.9633), type = "l",
+     lwd = 2, ylab = "", col = "red")
> arrows(-(2 * 1.9633), -0.05, -(2 * 1.9633), 0.2, angle = 0)
> arrows((2 * 1.9633), -0.05, (2 * 1.9633), 0.2, angle = 0)
> text(-4.5, 0.008, expression(alpha/2), cex = 1.5, col = "red")
> text(4.5, 0.008, expression(alpha/2), cex = 1.5, col = "red")
> text(0, 0.008, expression(1 - alpha), cex = 1.5, col = "red")
```

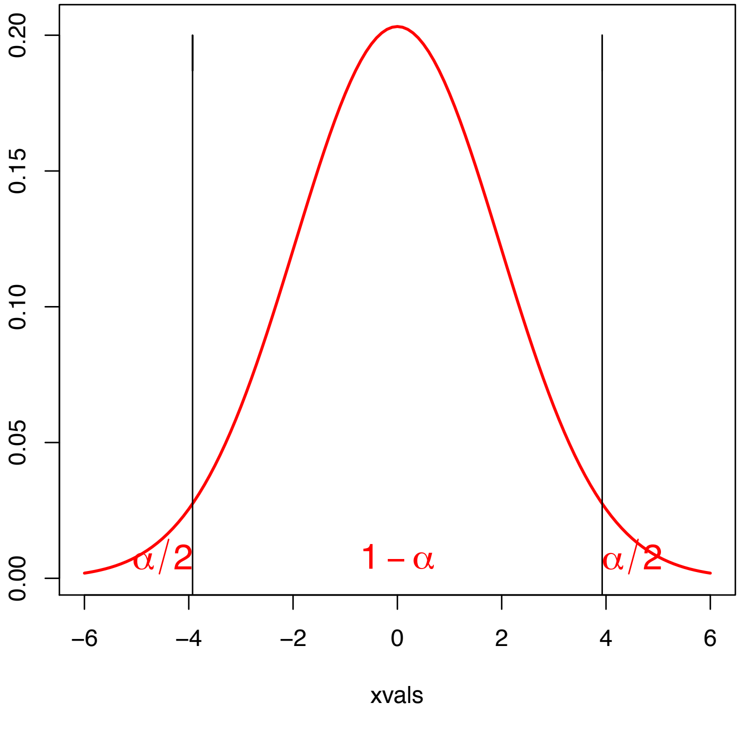


Figure 4.1: The distribution corresponding to the null hypothesis, along with rejection regions (95% confidence intervals).

```
> plot(xvals, dnorm(xvals, mean = 0, sd = 1.9633), type = "l",
+     ylab = "", col = "red")
> arrows(-(2 * 1.9633), -0.05, -(2 * 1.9633), 0.2, angle = 0)
> arrows((2 * 1.9633), -0.05, (2 * 1.9633), 0.2, angle = 0)
> lines(xvals, dnorm(xvals, mean = mean(d), sd = sd(d)), lwd = 2)
> text(-4.5, 0.008, expression(alpha/2), cex = 1.5, col = "red")
> text(4.5, 0.008, expression(alpha/2), cex = 1.5, col = "red")
> text(0, 0.008, expression(1 - alpha), cex = 1.5, col = "red")
```

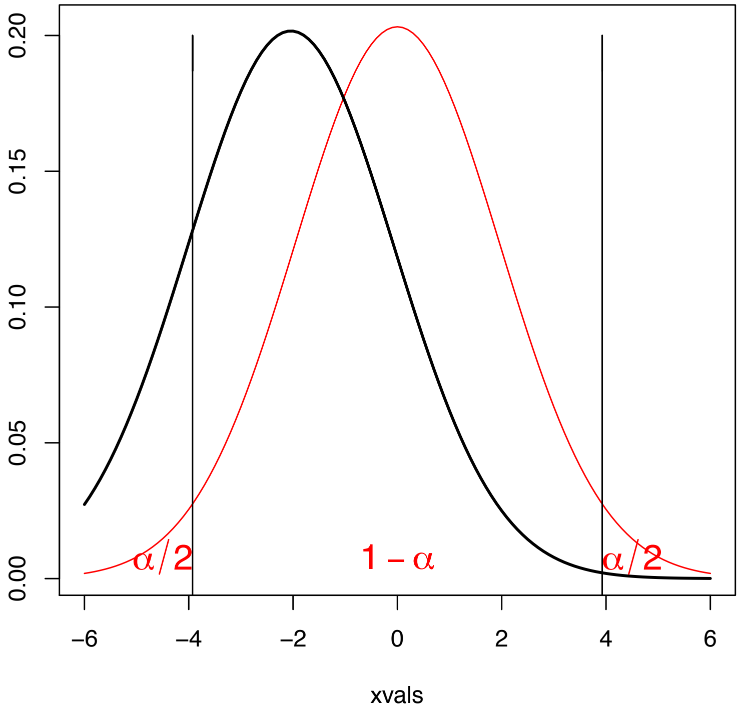


Figure 4.2: The distribution corresponding to the null hypothesis and the distribution corresponding to the true population scores.

```
> plot(xvals, dnorm(xvals, mean = mean(d), sd = sd(d)), lwd = 2,
+     type = "l")
> arrows(-(2 * 1.9633), -0.05, -(2 * 1.9633), 0.2, angle = 0)
> arrows((2 * 1.9633), -0.05, (2 * 1.9633), 0.2, angle = 0)
> text(-2, 0.08, expression(beta), cex = 1.5, col = "red")
> text(-5, 0.02, expression(1 - beta), cex = 1.5, col = "black")
> arrows((2 * 1.9633 + 1), 0.04, (2 * 1.9633 + 0.2), 0, angle = 45)
> text((2 * 1.9633 + 1), 0.04, expression(1 - beta), cex = 1.5,
+     col = "black")
```

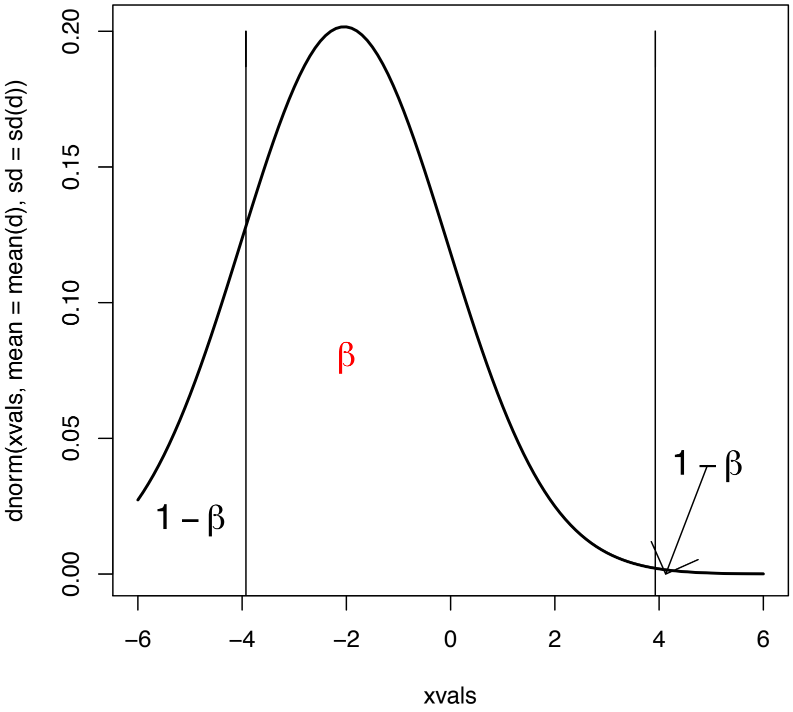
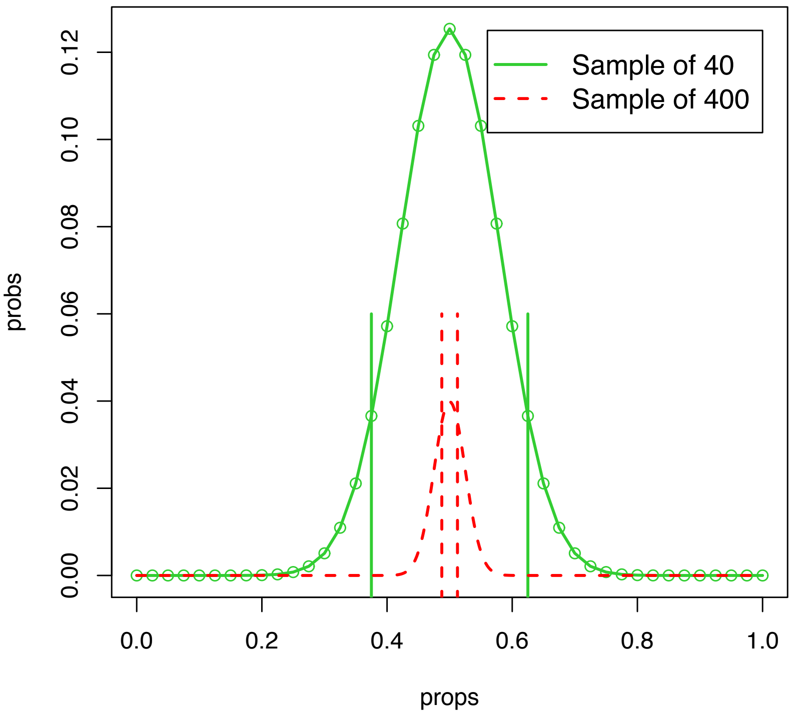


Figure 4.3: The distribution corresponding to the true population scores overlaid with the confidence interval from the distribution corresponding to the null hypothesis.

```
+     probs <- append(probs, currentk)
+ }
> props <- n/num
> plot(props, probs, type = "p", col = "limegreen")
> lines(props, probs, col = "limegreen", lwd = 2)
> segments(props[[(numdrops/2 + 1) - 5]], -0.5, props[[(numdrops/2 +
+     1) - 5]], 0.06, col = "limegreen", lty = 1, lwd = 2)
> segments(props[[(numdrops/2 + 1) + 5]], -0.5, props[[(numdrops/2 +
+     1) + 5]], 0.06, col = "limegreen", lty = 1, lwd = 2)
> numdrops <- 400
> p <- 0.5
> n <- c(0:numdrops)
> num <- numdrops
> probs <- c()
> for (k in n) {
+     currentk <- binomialprobability(num, p, k)
+     probs <- append(probs, currentk)
+ }
> props <- n/num
> lines(props, probs, col = "red", lwd = 2, lty = 2)
> segments(props[[(numdrops/2 + 1) - 5]], -0.5, props[[(numdrops/2 +
+     1) - 5]], 0.06, col = "red", lty = 2, lwd = 2)
> segments(props[[(numdrops/2 + 1) + 5]], -0.5, props[[(numdrops/2 +
+     1) + 5]], 0.06, col = "red", lty = 2, lwd = 2)
> leg.txt <- c("Sample of 40", "Sample of 400")
> legend(1, 0.125, legend = leg.txt, col = c("limegreen", "red"),
+     lty = c(1, 2), cex = 1.2, lwd = 2, xjust = 1, yjust = 1,
+     merge = TRUE)
```



As sample size increases from 40 to 400, we get proportionally tighter 95% confidence intervals. This fact is now relevant in Figure 4.3: the narrower the 95% CI, the higher the power; note that the value of β (The probability of a Type II error, P(A$|\neg\mathcal{H}_0$)) will obviously also go down.

So, if you have a relatively narrow CI, and a nonsignificant result ($p > .05$), you have relatively high power and a relatively low probability of making a Type II error (of accepting a null hypothesis as true when it is in fact not true).

A **heuristic** suggested by (?, ?) is: if you have a narrow CI, and a nonsignificant result, you have some justification for concluding that the null hypothesis may in fact be true. Conversely, if you have a wide CI and a nonsignificant result, all bets are off: *the result is inconclusive*.

The above heuristic seems a bit vague; how to define "narrow CI"? If the goal is to argue for the null hypothesis, one solution is equivalence/bioequivalence testing. The basic idea is to reverse the burden of proof. The null hypothesis becomes the alternative hypothesis and the alternative the null:

$$\mathcal{H}_0 : d \leq \Theta_L \text{ or } d \geq \Theta_U \tag{4.5}$$

$$\mathcal{H}_a : \Theta_L < d < \Theta_U \tag{4.6}$$

4.4 Equivalence testing

There are two techniques: TOST: Two one-sample t-tests; and confidence intervals approach.

4.4.1 Equivalence testing example

Let's look at equivalence testing using a concrete example. This example is taken from (Stegner, Bostrom, & Greenfield, 1996).

We have data on two kinds of case management randomly applied to 201 seriously mentally disordered patients: (a) Traditional Case Management or TCM (Control) (b) TCM plus trained service coordinators (Treatment). Treatment is costlier than Control, so if they're not significantly different the extra cost is money wasted. Dependent measure: Brief Psychiatric Rating Scale (for our purposes, it doesn't matter what exactly it is). (Some patients' data were not available). Data summary:

Group	n	Mean	SD
Control	64	1.5679	0.4285
Treatment	70	1.6764	0.4748
Total	134	1.6246	0.4533 (pooled)

Let $\bar{x}_C$ be the mean for controls, and $\bar{x}_T$ the mean for treatment, and let pooled standard deviation be s_{pooled}. Specifically, $\bar{x}_C = 1.5679, \bar{x}_T = 1.6764, s_{\text{pooled}} = 0.4533$

Therefore, the difference between the two means d is: $d = \bar{x}_T - \bar{x}_C = 1.6764 - 1.5679 = 0.1085$.

Here, the research goal is to find out if the treatment is effective or not; if it's not, the difference between the means should be "essentially" equivalent. In order to formally specify what is meant by "essentially" equivalent, we can specify an equivalence threshold Θ; if d lies within this threshold we accept the null. Suppose previous experience in the field suggests that a difference of 20% or less with respect to the control's mean can be considered to be equivalent. $\Theta = .2 \times 1.5679 = 0.3136$. There has to be some independent, prior criterion for deciding what $Theta$ will be.

4.4.2 TOST approach to the Stegner et al. example

Since $\Theta = 0.3136$, we can define two limits around 0 that constitute the equivalence threshold: $\Theta_L = -0.3136, \Theta_U = 0.3136$. If d lies within this region we reject the hypothesis that the two means are different. Thus, our null and alternative hypotheses are:

$$\mathcal{H}_0 : d \leq \Theta_L \text{ or } d \geq \Theta_U \tag{4.7}$$

$$\mathcal{H}_a : \Theta_L < d < \Theta_U \tag{4.8}$$

It follows that:

$$t = \frac{d - \Theta}{SE} = \frac{d - \Theta}{s_{\text{pooled}} / \sqrt{(1/n_1 + 1/n_2)}} = -2.616 \tag{4.9}$$

$$t = \frac{d + \Theta}{SE} = \frac{d + \Theta}{s_{\text{pooled}} / \sqrt{(1/n_1 + 1/n_2)}} = 5.384 \tag{4.10}$$

t(134-2)=1.6565 (`qt(.95, df = 132)`), so both parts of the null hypotheses (3) and (4) are rejected. Conclusion: The difference between the two population means is no greater than Θ; the extra cost is unjustified.

Summary of calculations:

- TOST(mean1,mean2,Θ, n_1, n_2, σ)

- Compute two one-way t-tests:

$$t_{d \leq \Theta_L} = \frac{d - \Theta}{s_{\text{pooled}} / \sqrt{(1/n_1 + 1/n_2)}} \tag{4.11}$$

$$t_{d \geq \Theta_U} = \frac{d + \Theta}{s_{\text{pooled}} / \sqrt{(1/n_1 + 1/n_2)}} \tag{4.12}$$

- Compute critical t-value t_{crit} (the 95% CI cutoff points). In R this is done as follows: $qt(.95, (n_1 + n_2 - 2))$

- Iff $t_{d \leq \Theta_L} < -t_{\text{crit}}$ and $t_{d \geq \Theta_U} > t_{\text{crit}}$, we can reject the null hypothesis.

It's easy to write a function that does this for us in general:

```
> TOST <- function(mean1, mean2, theta, n1, n2, sigma) {
+     d <- (mean2 - mean1)
+     t1 <- (d - theta)/(sigma * (sqrt((1/n1) + (1/n2))))
+     t2 <- (d + theta)/(sigma * (sqrt((1/n1) + (1/n2))))
+     tcrit <- qt(0.95, (n1 + n2 - 2))
+     if ((t1 < -tcrit) && (t2 > tcrit)) {
+         print(t1)
+         print(t2)
+         print(tcrit)
+         print(c("Equivalent"))
+     }
+     else {
+         print(c("Failed to show equivalence"))
+     }
+ }
```

4.4.3 Equivalence testing example: CIs approach

(?, ?) showed that TOST is operationally equivalent to determining whether 100(1-2α)% CIs fall within the range $-\Theta \cdots + \Theta$. Recall that $t_{\text{crit}} = 1.6565$. We can now compute the confidence intervals (CI):

$$CI = d \pm 1.6565 \times SE \tag{4.13}$$

$$= d \pm 1.6565 \times \left(\frac{\sigma}{\sqrt{(1/n_1 + 1/n_2)}}\right) \tag{4.14}$$

$$= d \pm 1.6565 \times \left(\frac{\sigma}{\sqrt{(1/n_1 + 1/n_2)}}\right) \tag{4.15}$$

$$= 0.1085 \pm 1.6565 \times \left(\frac{0.4533}{\sqrt{(1/64 + 1/70)}}\right) \tag{4.16}$$

$$= 0.1085 \pm 0.1299 \tag{4.17}$$

Since $(-0.0214, 0.2384)$ lies within the range $(-0.3136, +0.3136)$ we can declare equivalence. Recall now the heuristic we gave earlier: narrow CIs, accept null hypothesis; wide CIs, inconclusive. It's related to the above discussion.

4.5 Equivalence testing bibliography

Here is a short bibliography on articles on equivalence testing in case you are interested:

- Bruce Stegner et al. (1996). Equivalence testing for use in psychosocial and services research: An introduction with examples. Evaluation and Program Planning, Vol. 19, No. 3, pp. 193-198. (*easy to read introduction*).
- Graham McBride (1999). Equivalence tests can enhance environmental science and management. Austral. & New Zealand J. Statist., pp. 19-29. (*also easy to read, excellent graphics-based explanations*).
- John Hoenig et al. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. The American Statistician, Feb. 2001, Vol. 55, No. 1, pp. 19-24 (*more technical than the previous ones, but a classic in the field*).
- John Berger et al. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. Statistical Science, Vol. 11, No. 4, pp. 283-302. (*technical, review article; will put hair on your chest*).
- Walter Hauck et al. (1996). [Bioequivalence trials, intersection-union tests and equivalence confidence sets]: Comment. Statistical Science, Vol. 11, No. 4, p. 303.

4.6 Observed power and null results

Many journals and organizations ask you to compute "observed power" if you get a null result (Hoenig & Heisey, 2001). The logic is: if you got a null result ($p > .05$) and the "observed power" based on the sample is high ($> .80$), then you can safely accept the null result as true. After all, $P(\neg A \mid \neg \mathcal{H}_0) > .80$, so if you assert that A, you're fine, right? The problem with this is that the p-value and "observed power" are inversely related. "Observed power" provides no new information after the p-value is known. Let's convince ourselves this is true with a simple example.

Take the earlier example of a population with mean (μ_1) 60, sd (σ_1) 4, and another with mean (μ_2) 62, sd (σ_2) 6. If we compare the means from two samples, one taken from each population, and our null hypothesis is that the two samples come from the same population, we already know in this case that the null hypothesis is false. But suppose we didn't know this, and we got a sample mean difference of -3.5 and some p-value $p > 0.05$. We can compute "observed" power using this **observed** difference (cf. the **actual** difference -2 that we'd used earlier).

But if our sample mean difference had been -1 and the associated p-value p' had been greater than p, we could also have computed observed power.

Figure 4.4 shows that the area under the curve outside the vertical lines (power) decreases as p-values go up (as the difference in the sample means comes closer to zero). For this reason, computing observed power does not provide any new information: if the p-value is high, we already know the observed power is low, there is nothing gained by computing it.

Another commmon-enough approach is to keep increasing sample size n until you get a significant difference. Recall that $\sqrt{n}$ is inversely related to SE, which is used to compute 95% CIs. So, by increasing sample size, you are narrowing the CIs, thereby increasing power on the fly. A better approach is to do power analysis **before** running the experiment. R has `power.t.test` and `power.anova.test` for this purpose. Use it to compute sample size before running your experiment.

For example, suppose we about to run a self-paced reading experiment, and we expect (from previous work or from the predictions of some computational model) a reading time difference of about 200 msecs between two conditions. I also expect noisy data: SD about 200. I want to ensure that I have high power, say 0.80. How many subjects do I need?

```
> d1 <- rep(NA, 1000)
> for (i in c(1:1000)) {
+     sample1 <- rnorm(11, mean = 60, sd = 4)
+     sample2 <- rnorm(15, mean = 63.5, sd = 6)
+     d1[i] <- mean(sample1) - mean(sample2)
+ }
> d2 <- rep(NA, 1000)
> for (i in c(1:1000)) {
+     sample1 <- rnorm(11, mean = 60, sd = 4)
+     sample2 <- rnorm(15, mean = 61, sd = 6)
+     d2[i] <- mean(sample1) - mean(sample2)
+ }
> op <- par(mfrow = c(1, 2), pty = "s")
> plot(density(d1), xlab = "", main = "Smaller p-value, larger obs. power")
> arrows(-(2 * 1.9633), -0.05, -(2 * 1.9633), 0.2, angle = 0)
> arrows((2 * 1.9633), -0.05, (2 * 1.9633), 0.2, angle = 0)
> plot(density(d2), xlab = "", main = "Larger p-value, smaller obs. power")
> arrows(-(2 * 1.9633), -0.05, -(2 * 1.9633), 0.2, angle = 0)
> arrows((2 * 1.9633), -0.05, (2 * 1.9633), 0.2, angle = 0)
```

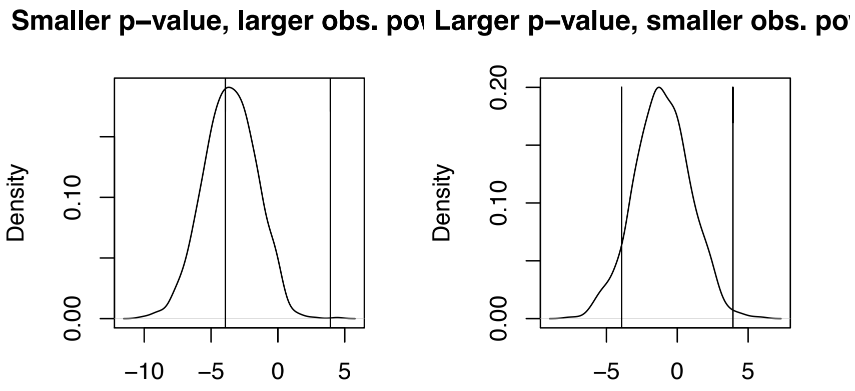


Figure 4.4: The relationship between observed power and p-values is inverse.

```
> power.t.test(n = NULL, delta = 200, sd = 200, sig.level = 0.05,
+     power = 0.8, type = c("two.sample"), alternative = c("two.sided"))

     Two-sample t test power calculation

              n = 16.71477
          delta = 200
             sd = 200
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

 NOTE: n is number in *each* group
```

Exercise 6

Recall an example from the last lecture:

group	sample size	$\bar{x}$ (VOT)	s
children	10	-3.67	33.89
adults	20	-23.17	50.74

Is the difference observed in the two sample means significantly different at α level 0.05? Give the t-score. Plot the t-distribution for the relevant degrees of freedom. To do this, you will need to compute the degrees of freedom using Welch's formula (look up the formula on the web). Given the t-score, can you safely (at α level 0.05) reject the null hypothesis that the two populations' means are identical? Explain why or why not.

Draw a bar-graph showing the sample means in the data, and plot 95% CIs around this mean using the relevant s and sample size. You will need to use the `arrows` command in R, and the parameter `angle` in that command.

Chapter 5

Analysis of variance

5.1 Comparing three populations

Consider three second-language vocabulary learning methods (I, II, III), three subjects assigned to each method. The relative effectiveness of learning methods is evaluated on some scale by scoring the increase in vocabulary after using the method.

	Group I	Group II	Group III
	9	10	1
	1	2	5
	2	6	0
$\bar{x}$	4	6	2

Suppose our research question is: is any one of the learning methods better than the others? We could reason as follows:

- Do a t-test on I vs. II, I vs. III, II vs. III.
- If *at least one* of the three null hypotheses can be rejected, we can safely reject the main research hypothesis that all the three groups' means are the same.

Here are the three null hypotheses:

$$A \rightarrow \mathcal{H}_{0_{G1,G2}} : \mu_{G1} = \mu_{G2} \tag{5.1}$$

$$B \rightarrow \mathcal{H}_{0_{G1,G3}} : \mu_{G1} = \mu_{G3} \tag{5.2}$$

$$C \rightarrow \mathcal{H}_{0_{G2,G2}} : \mu_{G2} = \mu_{G3} \tag{5.3}$$

And here (Table **??**) are all the logically possible outcomes (let ¬ X mean "reject X"):

1st Col	2nd Col	3rd Col	4th Col	5th Col	6th Col	7th Col	8th Col
A	¬A	A	A	¬A	A	¬A	¬A
B	B	¬B	B	¬B	¬B	B	¬B
C	C	C	¬C	C	¬C	¬C	¬C

Table 5.1: All the logically possible outcomes when we compare conditions pairwise.

Let the probability of rejecting the null hypothesis when it is true, i.e., the α level, be 0.05. The probability of rejecting at least one null hypothesis is the sum of the probabilities of each of the mutually exclusive events in the 2nd to 9th columns. Hence, the probability of rejecting at least one null hypothesis is:

$$(0.05)^3 + 3 \times 0.95 \times (0.05)^2 + 3 \times 0.05 \times (0.95)^2 = 0.142625 \tag{5.4}$$

So now our new α level is no longer 0.05. What to do? Recall our current running example, but now consider two separate samples.

	Group I	Group II	Group III	Group I	Group II	Group III
	9	10	1	3	7	1
	1	2	5	4	6	2
	2	6	0	5	5	3
$\bar{x}$	4	6	2	4	6	2

Group II seems to be doing consistently better. However, an important difference between the first and second sample is that in the first there is a lot more variance within groups. The variation in the mean scores (between-group variation) in the first sample could just be due to within-group variation.

In the second sample, there's a lot less within-group variation, but the between-group variation is just the same—maybe Group II really is doing significantly better. What we just did was analyze variance between and within groups—hence the name of this procedure: Analysis of variance or ANOVA.

5.2 ANOVA

Recall the earlier insight: averaging the observations leads to the true population parameter, i.e., errors tend to cancel out.

We can express the means $\bar{x}_1, \bar{x}_2, \bar{x}_3$ in terms of the population parameter and the errors:

$$\bar{x}_1 = \mu + \epsilon_1 \tag{5.5}$$

$$\bar{x}_2 = \mu + \epsilon_2 \tag{5.6}$$

$$\bar{x}_3 = \mu + \epsilon_3 \tag{5.7}$$

This is just in the present example; in general, for j groups, we'd have:

$$\bar{x}_j = \mu + \epsilon_j \tag{5.8}$$

Gauss noticed that the sampling *error* also has a normal distribution. We can see this in the simulation below, which should be self-explanatory. In Figure 5.1, Error = Sample mean - Population mean.

In what follows, we will use this fact to build statistical models.

5.3 Statistical models

Characterizing a sample mean as an error about a population mean is called building a statistical model:

$$\bar{x}_j = \mu + \epsilon_j \tag{5.9}$$

It's a powerful idea because it allows you to *compare* statistical models and decide which one better characterizes the data; we'll be looking at just how powerful this idea is, in subsequent

```
> error <- rep(NA, 1000)
> for (i in c(1:1000)) {
+     sample <- rnorm(10, mean = 0, sd = 1)
+     currentmean <- mean(sample)
+     error[i] <- currentmean - 0
+ }
> plot(density(error))
```

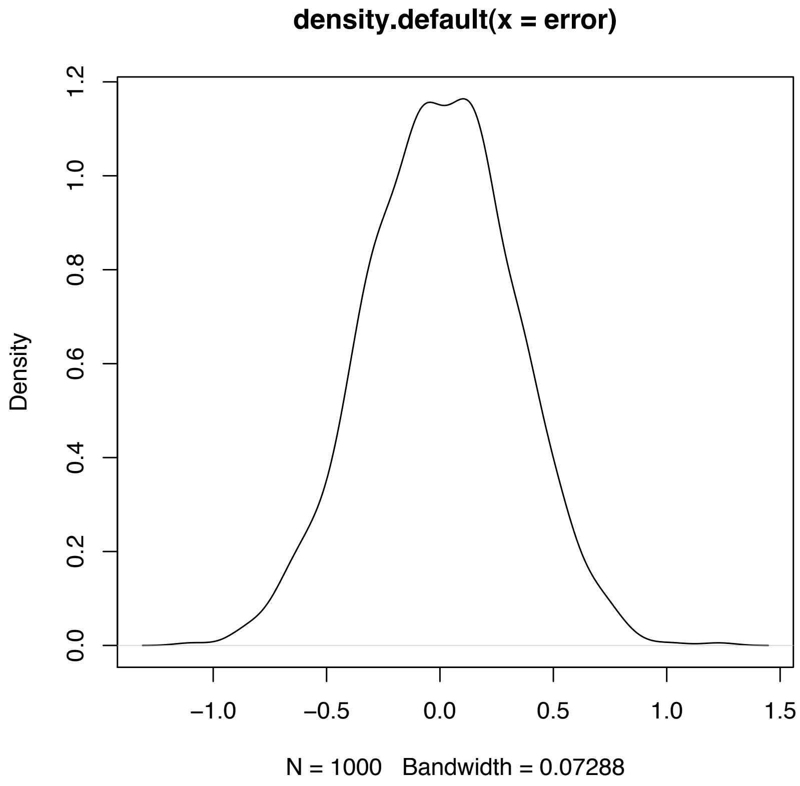


Figure 5.1: Errors tend to cancel out.

lectures. The way we'll use it is that one statistical model expresses the null hypothesis (no difference between means), and the other the alternative hypothesis (systematic difference between means).

I assume here that you know the concept of FACTORIAL DESIGN.[1] A given factor α has one or more levels (in the present case, three: α_1, α_2, α_3). (The use of alpha's here has nothing to do with the alpha-value.)

Let's build an idealized model first, using an idealized dataset, and labeling each cell:

	Group I	Group II	Group III
	$x_{1,1} = 4$	$x_{1,2} = 4$	$x_{1,3} = 4$
	$x_{2,1} = 4$	$x_{2,2} = 4$	$x_{2,3} = 4$
	$x_{3,1} = 4$	$x_{3,2} = 4$	$x_{2,3} = 4$
$\bar{x} = 4$	$\bar{x}_1 = 4$	$\bar{x}_2 = 4$	$\bar{x}_3 = 4$

More generally, this simplified model for i subjects, $i = 1, \ldots, n$, and j groups, $j = 1, \ldots, m$, is:

$$x_{i,j} = \mu \tag{5.10}$$

Now look at this slightly different data, the only change is that it has some variation in it:

	Group I	Group II	Group III
	$x_{1,1} = 4$	$x_{1,2} = 6$	$x_{1,3} = 2$
	$x_{2,1} = 4$	$x_{2,2} = 6$	$x_{2,3} = 2$
	$x_{3,1} = 4$	$x_{3,2} = 6$	$x_{2,3} = 2$
$\bar{x} = 4$	$\bar{x}_1 = 4$	$\bar{x}_2 = 6$	$\bar{x}_3 = 2$

Note that the GRAND MEAN $\bar{x}$ is still 4. The model now (in its most general form) is:

$$x_{i,j} = \mu + \alpha_j \tag{5.11}$$

"Unpacking" this model in the table we get $\alpha_1 = \mathbf{0}, \alpha_2 = \mathbf{2}, \alpha_3 =$ **-2**, see Table 5.2.

	Group I	Group II	Group III
	$x_{1,1} = 4 + \mathbf{0}$	$x_{1,2} = 4 + \mathbf{2}$	$x_{1,3} = 4 - \mathbf{2}$
	$x_{2,1} = 4 + \mathbf{0}$	$x_{2,2} = 4 + \mathbf{2}$	$x_{2,3} = 4 - \mathbf{2}$
	$x_{3,1} = 4 + \mathbf{0}$	$x_{3,2} = 4 + \mathbf{2}$	$x_{2,3} = 4 - \mathbf{2}$
$\bar{x} = 4$	$\bar{x}_1 = 6$	$\bar{x}_2 = 6$	$\bar{x}_3 = 2$

Table 5.2: Idealized situation: the effect of factors

Think about what α_j is: the variation between groups. Now, in real life, scores also show variation *within* a group of subjects—individual subjects differ. For each subject i in each group j, we can represent this within-subject variation as an error component ϵ_{ij}:

$$x_{i,j} = \mu + \alpha_j + \epsilon_{ij} \tag{5.12}$$

Coming back to our current running example of two separate samples, we can decompose the scores according to the model:

$$x_{i,j} = \mu + \alpha_j + \epsilon_{ij} \tag{5.13}$$

	Group I	Group II	Group III	Group I	Group II	Group III
	4+0+5=9	4+2+4=10	4-2-1=1	4+0-1=3	4+2+1=7	4-2-1=1
	4+0-3=1	4+2-4=2	4-2+3=5	4+0+0=4	4+2+0=6	4-2+0=2
	4+0-2=2	4+2+0=6	4-2-2=0	4+0+1=5	4+2-1=5	4-2+1=3
$\bar{x} = 4$	4	6	2	4	6	2

Here, we'd just assumed that $\mu = 4$, and $\alpha_1 = \mathbf{0}, \alpha_2 = \mathbf{2}, \alpha_3 = \mathbf{\text{-}2}$, but in real life we don't know these—we have to estimate them. Estimating α_j is equivalent to asking whether effect α_j exists or not.

Asking whether effect α_j exists is basically a matter of comparing two models:

$$\mathcal{H}_0 : x_{ij} = \mu + \epsilon_{ij} \tag{5.14}$$

$$\mathcal{H}_a : x_{ij} = \mu + \alpha_j + \epsilon_{ij} \tag{5.15}$$

If there is no effect α_j, then the variation observed between mean scores of groups is *due only to error variation.* If an effect α_j is present, the variation between groups increases because of the systematic differences between groups: the between-group variation is due to error variation plus variation due to α_j.

Exercise 7

Do all computations first by hand and then compare your result with computations using R's t.test function.

Given three second-language vocabulary learning methods (I, II, III), three subjects are assigned to each method. The relative effectiveness of learning methods is evaluated on some scale by scoring the increase in vocabulary after using the method.

	Group I	Group II	Group III
	9	10	1
	1	2	5
	2	6	0
$\bar{x}$	4	6	2

Evaluate the research question: is there any difference in the three learning methods? Do three pairwise t-tests. Can one conclude anything from the results? If so, what? If nothing, why not?

Now do the three t-tests to evaluate the same research question with this new sample from the same setup:

	Group I	Group II	Group III
	3	7	1
	4	6	2
	5	5	3
$\bar{x}$	4	6	2

Is anything significant? Can we conclude anything this time regarding the research question? If any of the tests are significant, is the p-value low enough that we can reject that particular null hypothesis at $\alpha = .05$?

Note: you will need either to look up a t-test table from a statistics textbook, or you can use R to ask: what's the critical t for n degrees of freedom at an alpha level of 0.05: $qt(.975, n)$.

Coming back to our current running example with two separate samples.

[1] If not, see (Ray, 2000) for a simple discussion.

	Group I	Group II	Group III	Group I	Group II	Group III
	9	10	1	3	7	1
	1	2	5	4	6	2
	2	6	0	5	5	3
$\bar{x}$	4	6	2	4	6	2

There is some population grand mean μ (we don't know what it is). To this mean, each of the separate sub-populations j (groups) might or might not contribute their own effect α_j.

Then there is the variation in individual subjects i within a sub-population j (group): the "subject error" ϵ_{ij}.

Our null hypothesis is that $\alpha_j = 0$, i.e., that the three groups (the three learning methods) have no effect on the score. We can write this as:

$$\mathcal{H}_0 : x_{ij} = \mu + \epsilon_{ij} \tag{5.16}$$

The alternative hypothesis is that $\alpha \neq 0$, and we can write this as:

$$\mathcal{H}_a : x_{ij} = \mu + \alpha_j + \epsilon_{ij} \tag{5.17}$$

Where do we go from here? How to conclude something about the null hypothesis?

5.4 Measuring variation

Recall the definition of variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{5.18}$$

Recall that $n-1$ is the DEGREES OF FREEDOM. Let's call the numerator the SUM OF SQUARES or SS as:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{5.19}$$

We can ask (for reasons that will become clear very soon) what the within-group and between-group SS's are:

	Group I	Group II	Group III
	9	10	1
	1	2	5
	2	6	0
$\bar{x} = 4$	4	6	2

- Group I's SS is: $\sum\limits_{i=1}^{n_1}(x_{i1} - \bar{x}_1)^2 = 38$

- Group II's SS is: $\sum\limits_{i=1}^{n_2}(x_{i2} - \bar{x}_2)^2 = 32$

- Group III's SS is: $\sum\limits_{i=1}^{n_3}(x_{i3}-\bar{x}_3)^2 = 10$
- Between-group SS: $\sum\limits_{j=1}^{3}(\bar{x}_j-\bar{x})^2 = 8$

So we have SS's for within-group variation in each group, and the SS for between-group variation.

We are going to use these to compare between- and within-group variation. But our approach to doing inference from the comparison is going to be identical to the z- and t-scores approach we did last time:

1. Find a statistic (or statistics) which relates to our data and whose distribution is known if the null hypothesis is true (=no effect α_j).
2. Plot this distribution and then note where a related test statistic falls in this distribution, and determine the probability of getting a test statistic as extreme or more extreme than we did, *assuming that the null hypothesis is true.*

5.5 A simple but useful manipulation

Before we get to the inference stage, we have to do some algebraic manipulation.

Notice that the following equality holds:

$$\begin{aligned}
x_{ij} &= x_{ij} \tag{5.20}\\
x_{ij}-\bar{x} &= x_{ij}-\bar{x} \tag{5.21}\\
&= x_{ij}+(-\bar{x}_j+\bar{x}_j)-\bar{x} \tag{5.22}\\
&= (x_{ij}-\bar{x}_j)+(\bar{x}_j-\bar{x}) \tag{5.23}\\
&= (\bar{x}_j-\bar{x})+(x_{ij}-\bar{x}_j) \tag{5.24}
\end{aligned}$$

For all i, j:

$$x_{ij}-\bar{x} = (\bar{x}_j-\bar{x})+(x_{ij}-\bar{x}_j) \tag{5.25}$$

Unpack the above equation:

$$\begin{aligned}
x_{11}-\bar{x} &= (\bar{x}_1-\bar{x})+(x_{11}-\bar{x}_1) \tag{5.26}\\
x_{21}-\bar{x} &= (\bar{x}_1-\bar{x})+(x_{21}-\bar{x}_1) \tag{5.27}\\
x_{31}-\bar{x} &= (\bar{x}_1-\bar{x})+(x_{31}-\bar{x}_1) \tag{5.28}\\
x_{12}-\bar{x} &= (\bar{x}_2-\bar{x})+(x_{12}-\bar{x}_1) \tag{5.29}\\
x_{22}-\bar{x} &= (\bar{x}_2-\bar{x})+(x_{22}-\bar{x}_1) \tag{5.30}\\
x_{32}-\bar{x} &= (\bar{x}_2-\bar{x})+(x_{32}-\bar{x}_1) \tag{5.31}\\
x_{13}-\bar{x} &= (\bar{x}_3-\bar{x})+(x_{13}-\bar{x}_1) \tag{5.32}\\
x_{23}-\bar{x} &= (\bar{x}_3-\bar{x})+(x_{23}-\bar{x}_1) \tag{5.33}\\
x_{33}-\bar{x} &= (\bar{x}_3-\bar{x})+(x_{33}-\bar{x}_1) \tag{5.34}
\end{aligned}$$

We can repackage these equations more concisely:

$$\sum_{j=1}^{I}\sum_{i=1}^{n_j}x_{ij}-\bar{x} = \sum_{j=1}^{I}\sum_{i=1}^{n_j}((\bar{x}_j-\bar{x})+(x_{ij}-\bar{x}_j)) \tag{5.35}$$

If we square the terms on both sides ...

$$\sum_{j=1}^{I}\sum_{i=1}^{n_j}(x_{ij}-\bar{x})^2 = \sum_{j=1}^{I}\sum_{i=1}^{n_j}((\bar{x}_j-\bar{x})+(x_{ij}-\bar{x}_j))^2 \tag{5.36}$$

It is easy to show that:

$$\sum_{j=1}^{I}\sum_{i=1}^{n_j}(x_{ij}-\bar{x})^2 = \sum_{j=1}^{I}\sum_{i=1}^{n_j}(\bar{x}_j-\bar{x})^2 + \sum_{j=1}^{I}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2 \tag{5.37}$$

Exercise 8

Prove that:
$\sum\limits_{j=1}^{I}\sum\limits_{i=1}^{n_j}(x_{ij}-\bar{x})^2 = \sum\limits_{j=1}^{I}\sum\limits_{i=1}^{n_j}(\bar{x}_j-\bar{x})^2 + \sum\limits_{j=1}^{I}\sum\limits_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2$
Try proving the above before looking at the solution below.

Solution:

$$\begin{aligned}\sum_{j=1}^{I}\sum_{i=1}^{n_j}(x_{ij}-\bar{x})^2 &= \sum_{j=1}^{I}\sum_{i=1}^{n_j}((\bar{x}_j-\bar{x})+(x_{ij}-\bar{x}_j))^2 && (5.38)\\ &= \sum_{j=1}^{I}\sum_{i=1}^{n_j}((\bar{x}_j-\bar{x})^2+(x_{ij}-\bar{x}_j)^2+\underline{2(\bar{x}_j-\bar{x})(x_{ij}-\bar{x}_j))} && (5.39)\end{aligned}$$

It's enough to show that the underlined part = 0.

$$\sum_{j=1}^{I}\sum_{i=1}^{n_j}2(\bar{x}_j-\bar{x})(x_{ij}-\bar{x}_j) = \sum_{j=1}^{I}2(\bar{x}_j-\bar{x})\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j) \tag{5.40}$$

Notice that for any group j the following holds (can you say why?):

$$\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)=0. \tag{5.41}$$

5.6 The total sum of squares

What we've just established is that the total sum of squares (SS) is the sum of the SS between- and SS within-groups:

$$\sum_{j=1}^{I}\sum_{i=1}^{n_j}(x_{ij}-\bar{x})^2 = \sum_{j=1}^{I}\sum_{i=1}^{n_j}(\bar{x}_j-\bar{x})^2 + \sum_{j=1}^{I}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2 \tag{5.42}$$

$$SS_{\text{total}} = SS_{\text{between}} + SS_{\text{within}} \tag{5.43}$$

You will need this equation from now on, so it's important to know where it came from.

The really interesting thing here is that we can compute three different variances, **total**, **between**, and **within**. We compute these in the equations below. One question you will have at this point is, how do we get the denominators? The denominators are called degrees of freedom (recall the discussion of the t-distribution). The number of scores minus the number of parameters estimated (here, the number of means) gives you the degrees of freedom for each variance. The logic for this is identical to the reason why we have $n-1$ as a denominator for variance σ^2.[2]

$$s^2_{\text{total}} = \frac{\sum_{j=1}^{I}\sum_{i=1}^{n_j}(x_{ij}-\bar{x})^2}{N-1} \tag{5.44}$$

$$s^2_{\text{between}} = \frac{\sum_{j=1}^{I}\sum_{i=1}^{n_j}(\bar{x}_j-\bar{x})^2}{I-1} \tag{5.45}$$

$$s^2_{\text{within}} = \frac{\sum_{j=1}^{I}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2}{N-I} \tag{5.46}$$

These estimated variances have a special name: MEAN SQUARE.[3] So we can say:

$$MS_{\text{total}} = s_{\text{total}} = \frac{\sum_{j=1}^{I}\sum_{i=1}^{n_j}(x_{ij}-\bar{x})^2}{N-1} \tag{5.47}$$

$$MS_{\text{between}} = s_{\text{between}} = \frac{\sum_{j=1}^{I}\sum_{i=1}^{n_j}(\bar{x}_j-\bar{x})^2}{I-1} \tag{5.48}$$

$$MS_{\text{within}} = s_{\text{within}} = \frac{\sum_{j=1}^{I}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2}{N-I} \tag{5.49}$$

5.7 Hypothesis testing

The interesting thing is that we can compute an estimate of within-group variance ($\text{MS}_{\text{within}}$) and of between group variance ($\text{MS}_{\text{between}}$).

Recall our null and alternative hypotheses:

$$\mathcal{H}_0 : x_{ij} = \mu + \epsilon_{ij} \tag{5.50}$$

$$\mathcal{H}_a : x_{ij} = \mu + \alpha_j + \epsilon_{ij} \tag{5.51}$$

The null hypothesis amounts to saying that there is no effect of α_j: any between-group variance we see is attributable to within-group variance. In other words:

$$\mathcal{H}_0 : MS_{\text{between}} = MS_{\text{within}} \tag{5.52}$$

Alternatively, we can state the null hypothesis as a ratio:

$$\mathcal{H}_0 : \frac{MS_{\text{between}}}{MS_{\text{within}}} = 1 \tag{5.53}$$

[2] To remind you: The sum of deviations from mean is always zero, so if we know $n-1$ of the deviations, the last deviation is predictable. The mean is an average of n unrelated numbers, but s is an average of $n-1$ unrelated numbers.

[3] No idea why.

Recall that variance (hence, MS) is an unbiased estimator, so MS is an appropriate statistic: just like the sample mean "points" to the population mean, it "points" to the true population parameter (here, σ).

The ratio of MS's is our test statistic:

$$\frac{MS_{\text{between}}}{MS_{\text{within}}} = \text{F-statistic} \tag{5.54}$$

Now we need a distribution that tells us the probability that such a value of the test statistic (or something greater than it) could be obtained if the null hypothesis were true. The statistical distribution we need turns out to be a distribution derived from the quotient of two variances. It is called the F-distribution.

The t-distribution takes one parameter (degrees of freedom), the F-distribution takes two: the DF of the numerator and the DF of the denominator. So we refer to a given F-distribution as F(df1,df2), where df1 is the degrees of freedom of the between-group variance and df2 that of the within-group variance. In our current example, df1=3-1, df2=9-3. Let's see what the plot of F(2,6) looks like.

5.8 Generating an F-distribution

We return to the motivation for the F-distribution below, but for now let us focus on the MS values and the calculation of the F-value from them.

5.9 Computing the F value using MS square and MS within

We know how to compute MSbetween and MSwithin for any sample now, so we can compute the ratio of these two. First we will do this "by hand", and then let R do it for us. You will see exactly where the R output comes from.

First, let's create the data set (this is the first of the two datasets we looked at in the beginning of this chapter):

```
> scores <- c(9, 1, 2, 10, 2, 6, 1, 5, 0)
> subj <- paste("s", rep(c(1:9), 1), sep = "")
> group <- paste("g", rep(c(1:3), 1, each = 3), sep = "")
> data1 <- data.frame(scores, group, subj)

> g1data1 <- subset(data1, group == "g1")$scores
> g2data1 <- subset(data1, group == "g2")$scores
> g3data1 <- subset(data1, group == "g3")$scores
> SSwithin <- sum((mean(g1data1) - g1data1)^2) + sum((mean(g2data1) -
+     g2data1)^2) + sum((mean(g3data1) - g3data1)^2)
> Dfwithin <- 9 - 3
> (MSwithin <- SSwithin/Dfwithin)

[1] 14

> grandmean <- mean(data1$scores)
> SSbetween <- 3 * (mean(g1data1) - grandmean)^2 + 3 * (mean(g2data1) -
+     grandmean)^2 + 3 * (mean(g3data1) - grandmean)^2
> Dfbetween <- 3 - 1
> (MSbetween <- SSbetween/Dfbetween)
> x <- seq(c(1:100, by = 0.005))
> plot(density(rf(10000, 2, 6)), xlim = range(0, 5), xlab = "",
+     main = "An F-distribution with parameters 2 and 6.")
```

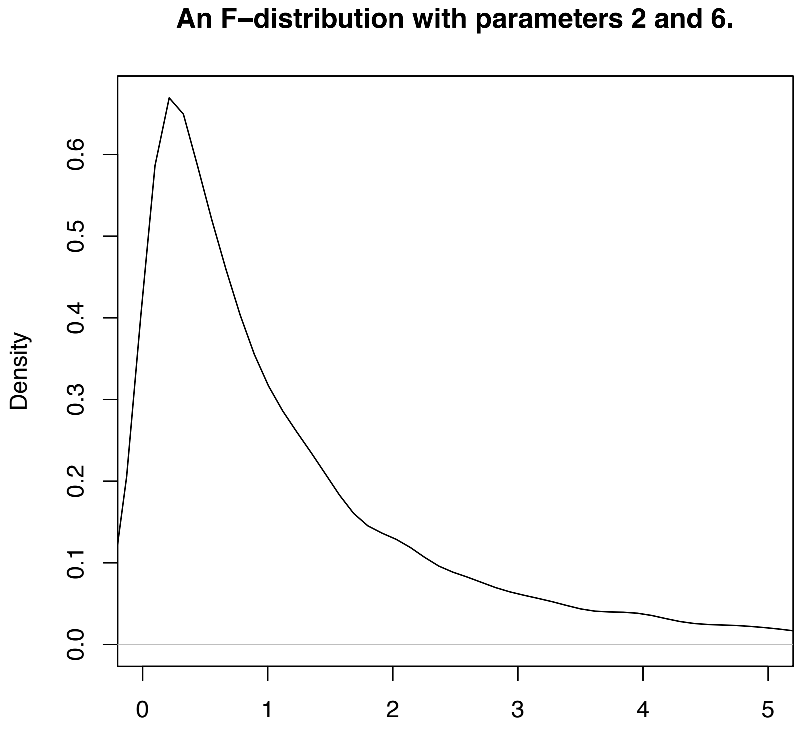


Figure 5.2: An F-distribution.

```
[1] 12
```

The critical thing to notice is that we computed the MSbetween and MSwithin, with 2 and 6 degrees of freedom respectively. Now we can compute the F-value by computing the ratio:

```
> (MSbetween/MSwithin)

[1] 0.8571429
```

And now, we take a look at the R output for ANOVA:

```
> aov.fm <- aov(scores ~ group + Error(subj), data1)
> summary(aov.fm)

Error: subj
          Df Sum Sq Mean Sq F value Pr(>F)
group      2     24      12  0.8571 0.4705
Residuals  6     84      14
```

It should come as no great surprise that we have exactly the same MSbetween. MSwithin, and F-value. As a bonus, R also gives us the p-value.

5.10 ANOVA as a linear model

You might at this point ask: can't we do the ANOVA based on our original null and alternative hypotheses. Just to remind you what these were:

$$\mathcal{H}_0 : x_{ij} = \mu + \epsilon_{ij} \tag{5.55}$$

$$\mathcal{H}_a : x_{ij} = \mu + \alpha_j + \epsilon_{ij} \tag{5.56}$$

The answer is: the MS-between and MS-within method we used is identical to the ANOVA based on the linear models above.

Take a look at this output:

```
> (aov.fm)

Call:
aov(formula = scores ~ group + Error(subj), data = data1)

Grand Mean: 4

Stratum 1: subj

Terms:
                group Residuals
Sum of Squares     24        84
Deg. of Freedom     2         6

Residual standard error: 3.741657
Estimated effects may be unbalanced
```

Towards the beginning the ANOVA output tells us that the formula is for the calculations:

```
Call:
aov(formula = scores ~ group + Error(subj), data = data1)
```

This is almost literally the alternative hypothesis as a system of linear equations. The only thing missing in the formula above is the term for the grand mean.

$$\mathcal{H}_a : x_{ij} = \mu + \alpha_j + \epsilon_{ij} \tag{5.57}$$

But the alternative hypothesis above is what R is actually using for computation. To see this, let's unpack ANOVA in R further. In reality, the ANOVA call in R is actually doing its computations based on a bunch of linear equations, one for each subject. Let's see if we can squeeze this information out of R.

First we are going to fit a linear model (with the function `lm`), and then examine the underlying equations. The code you see below seems obscure at first, but all will become clear in a moment.

```
> lm.fm <- lm(scores ~ group, data = data1)
> (mm.fm <- model.matrix(lm.fm))

  (Intercept) groupg2 groupg3
1           1       0       0
2           1       0       0
3           1       0       0
4           1       1       0
5           1       1       0
6           1       1       0
7           1       0       1
8           1       0       1
9           1       0       1
attr(,"assign")
[1] 0 1 1
attr(,"contrasts")
attr(,"contrasts")$group
[1] "contr.treatment"

> (cf.fm <- coefficients(lm.fm))

(Intercept)     groupg2     groupg3
          4           2          -2

> (res.fm <- residuals(lm.fm))

 1  2  3  4  5  6  7  8  9
 5 -3 -2  4 -4  0 -1  3 -2

> (cbind(mm.fm, res.fm))

  (Intercept) groupg2 groupg3 res.fm
1           1       0       0      5
2           1       0       0     -3
3           1       0       0     -2
4           1       1       0      4
5           1       1       0     -4
```

```
6             1        1        0      0
7             1        0        1     -1
8             1        0        1      3
9             1        0        1     -2
```

The matrix generated by the last call above shows almost all the terms of the nine equations corresponding to each subject. See if you can discover the relationship between cf.fm, mm.fm, and res.fm. To help you along, we provide the answer below:

$$9 = 4 \times 1 + 2 \times 0 + -2 \times 0 + 5.000000e + 00 \tag{5.58}$$
$$1 = 4 \times 1 + 2 \times 0 + -2 \times 0 + -3.000000e + 00 \tag{5.59}$$
$$2 = 4 \times 1 + 2 \times 0 + -2 \times 0 + -2.000000e + 00 \tag{5.60}$$
$$10 = 4 \times 1 + 2 \times 1 + -2 \times 0 + 4.000000e + 00 \tag{5.61}$$
$$2 = 4 \times 1 + 2 \times 1 + -2 \times 0 + -4.000000e + 00 \tag{5.62}$$
$$6 = 4 \times 1 + 2 \times 1 + -2 \times 0 + 1.526557e - 16 \tag{5.63}$$
$$1 = 4 \times 1 + 2 \times 0 + -2 \times 1 + -1.000000e + 00 \tag{5.64}$$
$$5 = 4 \times 1 + 2 \times 0 + -2 \times 1 + 3.000000e + 00 \tag{5.65}$$
$$0 = 4 \times 1 + 2 \times 0 + -2 \times 1 + -2.000000e + 00 \tag{5.66}$$

Each of these equations gives you the observed score of each subject as a function of the grand mean, the effect of each factor, and the residual error due to the subject in question. These equations are the exploded form of the compact one we saw earlier:

$$x_{ij} = \mu + \alpha_j + \epsilon_{ij} \tag{5.67}$$

It is probably not obvious what the connection between the system of equations above and this compact-form equation is. To see the connection, we can restate the system of equations above as a giant matrix-based equation:

$$Y_i = \beta_0 X_{0_i} + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \epsilon_i \tag{5.68}$$

- Y_i is the matrix containing the scores of all the 9 subjects.

  ```
  > (data1$scores)
  ```

  ```
  [1]  9  1  2 10  2  6  1  5  0
  ```

- X_{0_i} is the intercept column of 1s in the mm.fm matrix.

  ```
  > (mm.fm[, 1])
  ```

  ```
  1 2 3 4 5 6 7 8 9
  1 1 1 1 1 1 1 1 1
  ```

- X_{1_i} and X_{2_i} are dummy variables that help us code each subject as being in group 1, 2, or 3 (see below for an explanation).

  ```
  > (mm.fm[, 2:3])
  ```

```
  groupg2 groupg3
1       0       0
2       0       0
3       0       0
4       1       0
5       1       0
6       1       0
7       0       1
8       0       1
9       0       1
```

- β_0 is the grand mean (see the first element of cf.fm).

```
> (mm.fm[, 1])

1 2 3 4 5 6 7 8 9
1 1 1 1 1 1 1 1 1
```

- β_1 is the effect of group 2 (see the second element of cf.fm).

```
> (cf.fm[2])

groupg2
      2
```

- β_2 is the effect of group 3 (see the third element of cf.fm).

```
> (cf.fm[3])

groupg3
     -2
```

- Exercise: What is ϵ_i?

Note how the three α components (each corresponding to one of the three groups) are expressed in the system of linear equations above. To figure this out, look at the model matrix output from the linear model once again:

```
> (mm.fm)[, 2:3]

  groupg2 groupg3
1       0       0
2       0       0
3       0       0
4       1       0
5       1       0
6       1       0
7       0       1
8       0       1
9       0       1
```

Notice that the second and third columns uniquely classify each of the 9 rows as corresponding to subjects 1-9. Subject 1 has groupg2=0, and groupg3=0, same for subjects 2 and 3: i.e. these subjects are neither in group 2 or 3, that is, they are in group 1. And so on.

This kind of coding of the α component is called dummy coding.

You can ask R to compute an ANOVA using this linear model. Compare the output of the anova function (which uses the lm output) with the earlier anova we had found using the aov function:

```
> (anova(lm.fm))

Analysis of Variance Table

Response: scores
          Df Sum Sq Mean Sq F value Pr(>F)
group      2     24      12  0.8571 0.4705
Residuals  6     84      14

> summary(aov.fm)

Error: subj
          Df Sum Sq Mean Sq F value Pr(>F)
group      2     24      12  0.8571 0.4705
Residuals  6     84      14
```

Exercise 9

Compute the mean squares and F value for the second dataset in the same manner (by hand and using ANOVA). Isolate the underlying linear equations as we did above and identify all the coefficients and terms in the general form of the equation:

$$Y_i = \beta_0 X_{0_i} + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \epsilon_i \tag{5.69}$$

Here's the data to get you started off:

```
> scores <- c(3, 4, 5, 7, 6, 5, 1, 2, 3)
> subj <- paste("s", rep(c(1:9), 1), sep = "")
> group <- paste("g", rep(c(1:3), 1, each = 3), sep = "")
> data2 <- data.frame(scores, group, subj)
```

After that exciting interlude, we now return to thinking a bit more about the F value and the F-distribution. This distribution seemed to come out of thin air, but there is a principled motivation for it. Let's figure out the motivation for the F-distribution.

5.11 Motivation for the F-distribution

Even though there's a distribution involved as in the t-test, the ANOVA seems to have a rather different logic compared to z- and t-tests. The F-distribution seems to come out of nowhere. But actually the logic is identical to the z- and t-tests. Let's look at this logic next, using our usual tool: simulations.

5.12 A first attempt

Let's start all over again, using the same logic as in t-tests: To compare three (or more means), we need to find an UNBIASED ESTIMATOR, a STATISTIC for a corresponding population parameter. In the past we used the mean (single sample case), and d (the difference between means) for the two sample case: $\mathcal{H}_0 : \mu_1 - \mu_2 = 0 = \delta$.

We take a stab at this and ask: maybe the **variance of the sample means** would suffice. It satisfies some nice-looking properties: it gets bigger as the three+ sample means get spread further apart, and is equal to 0 if all the means are the same; just like the difference of the means d which is an unbiased estimator of δ.

But is the variance of the means an unbiased estimator of the population parameter? The sampling distribution of the sample mean variances is shown in Figure 5.3.

5.13 A second attempt

This failure to use variance of sample means as our test statistic motivates our real test statistic: the ratio of between- and within-group variances.

Let's study the sampling distribution of MS-between and MS-within.

5.13.1 A second attempt : MS within, three identical populations

```
> pop1 <- rnorm(1000, mean = 60, sd = 4)
> pop2 <- rnorm(1000, mean = 60, sd = 4)
> pop3 <- rnorm(1000, mean = 60, sd = 4)
> ss <- function(sample) {
+     m <- rep(mean(sample), length(sample))
+     m2 <- (sample - m)^2
+     result <- sum(m2)
+     result
+ }
> mswithin <- function(s1, s2, s3) {
+     N <- sum(length(s1), length(s2), length(s3))
+     DF <- N - 3
+     msw <- (ss(s1) + ss(s2) + ss(s3))/DF
+     msw
+ }
```

```
> mswithins <- rep(NA, 1000)
> for (i in c(1:1000)) {
+     s1 <- sample(pop1, 11)
+     s2 <- sample(pop2, 15)
+     s3 <- sample(pop3, 20)
+     mswithins[i] <- mswithin(s1, s2, s3)
+ }
> plot(density(mswithins))
> pop1 <- rnorm(1000, mean = 60, sd = 4)
> pop2 <- rnorm(1000, mean = 62, sd = 4)
> pop3 <- rnorm(1000, mean = 64, sd = 4)
> variances <- rep(NA, 1000)
> for (i in c(1:1000)) {
+     s1 <- sample(pop1, 11)
+     s2 <- sample(pop2, 11)
+     s3 <- sample(pop3, 11)
+     means1 <- mean(s1)
+     means2 <- mean(s2)
+     means3 <- mean(s3)
+     variances[i] <- var(c(means1, means2, means3))
+ }
> meanvar <- mean(variances)
> plot(density(variances), main = "", xlab = "")
```

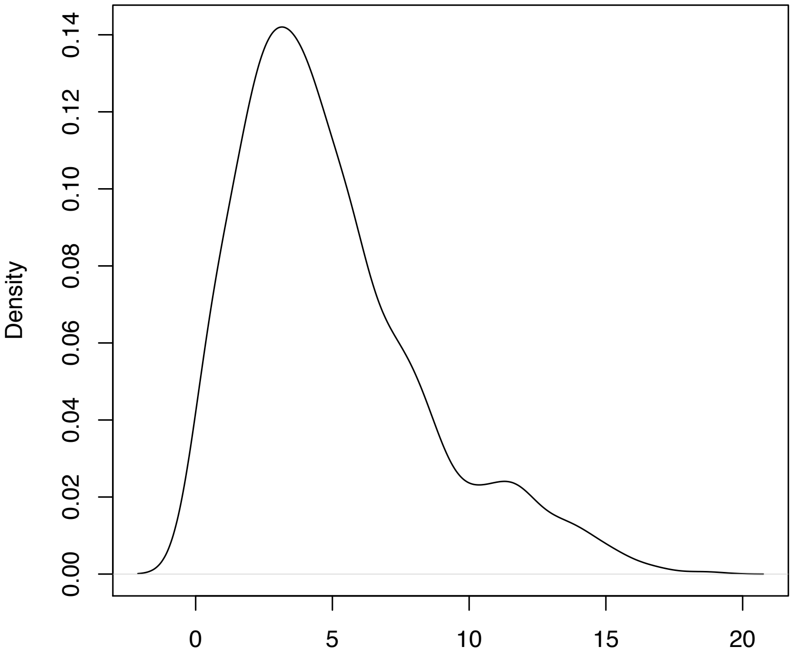
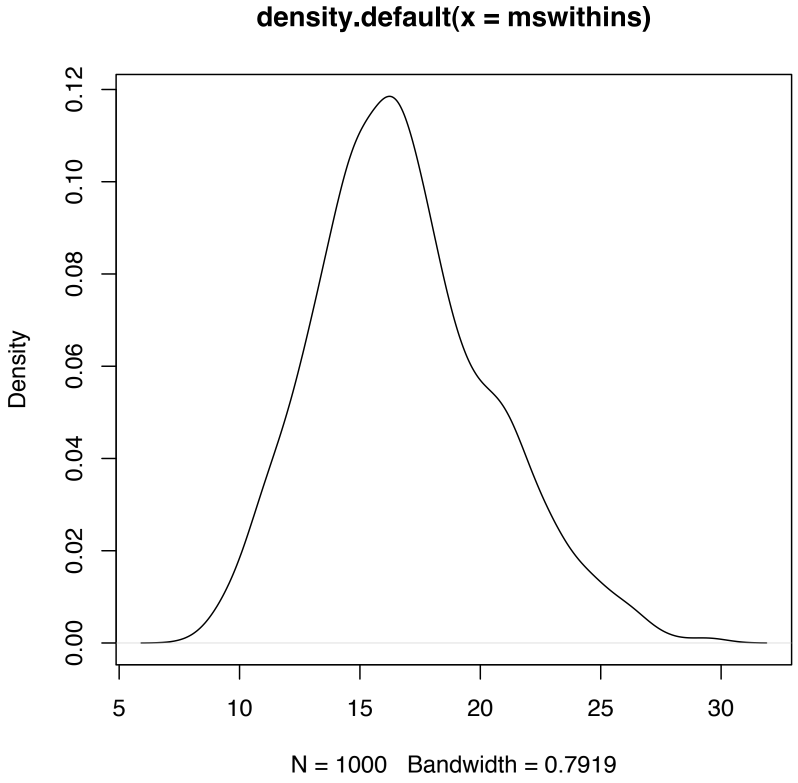


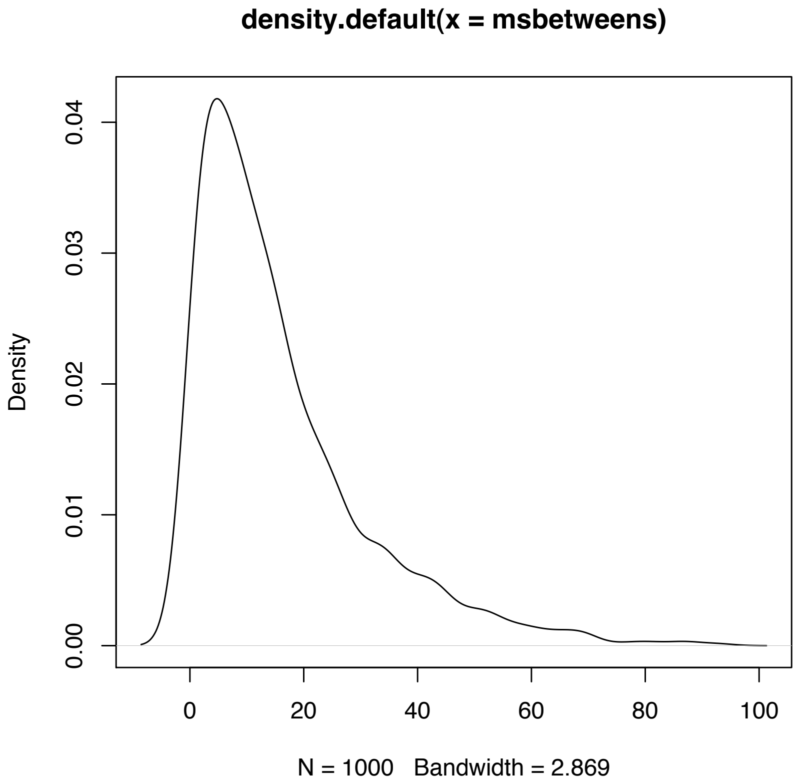
Figure 5.3: The sampling distribution of the sample mean variances.



This is hardly surprising. Standard deviation is an unbiased estimator. Here, we are just taking the variance of each of the three samples, adding them up and dividing by $N - I$: we're pooling variances to get an estimate of the population variance. A single sample's SD is an unbiased estimator, so it's no surprise that the pooled variance here is also an unbiased estimator.

$$MS_{\text{between}} = \frac{\sum\limits_{j=1}^{I}\sum\limits_{i=1}^{n_j}(\bar{x}_j - \bar{x})^2}{I-1} \tag{5.70}$$

```
> msbetween <- function(s1, s2, s3) {
+     gm <- mean(c(s1, s2, s3))
+     m1 <- mean(s1)
+     m2 <- mean(s2)
+     m3 <- mean(s3)
+     msb <- (length(s1) * (m1 - gm)^2 + length(s2) * (m2 - gm)^2 +
+         length(s3) * (m3 - gm)^2)/2
+     return(msb)
+ }
> msbetweens <- rep(NA, 1000)
> for (i in c(1:1000)) {
+     s1 <- sample(pop1, 11)
+     s2 <- sample(pop2, 15)
+     s3 <- sample(pop3, 20)
+     msbetweens[i] <- msbetween(s1, s2, s3)
+ }
> plot(density(msbetweens))
```



When the three populations have the same mean (60) and the same variance (16), the mean (or center) of the sampling distribution of the MS-within (15.7) points to the population variance, and the mean of the sampling distribution of the MS-between (16.42) also points to the population variance.

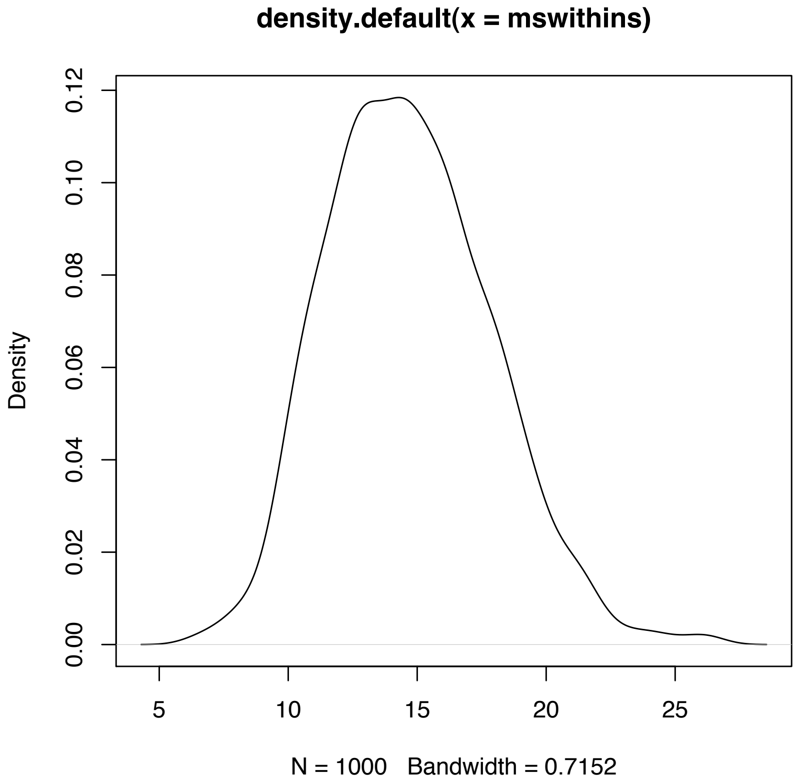
The key idea of ANOVA is this: when the populations' means are not different, the means of these two distributions are very close to the population variance (assuming, as we are, that the populations' variances are identical).

When the null hypothesis is true (population means are not different), these two statistics (MS-between, MS-within) are unbiased estimators of the same number – the population variance. (Keep in mind that the three populations' variances were assumed to be equal in the simulations.)

5.13.2 A second attempt: MS within, three non-identical populations

Suppose we have three populations with different means, same SDs (the null hypothesis is now in fact false).

```
> pop1 <- rnorm(1000, mean = 60, sd = 4)
> pop2 <- rnorm(1000, mean = 62, sd = 4)
> pop3 <- rnorm(1000, mean = 64, sd = 4)
> mswithins <- rep(NA, 1000)
> for (i in c(1:1000)) {
+     s1 <- sample(pop1, 11)
+     s2 <- sample(pop2, 15)
+     s3 <- sample(pop3, 20)
+     mswithins[i] <- mswithin(s1, s2, s3)
+ }
> plot(density(mswithins))
```



Why didn't the center change? Reflect upon MS-within for a minute:

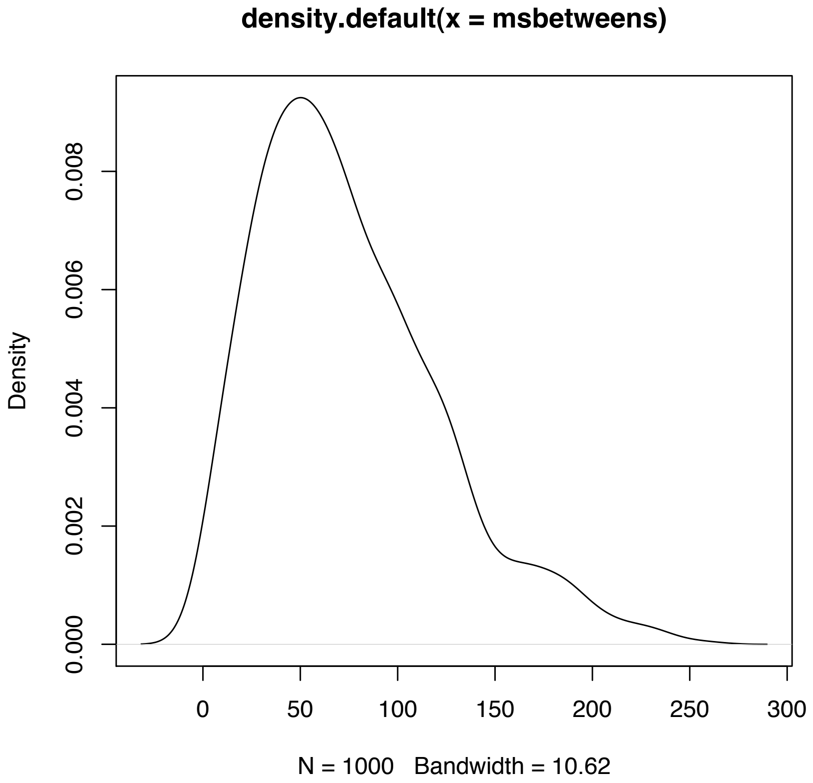
$$MS_{\text{within}} = \frac{\sum_{j=1}^{I}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2}{N - I} \tag{5.71}$$

MS-within is computing the spread about the mean in each sample, the location of the mean in that sample is irrelevant. As long as the populations spreads (variances) remain identical, MS-within will always estimate this variance in an unbiased manner.

So, MS-within is an invariant reference number for a comparison of populations with the same variances. Now let's look at how MS-*between* behaves with non-identical means in the three populations.

Suppose we have three populations with different means, same SDs (the null hypothesis is now in fact false).

```
> pop1 <- rnorm(1000, mean = 60, sd = 4)
> pop2 <- rnorm(1000, mean = 62, sd = 4)
> pop3 <- rnorm(1000, mean = 64, sd = 4)
> msbetweens <- rep(NA, 1000)
> for (i in c(1:1000)) {
+     s1 <- sample(pop1, 11)
+     s2 <- sample(pop2, 15)
+     s3 <- sample(pop3, 20)
+     msbetweens[i] <- msbetween(s1, s2, s3)
+ }
> plot(density(msbetweens))
```



5.14 MS-between and MS-within

The mean of the sampling distribution of MS-within is an invariant reference number for a comparison of populations with the same variances. The mean of the sampling distribution of MS-between is near identical to that of MS-within if the null hypothesis is true (identical population means). I.e., $\frac{MS_{\text{between}}}{MS_{\text{within}}} \approxeq 1$ if $\mathcal{H}_0 = T$. If the null hypothesis is in fact false (if the population means differ), then it's highly likely that MS-between > MS-within.

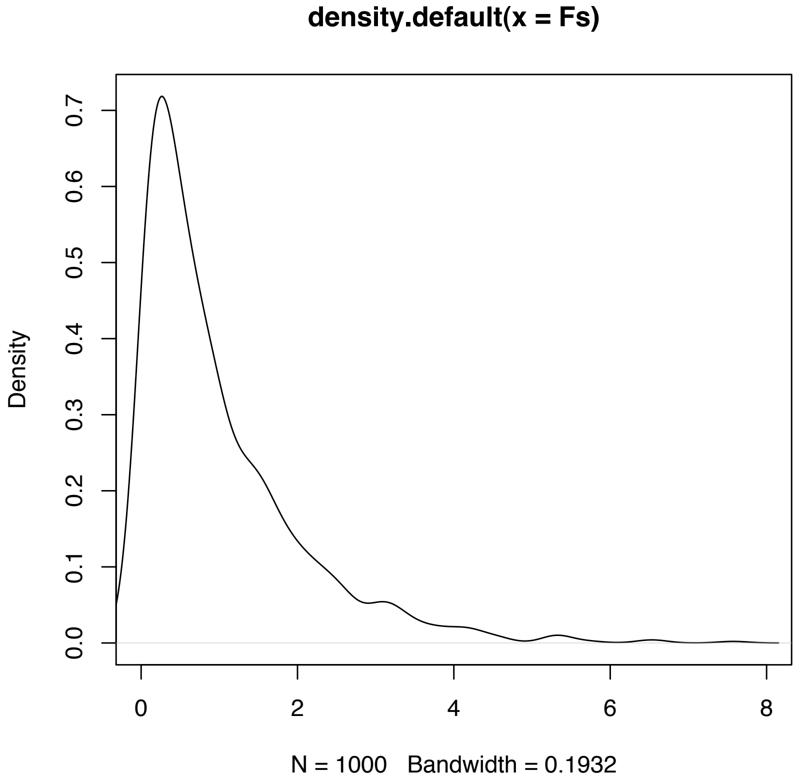
When population means actually differ, for a *given* sample it's theoretically possible that MS-between is lower (close to populations' variance) and that MS-within is higher than the populations' variances. . . But, *because of the shapes of sampling distributions we just saw*, the likelihood of each of these events happening is low, and therefore the co-occurrence of both these events is even less likely.

5.15 In search of a test statistic

We have two unbiased estimators, the statistics MS-between and MS-within. We know that MS-within never varies. We know that MS-between does, depending on whether the null hypothesis is in fact true or not. It turns out that the ratio of these gives a good test statistic; let's convince ourselves that this is true.

5.16 The F-distribution: identical populations

```
> Fratio <- function(msbetween, mswithin) {
+     Fvalue <- msbetween/mswithin
+     return(Fvalue)
+ }
> pop1 <- rnorm(1000, mean = 60, sd = 4)
> pop2 <- rnorm(1000, mean = 60, sd = 4)
> pop3 <- rnorm(1000, mean = 60, sd = 4)
> Fs <- rep(NA, 1000)
> for (i in c(1:1000)) {
+     s1 <- sample(pop1, 11)
+     s2 <- sample(pop2, 15)
+     s3 <- sample(pop3, 20)
+     Fs[i] <- msbetween(s1, s2, s3)/mswithin(s1, s2, s3)
+ }
> plot(density(Fs), xlim = range(0, 8))
```



5.17 Inference with the F-distribution

Assuming that the null hypothesis is true, we can construct an F-distribution—we just did this from first principles. If the null hypothesis is true, the distribution is centred around 1.

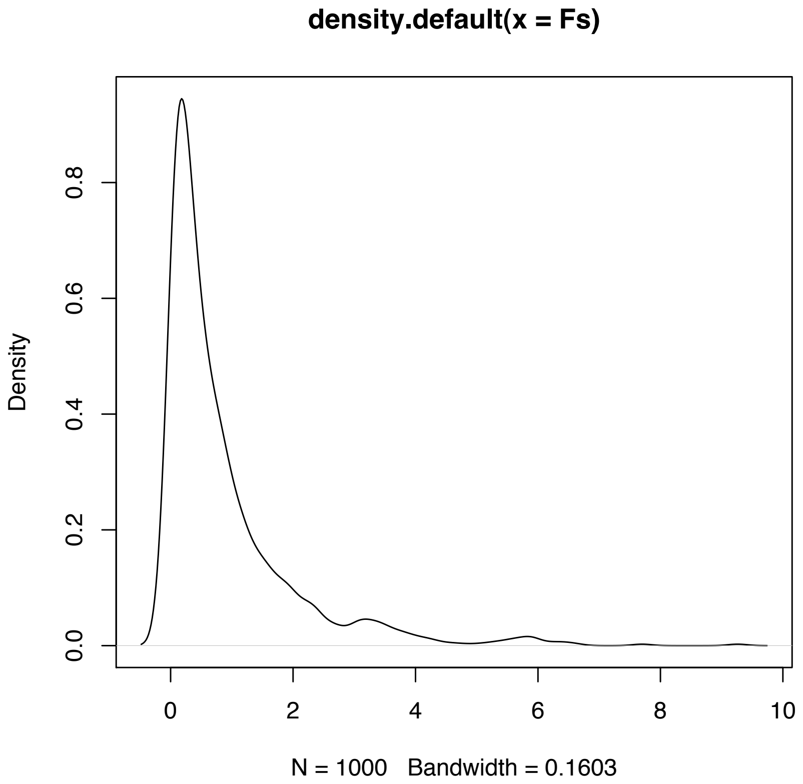
Given our *sample*'s F-ratio, we can now ask (as in the t-test): what's the likelihood of getting an F-ratio as far from 1 as it is, or further, *assuming that the null hypothesis is true*? That likelihood is what the P-value is giving you in the ANOVA test; just like in the t-test.

5.18 The F-ratio, three populations with wildly different σ, but identical means

```
> pop1 <- rnorm(1000, mean = 60, sd = 2)
> pop2 <- rnorm(1000, mean = 60, sd = 5)
> pop3 <- rnorm(1000, mean = 60, sd = 10)
> Fs <- rep(NA, 1000)
> for (i in c(1:1000)) {
+     s1 <- sample(pop1, 11)
+     s2 <- sample(pop2, 15)
+     s3 <- sample(pop3, 20)
+     Fs[i] <- msbetween(s1, s2, s3)/mswithin(s1, s2, s3)
+ }
```

The F-ratio, three populations with wildly different σ, but identical means

```
> plot(density(Fs))
```



So if you have different population variances and get disagreeing numbers for MS-between and MS-within, you can't conclude it's due to different population means. It could just be due to different variances.

Rule of thumb: if the largest s is less than twice the smallest s in a multiple population comparison, we can safely use ANOVA. The greater the divergence, the smaller the P-value we would hope for.

The F-ratio, three populations with wildly different σ, but identical means

Chapter 6

Bivariate statistics

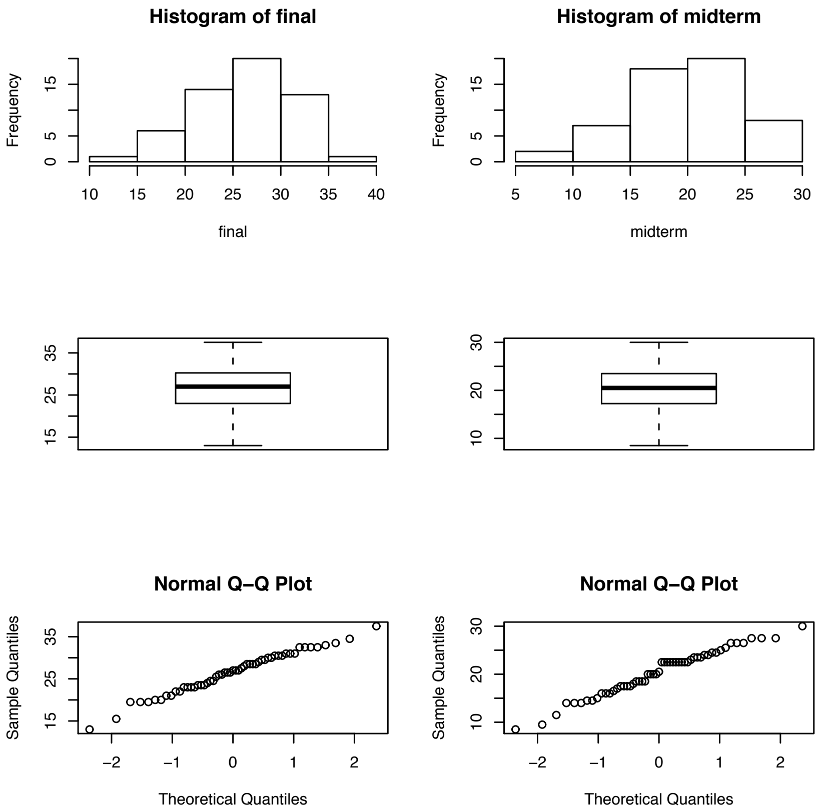
So far we've been studying univariate statistics: One population, one mean and SD; Or two populations, two means and SDs, etc. Now we consider the scenario where, **for each individual** in a population, we have two values.

For example, here are some (allegedly real) homework, midterm, and final exam scores from a statistics course (Faraway's dataset `stat500`). You should at this point make a text file containing this data.

```
   midterm final   hw total
1     24.5  26.0 28.5  79.0
2     22.5  24.5 28.2  75.2
3     23.5  26.5 28.3  78.3
4     23.5  34.5 29.2  87.2
[and so on]
```

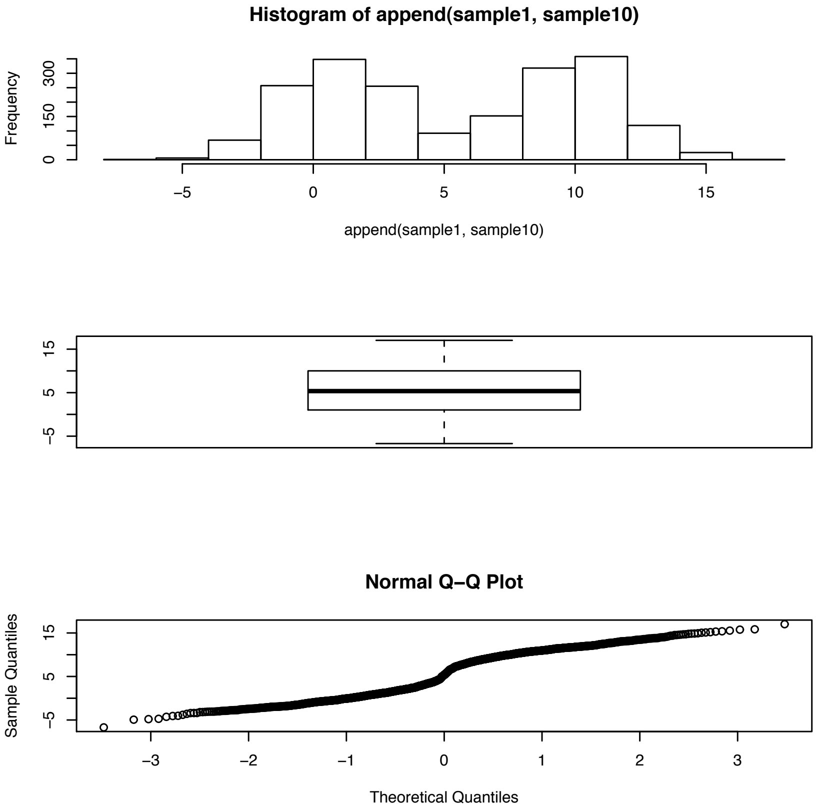
Research question: can we predict the performance of students on the final exam from their midterm scores? First, let's take a look at the distribution of these two scores, using hist, boxplot, and qqplot.

```
> library(faraway)
> data(stat500)
> attach(stat500)
> x <- mean(midterm)
> sdx <- sd(midterm)
> y <- mean(final)
> sdy <- sd(final)
> op <- par(mfrow = c(3, 2))
> hist(final)
> hist(midterm)
> boxplot(final)
> boxplot(midterm)
> qqnorm(final)
> qqnorm(midterm)
```



In this example, the distributions are approximately normal. But sometimes two distinct populations can be present in the sample. Suppose we have two normal populations, one with $\mu_1 = 1$ and the other with $\mu_2 = 10$. Suppose also that the data we have contains samples from both these populations. What would the distribution of the mixed-up sample look like?

```
> sample1 <- rnorm(1000, mean = 1, sd = 2)
> sample10 <- rnorm(1000, mean = 10, sd = 2)
> op <- par(mfrow = c(3, 1))
> hist(append(sample1, sample10))
> boxplot(append(sample1, sample10))
> qqnorm(append(sample1, sample10))
```



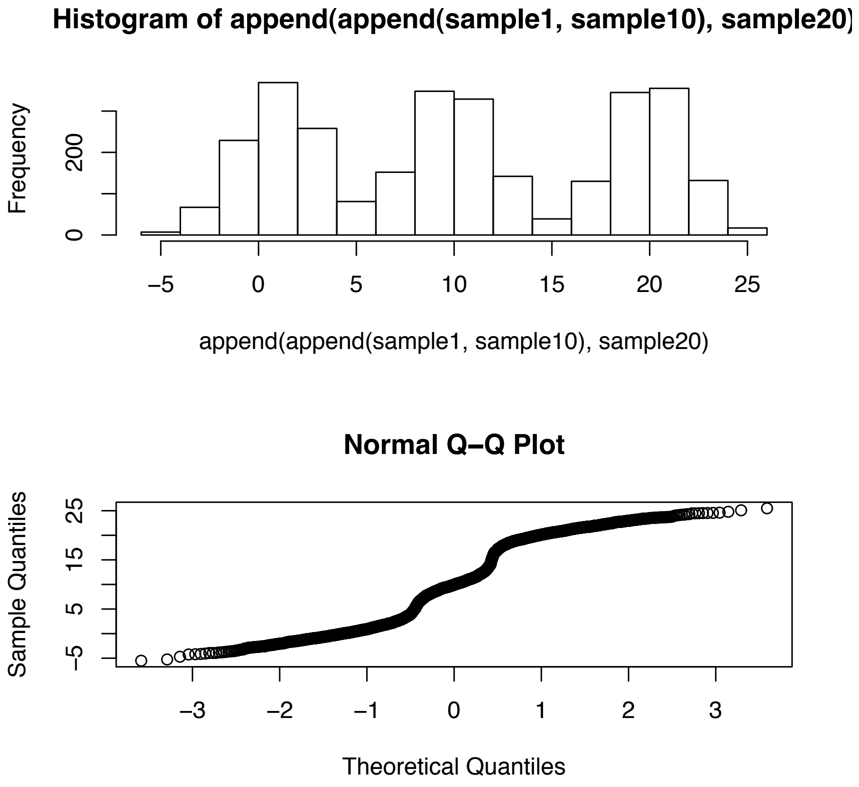
Now consider two cases:

- Case 1: we have three normal populations, one with $\mu_1 = 1$, $\mu_2 = 10$, $\mu_3 = 20$.

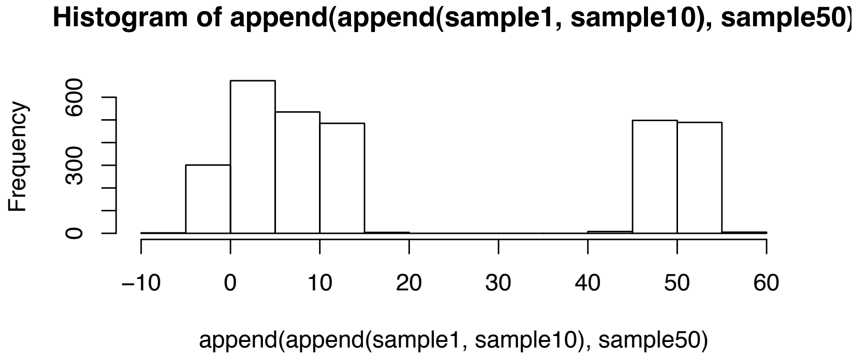
- Case 2: we have three normal populations, one with $\mu_1 = 1$, $\mu_2 = 10$, $\mu_3 = 50$.

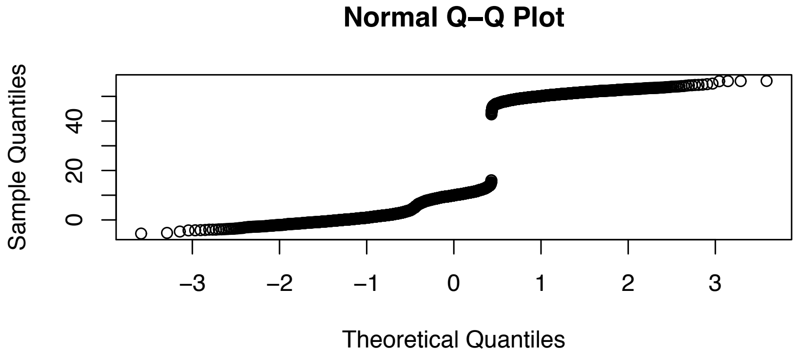
- What do the histograms and q-q plots look like in each case?

```
> sample50 <- rnorm(1000, mean = 50, sd = 2)
> sample20 <- rnorm(1000, mean = 20, sd = 2)
> op <- par(mfrow = c(2, 1))
> hist(append(append(sample1, sample10), sample20))
> qqnorm(append(append(sample1, sample10), sample20))
```



```
> op <- par(mfrow = c(2, 1))
> hist(append(append(sample1, sample10), sample50))
> qqnorm(append(append(sample1, sample10), sample50))
```



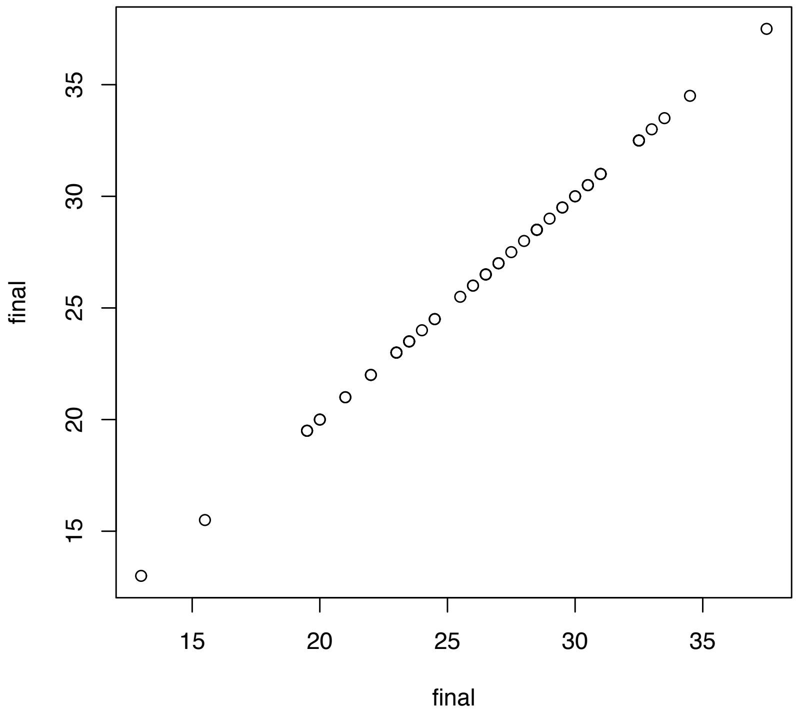


The point here is that one should explore your data before analyzing it using statistical tests. The logic underlying hypothesis testing assumes approximately normal residuals (more on this later). If this assumption is seriously compromised in the data, this can be a problem. What to do when we have non-normal distributions? We'll be adding a chapter on this in the book in the near future.

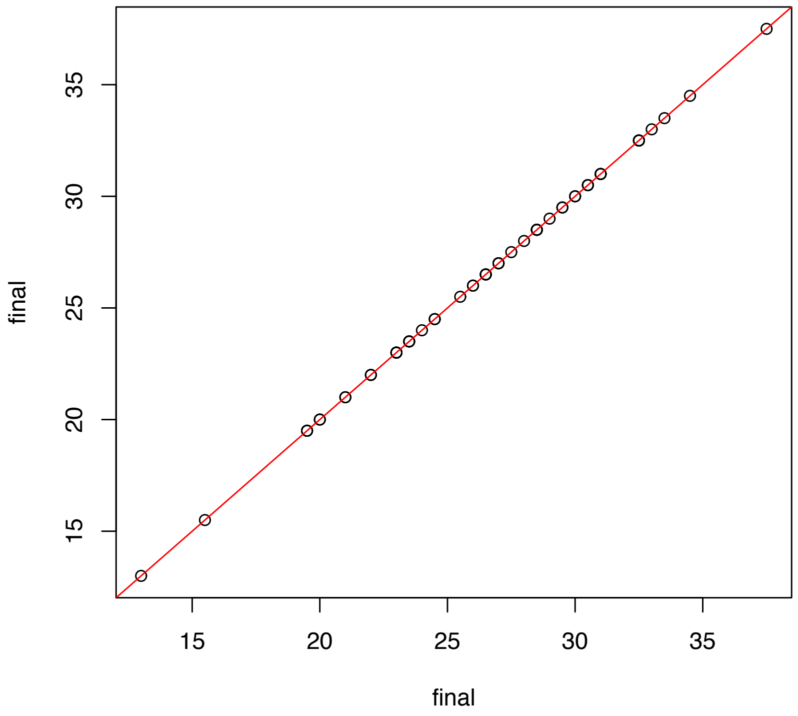
Back to the bivariate example. The research question is: can we predict the performance of students on the final exam from their midterm scores? Consider first a trivial variant of such a research question: can we predict the performance of students on the final exam from their **final exam** scores?

Trivial prediction:

```
> final2 <- final
> plot(final ~ final2, xlab = "final")
```

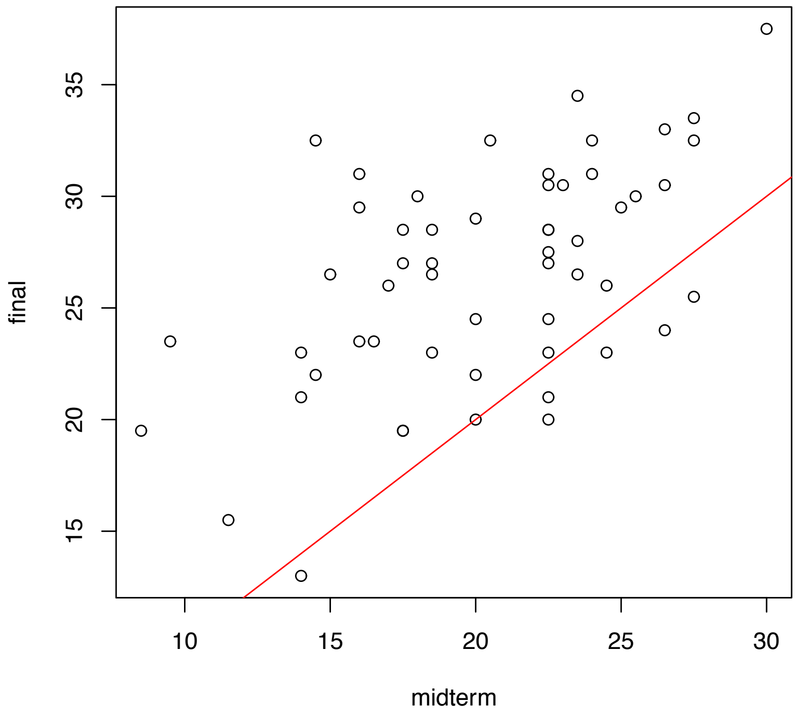


```
> final2 <- final
> plot(final ~ final2, xlab = "final")
> abline(0, 1, col = "red")
```



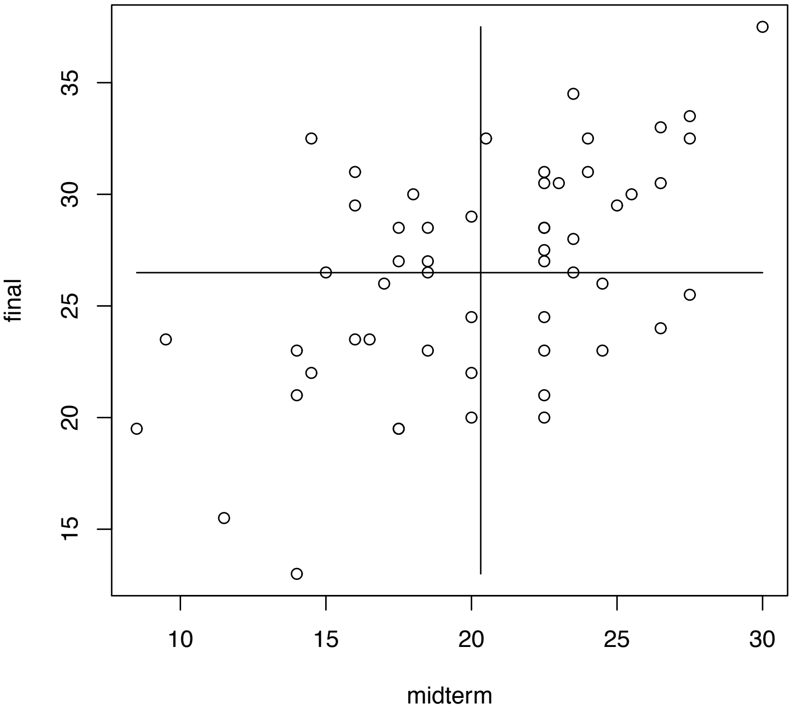
The real issue is however whether the midterm scores can predict the final scores:

```
> plot(final ~ midterm)
> abline(0, 1, col = "red")
```



The distributions and their means:

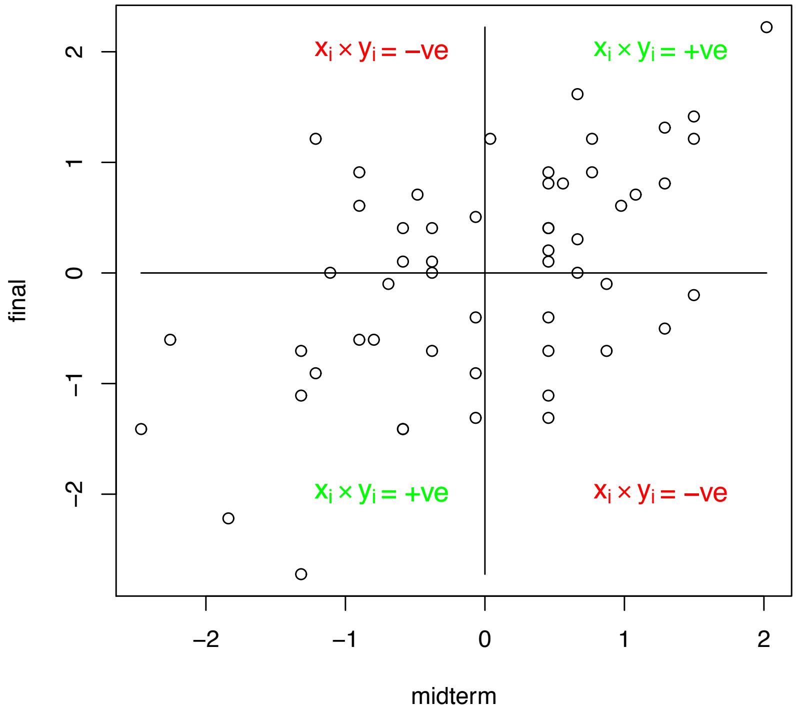
```
> plot(final ~ midterm)
> arrows(x, min(final), x, max(final), code = 0)
> arrows(min(midterm), y, max(midterm), y, code = 0)
> detach(stat500)
```



6.1 Summarizing a bivariate distribution

In order to predict the finals score from the midterm, we need to somehow summarize the relationship between the midterm and finals scores. One easy way to do this is to say: how linear is the relationship? To define this linearity, let's start by standardizing the final and midterm values, and then plot them.

```
> scaledstat500 <- data.frame(scale(stat500))
> attach(scaledstat500)
> plot(final ~ midterm)
> arrows(mean(midterm), min(final), mean(midterm), max(final),
+     code = 0)
> arrows(min(midterm), mean(final), max(midterm), mean(final),
+     code = 0)
> text(1, 2, labels = expression(x[i] %*% y[i]), cex = 1.2, col = "green")
> text(1.5, 2, labels = c("= +ve"), cex = 1.2, col = "green")
> text(-1, -2, labels = expression(x[i] %*% y[i]), cex = 1.2, col = "green")
> text(-0.5, -2, labels = c("= +ve"), cex = 1.2, col = "green")
> text(1, -2, labels = expression(x[i] %*% y[i]), cex = 1.2, col = "red")
> text(1.5, -2, labels = c("= -ve"), cex = 1.2, col = "red")
> text(-1, -2, labels = expression(x[i] %*% y[i]), cex = 1.2, col = "green")
> text(-0.5, -2, labels = c("= +ve"), cex = 1.2, col = "green")
> text(-1, 2, labels = expression(x[i] %*% y[i]), cex = 1.2, col = "red")
> text(-0.5, 2, labels = c("= -ve"), cex = 1.2, col = "red")
```



6.2 The correlation coefficient

Now if we multiply and sum each x-y pair of values in each quadrant, the grand sum of all these sums will be positive just in case more points lie in the first and third (positive) quadrants than in the other two (and there are no major "outliers" in the second and fourth quadrants).

Definition: Correlation r is defined as:

$$r = \frac{\sum_{i=1}^{n} (z_{x_i} \times z_{y_i})}{n-1} \tag{6.1}$$

(where z_{x_i} refers to the z-score of x_i, etc.)

Quick sanity check (note: final and midterm vectors refer here to the z-scores):

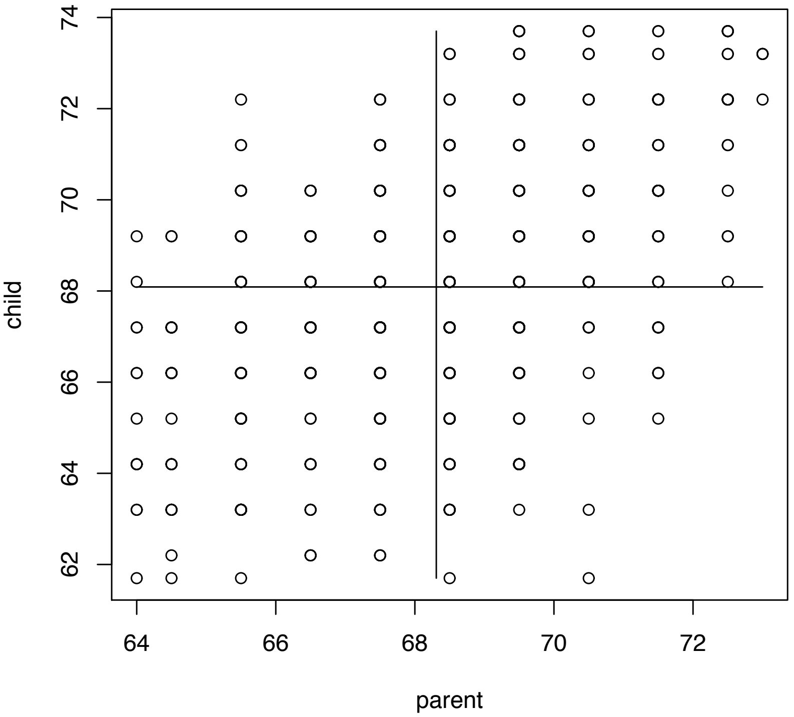
```
> sum(final * midterm)/(length(final) - 1)

[1] 0.5452277

> cor(midterm, final)

[1] 0.5452277
```

The positive correlation is telling us that the majority of the x-y pairs are located in the first and third quadrants. So we can say, roughly, that the higher the midterm score, the higher the final score is likely to be.

Here's a more subtle question: how much higher is an above-average final score for an above-average midterm score? I.e., how much higher is the final score for a higher midterm score? A possible answer: 1:1 ratio. This was a more relevant question in the original context that it was asked: Galton's parent-son data.

6.3 Galton's question

```
> library(UsingR)
> data(galton)
> attach(galton)
> gx <- mean(parent)
> gsdx <- sd(parent)
> gy <- mean(child)
> gsdy <- sd(child)
> plot(child ~ parent)
> arrows(gx, min(child), gx, max(child), code = 0)
> arrows(min(parent), gy, max(parent), gy, code = 0)
> detach(galton)
```



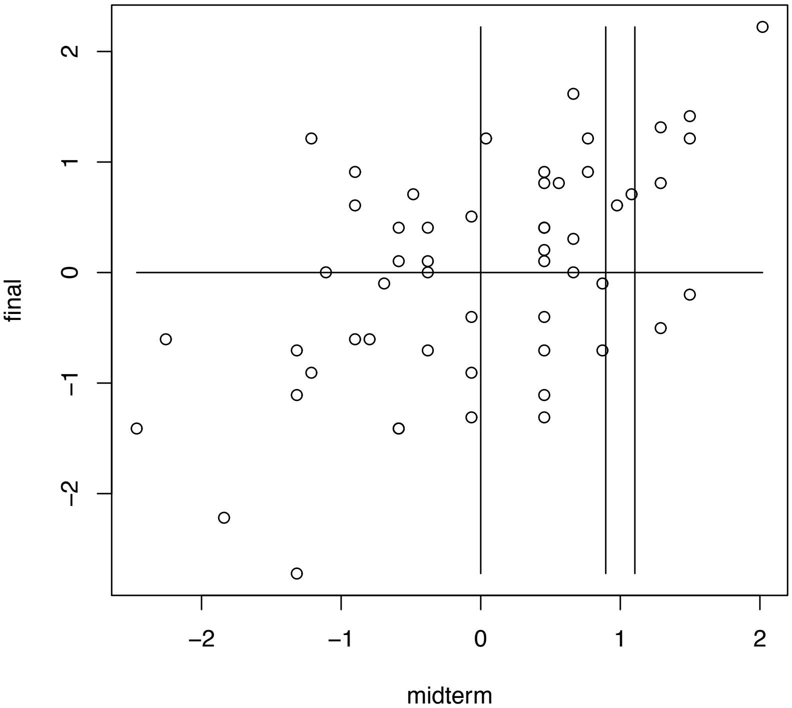
6.4 Regression

Theoretically, populations could diverge from, stay constant with respect to, or approach or *regress* to the mean. Galton found that they regress to the mean – hence the term regression.

Let's get back to the midterm-finals example to see how we can establish what happens in that case. To anticipate the results a bit, we'll find that here too the finals score regresses to the mean. The notion of "regressing" to the mean is less meaningful here – it's just a historically determined term we have to live with.

6.4.1 One SD above midterm means

```
> plot(final ~ midterm)
> arrows(mean(midterm), min(final), mean(midterm), max(final),
+     code = 0)
> arrows(min(midterm), mean(final), max(midterm), mean(final),
+     code = 0)
> arrows(1 - 0.5/sdx, min(final), 1 - 0.5/sdx, max(final), code = 0)
> arrows(1 + 0.5/sdx, min(final), 1 + 0.5/sdx, max(final), code = 0)
```

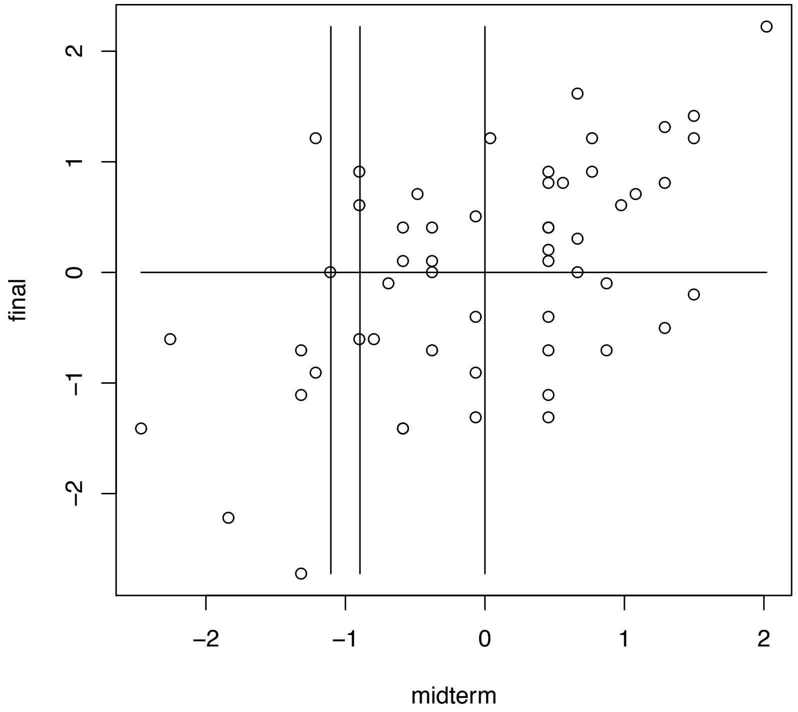


When the midterm is 1 SD above its mean, the finals scores are only 0.65 above the mean (recall that we're working with z-scores in the midterm and finals scores, mean 0 and sd 1).

```
> oneSDsubsampleabove <- subset(subset(scaledstat500, (1 - 0.5/sdx) <
+     midterm), (1 + 0.5/sdx) > midterm)
> yoneSDabove <- mean(oneSDsubsampleabove$final)
```

6.4.2 One SD below midterm means

```
> plot(final ~ midterm)
> arrows(mean(midterm), min(final), mean(midterm), max(final),
+     code = 0)
> arrows(min(midterm), mean(final), max(midterm), mean(final),
+     code = 0)
> arrows(-1 - 0.5/sdx, min(final), -1 - 0.5/sdx, max(final), code = 0)
> arrows(-1 + 0.5/sdx, min(final), -1 + 0.5/sdx, max(final), code = 0)
```



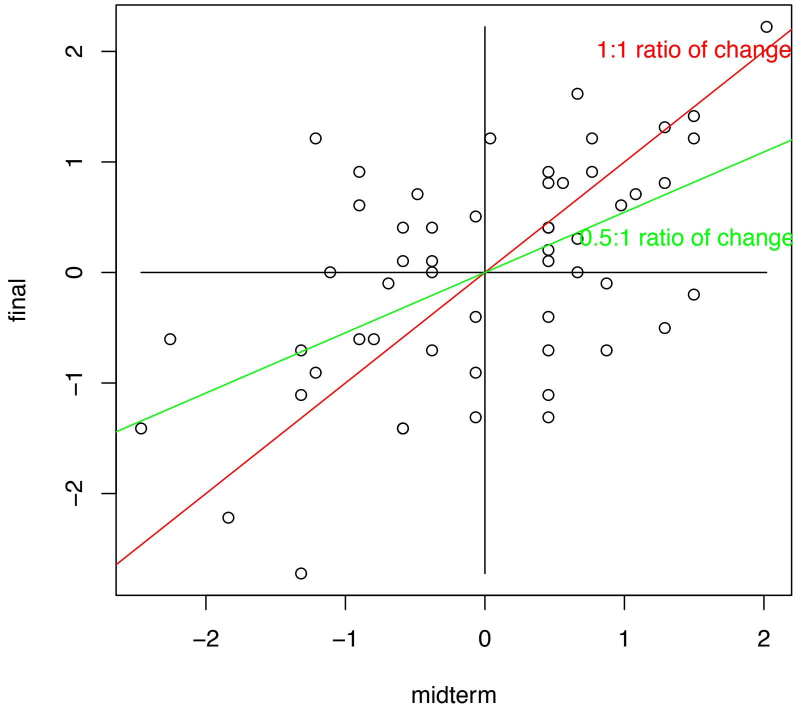
When the midterm is 1 SD below its mean, the finals scores are only 0.30 below the mean:

```
> oneSDsubsamplebelow <- subset(subset(scaledstat500, (-1 - 0.5/sdx) <
+     midterm), (-1 + 0.5/sdx) > midterm)
> oneSDsubsamplebelow

      midterm      final          hw      total
6  -0.9005185  0.9102018  0.31403725  0.1423654
39 -0.9005185  0.6074129 -3.73527088 -1.5987496
41 -0.9005185 -0.6037428 -0.01090723 -0.7232737

> yoneSDbelow <- mean(oneSDsubsamplebelow$final)
> yoneSDbelow

[1] 0.304624
```

6.4.3 Regression

- So, when the midterm score is 1 SD above the midterm mean, the final score is only 0.65 above the finals mean.
- When the midterm score is 1 SD below the midterm mean, the final score is only 0.30 below the finals mean.
- Recall that the correlation was .5 or so. Suppose now that we draw the line $y = 0.5 \times x$, and compare it with the line where a 1 SD change in midterm score results in a 1 SD change in the finals score (in the same direction): $y = x$.

```
> plot(final ~ midterm)
> arrows(mean(midterm), min(final), mean(midterm), max(final),
+     code = 0)
> arrows(min(midterm), mean(final), max(midterm), mean(final),
+     code = 0)
> abline(0, 1, col = "red")
> abline(0, 0.5452, col = "green")
> text(1.5, 2, labels = c("1:1 ratio of change"), col = "red")
> text(1.45, 0.3, labels = c("0.5:1 ratio of change"), col = "green")
```



When we averaged over a strip, we found a mean – the center of gravity, as it were – for those points in the strip. Recall now the method of least squares (Lecture 4, slide 18): the mean minimizes variance. The method of least squares applies in two dimensions as well: we can use it find a line that is at the center of a plane, i.e., closest to all the points than any other line. The R-command is `lm`:

This two-dimensional least-squares estimate is the line ($-3.997e - 16$ is essentially 0):

$$y = .5452 \times x \tag{6.2}$$

Recall that we'd previously calculated that $r = .5452$. We know that the slope of the line from the standardized data is $\frac{dy}{dx} = 0.5452$. In other words,

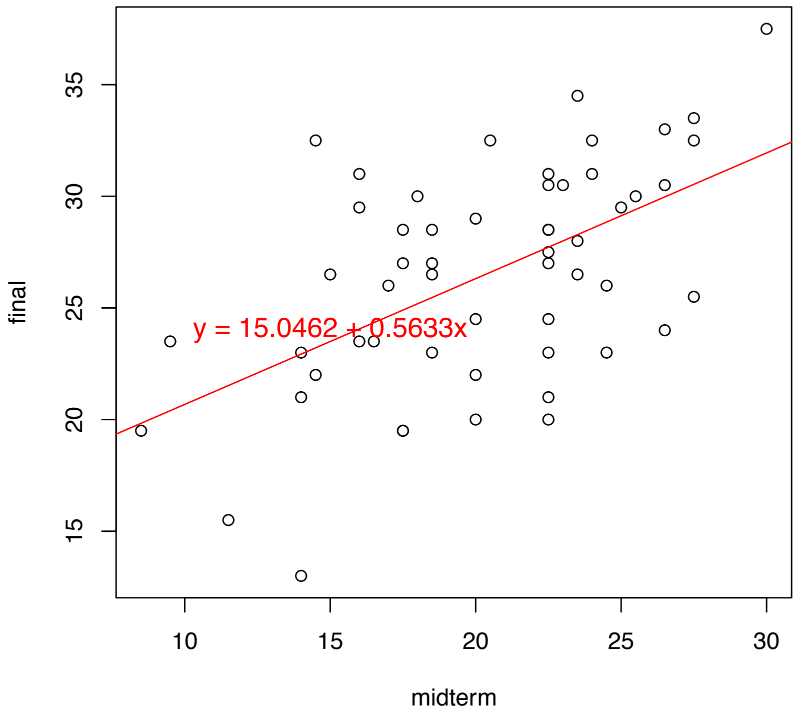
$$\frac{z_y}{z_x} = 0.5452 \tag{6.3}$$

To get the slope in the unstandardized data, we just need to undo the standardization:

$$\frac{z_y}{z_x} \times \frac{s_y}{s_x} = 0.5632756 \tag{6.4}$$

Equivalently, we can fit a least-squares line on the data:

```
> attach(stat500)


        The following object(s) are masked from scaledstat500 ( position 4 ) :

         final hw midterm total


        The following object(s) are masked from scaledstat500 ( position 5 ) :

         final hw midterm total


> lm.stat500 <- lm(final ~ midterm)
> plot(final ~ midterm)
> abline(lm.stat500, col = "red")
> text(15, 24, labels = c("y = 15.0462 + 0.5633x"), cex = 1.2,
+     col = "red")
```



6.5 Defining variance

The regression line is the 'bivariate mean'. What's the corresponding measure of variance in two-dimensions? Look at this equation again:

$$\hat{y} = 15.0462 + 0.5633 \times x \tag{6.5}$$

For any x, $\hat{y}$ is just the **predicted value**: in reality there's a considerable deviation from this $\hat{y}$. Call $y_i - \hat{y}_i$ the RESIDUAL ERROR. If the predicted value was perfect there would be no residual error – strictly analogous to variance.

Recall also that we minimized the sum of squared residual errors when we fit the regression line – that's what the Method of Least Squares is. So we can use residual error as our measure of variance about the regression line. There are several possibilities we can consider.

6.6 Defining variance and SD in regression

- Sum of squared residual error (SS$_{\text{res}}$): $\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Mean squared residual error (MSE$_{\text{res}}$): $\frac{\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} \rightsquigarrow s^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n}(x_i - \bar{x})^2$

ANOVA	Regression
$SS_{\text{Total}} = \sum_{j=1}^{I}\sum_{i=1}^{n_j}(x_{ij} - \bar{x})^2$	$SS_{\text{Total}} = \sum(y_i - \bar{y})^2$
$SS_{\text{Between}} = \sum_{j=1}^{I}\sum_{i=1}^{n_j}(\bar{x}_j - \bar{x})^2$	$SS_{\text{Regression}} = \sum(\hat{y}_i - \bar{y})^2$
$SS_{\text{Within}} = \sum_{j=1}^{I}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2$	$SS_{\text{Residual}} = \sum(y_i - \hat{y}_i)^2$

- Root mean squared residual error (RMSE$_{\text{res}}$): $\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} \rightsquigarrow s$
- Mean squared residual error (MSE$_{\text{res}}$): $\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$
- Recall ANOVA's MS-within: $\frac{\sum_{j=1}^{I}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2}{N-I}$
- The two are analogous: in MSE$_{\text{res}}$, we are taking the mean within all vertical strips, which are the $\hat{y}_i$'s, and taking the average of mean squared deviations from these means.
- We can go all the way to ANOVA, actually, because for regression as well we can define SS$_{\text{Total}}$, SS_{Between}, SS_{Within}.

6.7 Regression as hypothesis testing

Think of $\bar{y}$ as the grand mean, and $\hat{y}_i$ as the group mean – which is what they are. Note that (proof obvious): $SS_{\text{Total}} = SS_{\text{Regression}} + SS_{\text{Residual}}$

Recall ANOVA's null and alternative hypotheses:

$$\mathcal{H}_0 : x_{ij} = \mu + \epsilon_{ij} \tag{6.6}$$

$$\mathcal{H}_a : x_{ij} = \mu + \alpha_j + \epsilon_{ij} \tag{6.7}$$

Regression analysis uses the same logic:

$$\mathcal{H}_0 : y_i = \mu + \epsilon_i \tag{6.8}$$

$$\mathcal{H}_a : y_i = \hat{y}_i + \epsilon_i \tag{6.9}$$

Here, $\mu = \bar{y}$ and $\hat{y}_i = \beta_0 + \beta_1 x_i$.

So, given the SS's we can derive (as in ANOVA):

The "MS-between": MS$_{\text{Regression}}$: $\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{n-(n-1)}$.

The "MS-within": MS$_{\text{Residual}}$: $\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$.

So we can calculate the F-value:

$$\frac{MS_{\text{Regression}}}{MS_{\text{Residual}}} = \text{F-statistic} \tag{6.10}$$

6.8 Sum of squares and correlation

Notice that if all observed values of y had been on the line, then $y_i = \hat{y}_i$. I.e., $SS_{\text{Total}} = SS_{\text{Regression}}$. I.e., the regression would predict all the variation. If the observed values y_i spread out around the predicted values $\hat{y}_i$, then the regression would predict only part of the variation. We can state all this more succinctly:

$$
\begin{aligned}
SS_{\text{Total}} =& SS_{\text{Regression}} + SS_{\text{Residual}} && (6.11)\\
\frac{SS_{\text{Total}}}{SS_{\text{Total}}} =& \frac{SS_{\text{Regression}}}{SS_{\text{Total}}} + \frac{SS_{\text{Residual}}}{SS_{\text{Total}}} && (6.12)\\
1 =& \frac{SS_{\text{Regression}}}{SS_{\text{Total}}} + \frac{SS_{\text{Residual}}}{SS_{\text{Total}}} && (6.13)
\end{aligned}
$$

Clearly, $\frac{SS_{\text{Regression}}}{SS_{\text{Total}}}$ is telling you what proportion of the variance the regression equation can predict. This ratio is just the square of our old friend, r, the correlation coefficient.

```
> anova(lm.stat500)

Analysis of Variance Table

Response: final
          Df Sum Sq Mean Sq F value    Pr(>F)
midterm    1 393.96  393.96  22.421 1.675e-05 ***
Residuals 53 931.29   17.57
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

The above could have been done directly in R without fitting the linear model:

```
> summary(aov(final ~ midterm, stat500))

            Df Sum Sq Mean Sq F value    Pr(>F)
midterm      1 393.96  393.96  22.421 1.675e-05 ***
Residuals   53 931.29   17.57
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

We now have several ways of understanding regression:

- As the correlation of an explanatory variable to a predicted value: r.
- As an ANOVA: with F scores.
- As a proportion: the amount of variance explained by the regression equation: r^2.
- In R, the outputs of the calls to `summary(aov...)` and `summary(lm...)` give us all this information.

  ```
  > summary(lm.stat500)
  ```

```
Call:
lm(formula = final ~ midterm)

Residuals:
   Min     1Q Median     3Q    Max 
-9.932 -2.657  0.527  2.984  9.286 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  15.0462     2.4822   6.062 1.44e-07 ***
midterm       0.5633     0.1190   4.735 1.67e-05 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 4.192 on 53 degrees of freedom
Multiple R-squared: 0.2973,        Adjusted R-squared: 0.284 
F-statistic: 22.42 on 1 and 53 DF,  p-value: 1.675e-05
```

Chapter 7

Linear models and ANOVA

Having understood the foundational ideas behind ANOVA. now we look at ANOVA from a model comparison perspective. This section draws heavily from (Maxwell & Delaney, 2000). It's a bit technical, but the mathematics does not go beyond the 10th grade level (no calculus or linear algebra). So don't be put off; it's worth the effort to go through this chapter because it will solidify your understanding of how linear regression models and ANOVA fit together.

7.1 One way between subject designs

A typical experiment (in phonetics, psychology, psycholinguistics) has the following general format. We obtain a response (DV, for dependent variable) from participants in an experiment, and we have a hypothesis that this response depends on ("is a function of") several factors, which I will call A and B. This can be stated as an equation, as shown in (7.1). The word *others* refers to other sources that might affect the DV.

$$DV = \text{baseline DV} + \text{A} + \text{B} + \text{others} \tag{7.1}$$

The equation (7.1) can be written more precisely:

$$Y_i = \beta_0 X_{0_i} + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \epsilon_i \tag{7.2}$$

We have already seen this equation in the previous chapter, and how it relates to ANOVA.

- Y_i is the score of subject/participant i on the dependent variable, the observed values.
- β_0 is the *population* mean μ, and its coefficient X_{0_i} is (usually) 1.
- β_1 and β_2 are PARAMETERS that have to be estimated (we'll see later how this is done); they signify the contributions of the two factors A and B. The VARIABLES X_{1_i} and X_{2_i} are coefficients of the parameters.
- ϵ_i is the residual (everything that we cannot account for based on the factors A and B).

The equation is completely general: if you have p factors, the equation is:

$$Y_i = \beta_0 X_{0_i} + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \cdots + \beta_p X_{p_i} + \epsilon_i \tag{7.3}$$

In the following discussion, we build up the underlying details by beginning with a simple example, and then successively look at increasingly complex situations.

One group example

Consider first the simple situation when you have only one group. I.e., instead of two factors affecting the DV, we consider the case where there is only one.

$$Y_i = \mu + \epsilon_i, i = 1, \ldots, n \text{ where } n = \text{no. of subjects} \tag{7.4}$$

The equation asserts that the variable Y has some unknown typical value μ (the population mean), and that deviations from this typical value are due to random, uncontrolled factors: ϵ_i for each subject i.

Notice that one could have written (7.4) as

$$Y_i = \beta_0 X_{0_i} + \epsilon_i \tag{7.5}$$

Here, $\beta_0 = \mu$, and setting $X_0 = 1$ simply amounts to saying that μ has to be used as a prediction for each subject's equation.

Equation (7.4) is actually a series of equations, one for each i. So you have n equations, and $n+1$ unknowns (n unknowns are the ϵ_i's, and there is one unknown population mean μ). We could use *any* of a number of possible values of μ and ϵ_i, but we want a unique solution.

To get the unique, optimal solution, we can treat equation (7.4) as a prediction equation, where one tries to guess a value for μ that is *as close as possible* to the observed values Y_i. It's easy to quantify "as close as possible": minimize ϵ_i.

Suppose we take the mean of our observed values as our predicted (or estimated) value of μ; call this $\hat{\mu}$ (pronounced "μ hat"). Then we want to minimize e_i, which is defined as follows:

$$e_i = \hat{\epsilon_i} = Y_i - \hat{\mu} \tag{7.6}$$

Obviously, what this means is:

$$Y_i = \hat{\mu} + \hat{\epsilon_i} \tag{7.7}$$

If you try to get your guess $\hat{\mu}$ as close as possible to the observed Y_i, the extent to which you "missed" is $\hat{\epsilon_i}$. You want to minimize this miss. $\hat{\epsilon_i}$ is thus the error of prediction for each subject, and is estimated by e_i. We can use $e_i^2 = (Y_i - \hat{\mu})^2$ as a measure of our accuracy (or lack thereof), since squaring it gets rid of negative signs, and emphasizes large errors.

So, minimizing e_i is the same as stipulating that the sum or average of e_i^2 is as small as possible. Choosing parameter estimates to minimize squared errors of prediction is called the LEAST SQUARES CRITERION or LSC.

The least squares estimate (LSE) has some advantages:

- It's always unbiased.
- LSEs are minimum variance unbiased linear estimates, i.e., in replications, the LSE of the population parameter will have the least variability (= is more efficient) than any other estimator that's a linear combination of the observations in the sample (irrespective of whether ϵ_i is normally distributed or not).

In LSE we minimize

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\mu})^2 \tag{7.8}$$

by choosing the appropriate $\hat{\mu}$. Recall that the *sample* mean $\bar{Y}$ has the property that the sum of squared deviations from it are smaller than around any other value. So $\bar{Y}$ is really the best estimator for $\hat{\mu}$.

An important by-product of using LSE to estimate parameters is that it yields a measure of the ACCURACY of the model that is as fair as possible. That is, $\sum_{i=1}^{n} e_i^2$ is as small as it could be for this model.

Suppose now that we know (from previous experiments) that the mean of a population is known to be μ_0, and we wonder whether it is μ_0 for a particular sample of the population too. To give a concrete example, suppose we know that the mean IQ of all children of a particular age group is μ_0, and we want to know if the mean IQ μ_i of a specific group of i hyperactive children (of the same age group) is also μ_0.

The hypothesis that $\mu_0 = \mu_i$ is the NULL HYPOTHESIS H_0. This null hypothesis can be written as:

$$H_0 : Y_i = \mu_0 + \epsilon_i \tag{7.9}$$

Note that we're estimating *no* parameters here (call this the RESTRICTED MODEL). Compare equation (7.9) with the earlier one (7.7), repeated below:

$$Y_i = \hat{\mu} + \hat{\epsilon}_i \tag{7.10}$$

where we're estimating $\hat{\mu}$. Call this the UNRESTRICTED MODEL.

Now, for the restricted model,

$$e_i = \epsilon_i = Y_i - \mu_0 \tag{7.11}$$

which means that

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \mu_0)^2 \tag{7.12}$$

With some algebraic manipulation (exercise) for the restricted model you get

$$\sum_{i=1}^{n} (Y_i - \mu_0)^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + n(\bar{Y} - \mu_0)^2 \tag{7.13}$$

Now, the minimal error made in the *un*restricted model is:

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{7.14}$$

Suppose our null hypothesis were true (i.e., $\mu_0 = \mu_i = \bar{Y}$). Then, there would be no difference between the restricted model's error e_{i_R} and the unrestricted model's error e_{i_U}:

$$e_{i_R} - e_{i_U} = 0 \tag{7.15}$$

This is obvious since $\mu_0 = \bar{Y}$:

$$\left(\sum_{i=1}^{n} (Y_i - \bar{Y})^2 + n(\bar{Y} - \mu_0)^2 \right) - \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \tag{7.16}$$

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 - \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + n(\bar{Y} - \mu_0)^2 = \tag{7.17}$$

$$n(\bar{Y} - \mu_0)^2 = \tag{7.18}$$

$$n(\mu_0 - \mu_0)^2 = 0 \tag{7.19}$$

It also follows that if the null hypothesis is not true, then

$$e_{i_R} - e_{i_U} \neq 0 \tag{7.20}$$

This is also obvious since $\mu_0 \neq \bar{Y}$:

$$\left(\sum_{i=1}^{n}(Y_i - \bar{Y})^2 + n(\bar{Y} - \mu_0)^2\right) - \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \tag{7.21}$$

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + n(\bar{Y} - \mu_0)^2 = \tag{7.22}$$

$$= n(\bar{Y} - \mu_0)^2 \tag{7.23}$$

In other words, the further away $\bar{Y}$ is from our hypothesized value μ_0, the larger the difference in errors.

The key inferential step comes at this point. How much must the error increase for our assumption to be false that μ_0 is the mean of the subset we're interested in? We can take proportional increase in error (PIE):

$$PIE = \frac{\text{increase in error}}{\text{minimal error}} \tag{7.24}$$

This leads to the idea of a TEST STATISTIC. First, let's fix some terminology. Call the unrestricted model the FULL MODEL $\mathcal{F}$ because it is full of parameters, the number of parameters frequently equaling the number of groups in the design. Call the restricted model $\mathcal{R}$; recall that in $\mathcal{R}$ we've placed restrictions on the parameters of $\mathcal{F}$. For example, we've deleted a parameter (in the above one-group example). This restriction is our null hypothesis (specifically, in our example, the hypothesis that μ_0 is the subject's mean).

To summarize:

	Model	LSE	Errors
$\mathcal{F}$	$Y_i = \hat{\mu} + \epsilon_{i_\mathcal{F}}$	$\hat{\mu} = \bar{Y}$	$E_\mathcal{F} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
$\mathcal{R}$	$Y_i = \mu_0 + \epsilon_{i_\mathcal{R}}$	No parameters estimated	$E_\mathcal{R} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(Y_i - \mu_0)^2$

It follows that

$$PIE = \frac{(E_\mathcal{R} - E_\mathcal{F})}{E_\mathcal{F}} = \frac{n(\bar{Y} - \mu_0)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{7.25}$$

PIE compares the *adequacy* of the models, but ignores their *relative* complexity. We know already that $\mathcal{R}$ has to be less adequate than our $\mathcal{F}$ (why? The answer is: if the restricted model is less adequate than the full model, $E_\mathcal{F} < E_\mathcal{R}$). Notice that we're gaining simplicity but losing adequacy in moving from $\mathcal{F}$ to $\mathcal{R}$ (the restricted model has fewer parameters, so it's simpler, in a sense). If we could find out what the loss in adequacy was *per additional unit of simplicity*, we have a measure of the relative adequacy of the models $\mathcal{F}$ and $\mathcal{R}$, taking their relative simplicity into account.

If, in transitioning from $\mathcal{F}$ to $\mathcal{R}$, the loss in adequacy *per unit gain in simplicity* is large, then we have some reason to believe that our null hypothesis was false.

Quantifying simplicity of a model is the key problem now. The fewer the parameters, the simpler the model. Conversely, the more the number of parameters, the more complex the model.

To quantify simplicity, we want a number that should increase as the number of parameters decrease. Call this "degrees of freedom", *df*, and define it as follows:

$$\text{df} = (\text{no. of independent observations}) - (\text{no. of indep. parameters estimated}) \tag{7.26}$$

We can use *df* as our index of simplicity. So now we have the right measure – PIE relativized to simplicity:

$$PIE = \frac{(E_{\mathcal{R}} - E_{\mathcal{F}})/(df_{\mathcal{R}} - df_{\mathcal{F}})}{E_{\mathcal{F}}/df_{\mathcal{F}}} = F = t^2 \tag{7.27}$$

The interesting thing with this presentation is that *all* tests in ANOVA, ANCOVA, bivariate and multiple regression can be computed using this formula. Every new setup discussed after this point depends on the above result.

Notice that if the adequacy of $\mathcal{R}$ and $\mathcal{F}$ per degree of freedom is the same, $F = 1$. In that case, we'd prefer the simpler model $\mathcal{R}$ (more on this later). On the other hand, if the error per *df* of $\mathcal{R}$ is larger, the simpler model is inadequate; this amounts to saying that there is a significant difference between the population mean and the mean of the subset we're interested in. For example, if $F = 9$, that means that the additional error of the simpler, restricted model per its additional *df* is nine times larger than we would expect it to be on the basis of the error for the full model per degree of freedom. That is, the restricted model is considerably worse per extra degree of freedom in describing the data than is the full model relative to its *df*.

This can be re-stated as follows:

Hypothesis	Model
$H_1 : \mu \neq \mu_0$	Full : $Y_i = \mu + \epsilon_{i\mathcal{F}}$
$H_0 : \mu = \mu_0$	Restricted: $Y_i = \mu_0 + \epsilon_{i\mathcal{R}}$

7.2 Extending Linear Models to two groups

The above can be extended to two groups. The situation now is summarized as follows. Let μ_1 be the population mean for one group, μ_2 the population mean for the other group. More generally, let μ_j be the population mean for the jth group (i.e., $j = 1, 2$), and $i = 1, \ldots, n_j$.

Hypothesis	Model
$H_1 = \mu_1 \neq \mu_2$	Full $= Y_{ij} = \mu_j + \epsilon_{ij_{\mathcal{F}}}$
$H_0 = \mu_1 = \mu_2$	Restricted $= Y_{ij} = \mu + \epsilon_{ij_{\mathcal{R}}}$

To take a concrete example, we could be comparing the IQs of two groups of children, one "normal", and the other hyperactive. The normal group of n_1 children ($j = 1$) would have IQ μ_1, and the hyperactive group consisting of n_2 children ($j = 2$) would have IQ μ_2. We want to know if $\mu_1 = \mu_2$ (the null hypothesis).

After a little bit of mathematics (interesting, but we can skip it; see (Maxwell & Delaney, 2000, 77-80) for details), we get the following for the two-group situation:

$$PIE = \frac{(E_{\mathcal{R}} - E_{\mathcal{F}})/(df_{\mathcal{R}} - df_{\mathcal{F}})}{E_{\mathcal{F}}/df_{\mathcal{F}}} = \frac{\sum\limits_{j} n_j(\bar{Y}_j - \bar{Y})^2}{\sum\limits_{j}\sum\limits_{i}(Y_{ij} - \bar{Y}_j)^2/(N-2)} \tag{7.28}$$

7.2.1 Traditional terminology of ANOVA and the model comparison approach

Traditionally, F tests are supposed to indicate whether between-group variability is greater than within-group variability:

$$
\begin{aligned}
F &= \frac{\text{Variability between groups}}{\text{Variability within groups}} && (7.29)\\
&= \frac{\text{Mean square error between groups}}{\text{Mean square error within groups}} && (7.30)\\
&= \frac{MS_b}{MS_w} && (7.31)
\end{aligned}
$$

Intuitively, the logic is as follows. Given two groups with means μ_1 and μ_2, it is almost certain that $\mu_1 \neq \mu_2$, because of sampling variability. So the question really is: is the difference between treatment groups greater than within each group? The latter would be due to sampling variability. The equation in (7.29) reflects this.

The difference between μ_1 and μ_2 depends on the variability of the population, which can be estimated: take either of the groups' variance, or a weighted average of the two (weighted by the number of scores in each group).

Suppose each group's variance is s_j^2:

$$
s_j^2 = \frac{\sum\limits_i (Y_{ij} - \bar{Y}_j)^2}{n_j - 1} \tag{7.32}
$$

Then, σ^2, the weighted[1] average (or pooled estimate) of the two variances, s_1^2 and s_2^2, is:

$$
\sigma^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)}{n_1 + n_2 - 2} \tag{7.33}
$$

From equation (7.32) it follows that (**make sure that you see why it follows!**):

$$
\begin{aligned}
\sigma^2 &= \frac{\sum\limits_i (Y_{i1} - \bar{Y}_1)^2 + \sum\limits_i (Y_{i2} - \bar{Y}_2)^2}{n_1 + n_2 - 2} && (7.34)\\
&= \frac{\sum\limits_j \sum\limits_i (Y_{ij} - \bar{Y}_j)^2}{\sum\limits_j (n_j - 1)} && (7.35)
\end{aligned}
$$

Since the last result above is an average or mean squared deviation *within* the groups, we have MS_w:

$$
MS_w = \frac{\sum\limits_j \sum\limits_i (Y_{ij} - \bar{Y}_j)^2}{\sum\limits_j (n_j - 1)} \tag{7.36}
$$

Notice here that the sum of squares within the groups, call it SS_w, is:

$$
SS_w = \sum_j \sum_i (Y_{ij} - \bar{Y}_j)^2 \tag{7.37}
$$

[1] Weighted by the number of free parameters: this is $n_j - 1$ since we've already "used up" one parameter to estimate s_j^2 – the mean $\bar{Y}_j$.

We will be using SS_w a lot in the future, so it's a good idea to internalize what this means.

Next, we calculate the MS_b, the mean standard deviation *between* the two groups.

Suppose the null hypothesis is true: $\mu_1 = \mu_2$. What is the variability between the sample means μ_1 and μ_2? I.e., what is the variance of the means? The answer is:

$$\frac{\sum_j (\bar{Y}_j - \bar{Y})^2}{a-1} \tag{7.38}$$

This is the variance of the sample means; call it $\sigma^2_{\bar{Y}}$. We know that (assuming an equal number of scores n in both groups):

$$\sigma^2_Y = n \times \sigma^2_{\bar{Y}}, \text{where } \sigma^2_Y \text{ is the population variance} \tag{7.39}$$

It follows that

$$\sigma^2_{\bar{Y}} = n \times \frac{\sum_j (\bar{Y}_j - \bar{Y})^2}{a-1} \tag{7.40}$$

This is the mean squared deviation between groups:

$$MS_b = n \times \frac{\sum_j (\bar{Y}_j - \bar{Y})^2}{a-1} \tag{7.41}$$

Again, here the sum of squares between groups, call it SS_b, is:

$$SS_b = n \times \sum_j (\bar{Y}_j - \bar{Y})^2 \tag{7.42}$$

We will need SS_b again later.

Regarding MS_b and SS_b, note that when you have $j = a$ groups with n_j subjects in each group, the equation generalizes to:

$$MS_b = \frac{\sum_{j=1}^{a} n_j (\bar{Y}_j - \bar{Y})^2}{a-1} \tag{7.43}$$

$$SS_b = \sum_{j=1}^{a} n_j (\bar{Y}_j - \bar{Y})^2 \tag{7.44}$$

To summarize:

- Sum of squares between groups:

$$SS_b = \sum_{j=1}^{a} n_j(\bar{Y}_j - \bar{Y})^2 \tag{7.45}$$

- Mean square deviation between groups:

$$MS_b = \frac{\sum\limits_{j=1}^{a} n_j(\bar{Y}_j - \bar{Y})^2}{a-1} \tag{7.46}$$

- Sum of squares within groups:

$$SS_w = \sum_{j}\sum_{i}(Y_{ij} - \bar{Y}_j)^2 \tag{7.47}$$

- Mean square deviation within groups:

$$MS_w = \frac{\sum\limits_{j}\sum\limits_{i}(Y_{ij} - \bar{Y}_j)^2}{\sum\limits_{j}(n_j - 1)} \tag{7.48}$$

Now we can see the connection between the model-comparison approach and the traditional view of F tests:

$$\begin{aligned} F &= \frac{MS_b}{MS_w} && (7.49)\\ &= \frac{\sum\limits_{j=1}^{a} n_j(\bar{Y}_j - \bar{Y})^2}{a-1} \div \frac{\sum\limits_{j}\sum\limits_{i}(Y_{ij} - \bar{Y}_j)^2}{\sum\limits_{j}(n_j - 1)} && (7.50)\\ &= PIE && (7.51) \end{aligned}$$

This is because (recall (7.28), repeated below):

$$PIE = \frac{(E_{\mathcal{R}} - E_{\mathcal{F}})/(df_{\mathcal{R}} - df_{\mathcal{F}})}{E_{\mathcal{F}}/df_{\mathcal{F}}} = \frac{\sum\limits_{j} n_j(\bar{Y}_j - \bar{Y})^2}{\sum\limits_{j}\sum\limits_{i}(Y_{ij} - \bar{Y}_j)^2/(N-2)} \tag{7.52}$$

7.3 Individual comparisons of means – between subject data

Our null hypothesis (in an a-group study) has so far been:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a \tag{7.53}$$

The full and restricted models were:

$$Y_{ij} = \mu_j + \epsilon_{ij_{\mathcal{F}}} \tag{7.54}$$

$$Y_{ij} = \mu + \epsilon_{ij_{\mathcal{R}}} \tag{7.55}$$

Suppose now that our null hypothesis is: Do the means of two of the groups (say groups 1 and 2) differ? I.e.,

$$H_0 : \mu_1 = \mu_2 \tag{7.56}$$

The *restricted* model now changes to

$$Y_{ij} = \mu_i + \epsilon_{ij_{\mathcal{R}}}, \text{where } \mu_1 = \mu_2 \tag{7.57}$$

We could re-write this as:

$$Y_{i1} = \mu^* + \epsilon_{i1_{\mathcal{R}}} \tag{7.58}$$

$$Y_{i2} = \mu^* + \epsilon_{i2_{\mathcal{R}}} \tag{7.59}$$

$$Y_{ij} = \mu_j + \epsilon_{ij_{\mathcal{R}}}, j = 3, 4, \ldots, a \tag{7.60}$$

The new means μ^* refers to the mean of the two groups' scores; groups 3 to a can have their own potentially unique means. Since we've identified the restricted and full models, determining the F value is simply a matter of algebraic manipulation.

$$E_F = SS_w = \sum_j \sum_i (Y_{ij} - \bar{Y}_j)^2 \tag{7.61}$$

$$E_R = \sum_{j=1}^{2} \sum_{i=1}^{n_j} \left(Y_{ij} - \bar{Y}^*\right)^2 + \sum_{j=3}^{n} \sum_{i=1}^{n_j} \left(Y_{ij} - \bar{Y}_j\right)^2 \tag{7.62}$$

$$F = \frac{(E_{\mathcal{R}} - E_{\mathcal{F}})/(df_{\mathcal{R}} - df_{\mathcal{F}})}{E_{\mathcal{F}}/df_{\mathcal{F}}} \tag{7.63}$$

Recall the definition of *df*:

$$df = \text{no. of independent observations} - \text{no. of parameters} \tag{7.64}$$

Since $df_F = N - a$, where N is the total number of scores (across all groups, and $df_R = N - (a - 1)$, we can rewrite F as follows (after some algebraic messing around, that is):

$$F = \frac{n_1 n_2 (\bar{Y}_1 - \bar{Y}_2)^2}{(n_1 + n_2) MS_w} \tag{7.65}$$

7.4 Complex comparisons

Suppose we administer a blood pressure treatment study. There are four treatments, and one is called a "combination treatment". Suppose our research question was: "Is the combination treatment more effective than the average of the other three?" The corresponding null hypothesis is simply the negation of this statement, and is expressed as shown below:

$$H_0 : \frac{1}{3}(\mu_1 + \mu_2 + \mu_3) = \mu_4 \tag{7.66}$$

The full model remains unchanged:

$$Y_{ij} = \mu_j + \epsilon_{ij_{\mathcal{F}}} \tag{7.67}$$

but the corresponding restricted model is:

$$Y_{ij} = \mu_j + \epsilon_{ij_{\mathcal{R}}}, \text{ where } \frac{1}{3}(\mu_1 + \mu_2 + \mu_3) = \mu_4 \tag{7.68}$$

Suppose we re-write the null hypothesis as:

$$H_0 : \frac{1}{3}\mu_1 + \frac{1}{3}\mu_2 + \frac{1}{3}\mu_3 - \mu_4 = 0 \tag{7.69}$$

A more general form of this hypothesis would be:

$$H_0 : c_1\mu_1 + c_2\mu_2 + c_3\mu_3 + c_4\mu_4 = 0 \tag{7.70}$$

In the present case, $c_1 = c_2 = c_3 = \frac{1}{3}$ and $c_4 = -1$. Let us stipulate that the situation in equation (7.70) is a CONTRAST (or COMPARISON). In other words, let us define contrast ψ as a linear combination of population means in which the coefficients of the means sum to zero.

$$\psi = \sum_j^a c_j\mu_j \text{ where } \sum_j^a c_j = 0 \tag{7.71}$$

Such a general definition of contrasts allows us to test *any* contrast at all.

Note that the coefficients need not sum to zero; this is just a stipulation. However, when they don't, it is often the case that the contrast doesn't really mean anything. For example, in the blood pressure example, one could ask if the combination treatment is four times better than the average of the other three means. Here, $c_1 = c_2 = c_3 = 1/3$ and $c_4 = -4$, and the sum of coefficients is $1 - 4 = -3$. Maybe we do want to know the answer to such a question; it depends on the situation.

By the way, now our null hypothesis can simply be stated in terms of the contrast of interest:

$$H_0 : \psi = 0 \tag{7.72}$$

It is possible, but difficult, to find $E_{\mathcal{R}}$. But what we really need is only $E_{\mathcal{R}} - E_{\mathcal{F}}$, so that's what we'll derive (take it on trust for now).

$$E_{\mathcal{R}} - E_{\mathcal{F}} = \frac{(\hat{\psi})^2}{\sum\limits_{j=1}^{a}(c_j^2/n_j)} \tag{7.73}$$

Here, $\hat{\psi}$ is a sample estimate of the population parameter ψ:

$$\hat{\psi} = \sum_{j=1}^{a} c_j\bar{Y}_j \tag{7.74}$$

Now we're almost ready to replace the terms in the equation for F:

$$F = \frac{(E_{\mathcal{R}} - E_{\mathcal{F}})/(df_{\mathcal{R}} - df_{\mathcal{F}})}{E_{\mathcal{F}}/df_{\mathcal{F}}} \tag{7.75}$$

Recall again the definition of df:

$$df = \text{no. of independent observations} - \text{no. of parameters} \tag{7.76}$$

Let's compute the dfs. For a groups, we will have $a - 1$ parameters. This is because the ath parameter is predictable if we fix the others: if you have four groups, and your null hypothesis is

$$H_0 : \frac{1}{3}\mu_1 + \frac{1}{3}\mu_2 + \frac{1}{3}\mu_3 - \mu_4 = 0 \tag{7.77}$$

then μ_4 is predictable if you know (or have estimated) μ_1 to μ_3. So: df_R is $N-(a-1)$, where N is the number of independent observations. df_F is simply $N-a$, because you have a parameters.

So:

$$df_{\mathcal{R}} - df_{\mathcal{F}} = (N-(a-1)) - (N-a) = 1 \tag{7.78}$$

Also, recall that:

$$E_F/df_{\mathcal{F}} = MS_w \tag{7.79}$$

Finally, we're there:

$$F = \frac{\hat{\psi}^2}{MS_w \sum\limits_{j=1}^{a} (c_j^2/n_j)} \tag{7.80}$$

7.5 Generalizing the model comparison technique for any number of a groups

There is an easier way of talking about the effect a given factor has on the dependent variable (we will use this technique a lot later on, so it's useful to learn it now).

Instead of writing our full model for a groups as

$$Y_{ij} = \mu_j + \epsilon_{ijF} \tag{7.81}$$

we can write

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij_F} \tag{7.82}$$

The above reformulation amounts to saying that each factor $j = 1 \dots a$, contributes α_j to the mean: $\alpha_1 \dots \alpha_a$ are the effects of each of the factors.

In order to solve the equation above with a unique solution, we impose (as earlier) a side condition:

$$\sum_{j=1}^{a} \alpha_j = 0 \tag{7.83}$$

This is not a random choice. Notice that:

$$\mu_j = \mu + \alpha_j \tag{7.84}$$

That is, α_j is simply the deviation from the mean:

$$\alpha_j = \mu_j - \mu \tag{7.85}$$

We know that the sum of deviations from the mean sum to zero, and so

$$\sum_{j=1}^{a} \alpha_j = \sum_{j=1}^{a} (\mu_j - \mu) = 0 \tag{7.86}$$

Also, notice that it follows that the grand mean μ is:

$$\mu = \frac{\sum\limits_{j=1}^{a} \mu_j}{a} \tag{7.87}$$

Now we estimate parameters in the style of our one group example earlier (116):

Our equation for the full model, i.e.,

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij_F} \tag{7.88}$$

is a system of linear equations, one for each subject $i = N$. Notice that we have to estimate $N + 1 + (a - 1)$ parameters, where a is the number of groups. The number of parameters to be estimated is $a - 1$, and not a because all the a's sum to zero, so if we know any three the fourth is predictable.

As before, since we want a unique solution, we guess a value of $\mu + \alpha$ that is as close as possible to Y_{ij}. So, as usual, we minimize ϵ_{ij}:

$$\epsilon_{ij\mathcal{F}} = \sum_j \sum_i [Y_{ij} - (\hat{\mu} + \alpha_j)]^2 \tag{7.89}$$

How to estimate α? Notice that

$$\hat{\mu} = \frac{\sum\limits_{j=1}^{a} \bar{Y}_j}{a} = \bar{Y}_u \tag{7.90}$$

where $\bar{Y}_u$ is the unweighted mean.

Since $\bar{Y}_j = \hat{\mu} + \alpha_j$, it follows that

$$\alpha_j = \bar{Y}_j - \hat{\mu} \tag{7.91}$$

or (replacing $\hat{\mu}$):

$$\alpha_j = \bar{Y}_j - \frac{\sum\limits_{j=1}^{a} \bar{Y}_j}{a} \tag{7.92}$$

The restricted model's degrees of freedom are $N - 1$, and the full model's are $N - (a - 1) + 1) = N - a$.

When we have equal n in each group,

$$\begin{aligned} E_{\mathcal{R}} - E_{\mathcal{F}} =& \sum_j \sum_i \hat{\alpha_j}^2 && (7.93) \\ =& n \sum_j \hat{\alpha_j}^2 && (7.94) \end{aligned}$$

With unequal n:

$$\begin{aligned} E_{\mathcal{R}} - E_{\mathcal{F}} =& \sum_j \sum_i (Y_j - \hat{\mu})^2 && (7.95) \\ =& n_j (Y_j - \hat{\mu})^2 && (7.96) \\ =& n_j (\hat{\alpha})^2 && (7.97) \end{aligned}$$

Now it's straightforward to compute F, but the reason this technique was introduced here is that it's very useful in within subject designs' analyses. That's discussed in the next section.

7.6 Within subjects, two level designs

Suppose we did an experiment involving one factor with two levels. Recall the model for one way between subjects:

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij} \tag{7.98}$$

A problem here is that for any given subject the errors in group 1 and 2 are likely to be correlated: a subject who gives a high score in one group is likely to give a high score in the other. This is a problem because ANOVA assumes independence of errors, and here they're anything but.

What to do? Notice that the problem is that we have two errors per subject, and they're correlated. If we could get only one error per subject, the correlated-error problem is gone. We can rewrite (7.98) as:

$$Y_{i1} = \mu + \alpha_1 + \epsilon_{i1} \tag{7.99}$$

$$Y_{i2} = \mu + \alpha_2 + \epsilon_{i2} \tag{7.100}$$

Subtracting (7.100) from (7.99), we get:

$$Y_{i2} - Y_{i1} = \alpha_2 - \alpha_1 + \epsilon_{i2} - \epsilon_{i1} \tag{7.101}$$

and this is our full model ($M_{\mathcal{F}}$):

$$D_i = \mu + \epsilon_i \tag{7.102}$$

Our null hypothesis earlier was stated as $\alpha_1 = \alpha_2 = 0$, but now it is:

$$H_0 = \mu = 0 \tag{7.103}$$

and our restricted model $M_{\mathcal{R}}$ is:

$$D_i = 0 + \epsilon_i \tag{7.104}$$

or simply

$$D_i = \epsilon_i \tag{7.105}$$

Now we compute the terms of our F-statistic. The LSE of μ is $\bar{D}$.

$$E_F = \sum_i (D_i - \bar{D})^2 \tag{7.106}$$

$$E_R = \sum_i (D_i - 0)^2 = \sum_i D_i^2 \tag{7.107}$$

What's $E_R - E_F$? Obviously:

$$E_R - E_F = \sum_i (D_i - \bar{D})^2 - \sum_i D_i^2 \tag{7.108}$$

Perhaps less obviously, this reduces to:

$$E_R - E_F = n\bar{D}^2 \tag{7.109}$$

Exercise 10

1. Prove that $E_R - E_F = n\bar{D}^2$.
2. What are the degrees of freedom for the full and restricted models here?

Now, we're ready to compute F:

$$\begin{aligned} F =& \frac{n\bar{D}^2/n-(n-1)}{\sum\limits_i (D_i-\bar{D})^2/(n-1)} && (7.110) \\ =& \frac{n\bar{D}^2}{s_D^2} && (7.111) \end{aligned}$$

where

$$s_D^2 = \frac{\sum D_i^2 - n\bar{D}^2}{n-1} \qquad (7.112)$$

is the unbiased estimate of the population variance of the D scores.

Exercise 11

Prove that $F = \frac{n\bar{D}^2}{s_D^2}$.

Notice that F could be rewritten as:

$$t = \frac{\sqrt{n}\bar{D}}{s_D} \qquad (7.113)$$

This is the well-known formula for a dependent t-test. With two levels of the repeated factor, the model-comparisons test reduces to the dependent t-test.

7.7 R example for within-subjects designs

This is an example from Hays' book (1988, Table 13.21.2, p. 518) and was used in the Baron and Li notes on CRAN. A 2×2 within subjects design. We begin by setting up the data:

```
> data1 <- c(49, 47, 46, 47, 48, 47, 41, 46, 43, 47, 46, 45, 48,
+     46, 47, 45, 49, 44, 44, 45, 42, 45, 45, 40, 49, 46, 47, 45,
+     49, 45, 41, 43, 44, 46, 45, 40, 45, 43, 44, 45, 48, 46, 40,
+     45, 40, 45, 47, 40)
> Hays.mul.df <- as.data.frame(matrix(data1, ncol = 4, dimnames = list(paste("subj",
+     1:12), c("Shape1.Color1", "Shape2.Color1", "Shape1.Color2",
+     "Shape2.Color2"))))
> Hays.df <- data.frame(rt = data1, subj = factor(rep(paste("subj",
+     1:12, sep = ""), 4)), shape = factor(rep(rep(c("shape1",
+     "shape2"), c(12, 12)), 2)), color = factor(rep(c("color1",
+     "color2"), c(24, 24))))
```

The ANOVA call in R gives you the following output:

```
> anova.fm <- aov(rt ~ shape * color + Error(subj/(shape * color)),
+     data = Hays.df)
> summary(anova.fm)
```

```
Error: subj
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals 11 226.500  20.591

Error: subj:shape
          Df  Sum Sq Mean Sq F value  Pr(>F)
shape      1 12.0000 12.0000  7.5429 0.01901 *
Residuals 11 17.5000  1.5909
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Error: subj:color
          Df  Sum Sq Mean Sq F value   Pr(>F)
color      1 12.0000 12.0000  13.895 0.003338 **
Residuals 11  9.5000  0.8636
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Error: subj:shape:color
            Df    Sum Sq   Mean Sq   F value Pr(>F)
shape:color  1 1.246e-27 1.246e-27 4.495e-28      1
Residuals   11   30.5000    2.7727

> coefficients(anova.fm)

(Intercept) :
(Intercept)
         45

subj :
numeric(0)

subj:shape :
shapeshape2
         -1

subj:color :
colorcolor2
         -1

subj:shape:color :
shapeshape2:colorcolor2
           2.038340e-14
```

Now we compute the ANOVA “by hand”, using the equations worked out in this chapter.

First we compute the Sum of Squares within:

```
> c1 <- Hays.mul.df$Shape1.Color1
> c2 <- Hays.mul.df$Shape1.Color2
> c3 <- Hays.mul.df$Shape2.Color1
> c4 <- Hays.mul.df$Shape2.Color2
```

Now we look at the main effect of shape using the formula for F that we just derived in the sections above.

```
> Shape1 <- (c1 + c2) * 0.5
> Shape2 <- (c3 + c4) * 0.5
> DShape <- Shape2 - Shape1
> SumSqShape <- sum((mean(DShape) - DShape)^2)
> sdShape <- sd(DShape)
> n <- 12
> barD <- mean(DShape)
> (F <- (n * (barD^2))/(sdShape^2))

[1] 7.542857
```

Notice that the F value is exactly what R's ANOVA gives us. Now let's look at the main effect of color:

```
> Color1 <- (c1 + c3) * 0.5
> Color2 <- (c2 + c4) * 0.5
> DColor <- Color2 - Color1
> sum((mean(DColor) - DColor)^2)

[1] 9.5

> sdColor <- sd(DColor)
> n <- 12
> barD <- mean(DColor)
> (F <- (n * (barD^2))/(sdColor^2))

[1] 13.89474
```

Again, we get the F value that R gives us. Finally, look at Color and Shape interaction:

```
> Shapes <- (c1 - c2) * 0.5
> Colors <- (c3 - c4) * 0.5
> DSC <- (Shapes - Colors)
> sum((mean(DSC) - DSC)^2)

[1] 30.5

> sdDSC <- sd(DSC)
> n <- 12
> barD <- mean(DSC)
> (F <- (n * (barD^2))/(sdDSC^2))

[1] 0
```

The F-value for the interaction in the R code is not exactly zero but it's close enough. If you can't guess the reason why R would give a non-zero number, don't worry about it. It's only important to note that R-s F-value, 6.947e-29, is essentially 0.

Chapter 8

Linear mixed-effects models: An introduction

This chapter introduces linear mixed-effects models. Prerequisites for understanding this presentation are fearlessness, and a basic understanding of linear regression. The content is not easy, but comments on improving accessibility and understandability are most welcome.

8.1 Introduction

The standard linear model has only one random component, that is, the error term ϵ_i and one variance $\mathrm{Var}(\epsilon_i)$. In this chapter we consider a more sophisticated model.

We begin with a dataset discussed in Bryk and Raudenbush (1992). This dataset contains math achievement scores for subjects in 160 schools, and also provides the sex, Socio-economic status (SES), and minority status of each student.

```
> library(lme4)
> MathAchieve <- read.table("mathachieve.txt")
> colnames(MathAchieve) <- c("School", "Minority", "Sex", "SES",
+     "MathAch", "MEANSES")
> head(MathAchieve)
```

```
  School Minority    Sex    SES MathAch MEANSES
1   1224       No Female -1.528   5.876  -0.428
2   1224       No Female -0.588  19.708  -0.428
3   1224       No   Male -0.528  20.349  -0.428
4   1224       No   Male -0.668   8.781  -0.428
5   1224       No   Male -0.158  17.898  -0.428
6   1224       No   Male  0.022   4.583  -0.428
```

The SESes are sometimes negative because they are centered SESes: actual SES minus the mean SES of a particular school. The reason for this will become clear in a moment.

8.2 Simple linear model

Suppose our research question is: Does Socio-economic status (SES) affect math achievement?

A garden-variety linear model for this dataset would predict math achievement as a function of SESs:

$$Y_i = \beta_o + \beta_1 X_i + \epsilon_i \tag{8.1}$$

```
> lm0 <- lm(MathAch ~ SES, data = MathAchieve)
> summary(lm0)

Call:
lm(formula = MathAch ~ SES, data = MathAchieve)

Residuals:
     Min       1Q   Median       3Q      Max
-19.4382  -4.7580   0.2334   5.0649  15.9007

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.74740    0.07569  168.42   <2e-16 ***
SES          3.18387    0.09712   32.78   <2e-16 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 6.416 on 7183 degrees of freedom
Multiple R-squared: 0.1301,        Adjusted R-squared:  0.13
F-statistic:  1075 on 1 and 7183 DF,  p-value: < 2.2e-16

> coefficients(lm0)

(Intercept)         SES
   12.74740     3.18387
```

The coefficients tell us that the linear model is:

$$Y_i = 12.74 + 3.18 X_i + \epsilon_i \tag{8.2}$$

where X_i is the SES of each student, ϵ_i is the random error associated with each student. The error is assumed to be independent and identically distributed (iid for short), and is assumed to have a normal distribution centered around 0 with standard deviation σ (This mouthful about the errors can be written as $\epsilon_i \sim N(0, \sigma^2)$).

The term iid means that each error value in the sequence of errors comes from the same probability distribution (hence identical) and each error value is independent of the other values.

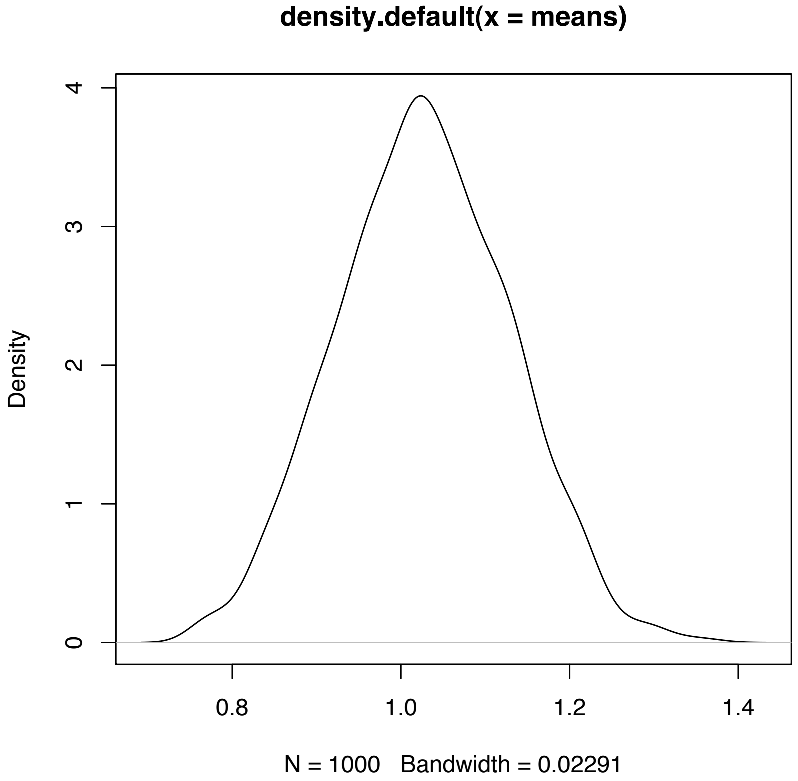
The iid assumption is necessary because the standard hypothesis testing procedure rests on the the Central Limit Theorem. This theorem says that the distribution of the mean of iid variables approaches a normal distribution given a large enough sample size.

Let us take a minute to confirm the Central Limit Theorem. First, we create a non-normal population:

```
> population <- rexp(1000)
> plot(density(population))
```

Then we sample 1000 times from it, taking a sample of size 100 each time, and plot the distribution of the 1000 means (one from each sample):

```
> sample.size <- 100
> n.sim <- 1000
> means <- rep(NA, n.sim)
> for (i in c(1:n.sim)) {
+     sample100 <- sample(population, sample.size)
+     means[i] <- mean(sample100)
+ }
> plot(density(means))
```



As an exercise, try making a sequence of plots of the distribution of means where the sample size goes from 1 to 100. How does the distribution of means change?

Returning to the linear model above, the model helps us answer the question about the relationship between SES and math achievement. The math achievment for any school j is predicted by a constant term 12.74, a factor 3.18 that is multiplied with their SES, plus some error associated with that particular school j.

But the above model cannot answer some other, perhaps more interesting questions:

- Do schools with higher mean math achievement also have stronger associations between SES and achievement (than schools with lower mean achievement scores)?

- Does SES affect math achievement to the same extent in each school? You can guess that this is probably not true, but how to find this out? If SES is not an important predictor in some schools but is in some others, this potentially is an important issue we should not ignore.

- Suppose schools can be separated out into different types, say Public versus Catholic. After we control for mean SES, do the two school types differ in terms of mean math achievement and the strength of the SES-math achievement relationship?

Mixed-effects models come in at this point; they help us answer such questions.

Our goal in the coming pages is to fit a more articulated linear model, where we have a separate intercept and slope for each school. Remember that the linear model above is fitting a single intercept and slope for all scores; it does not take individual variation into account at all. It is quite likely that the schools are quite a bit different from each other; if so, our simple model is an oversimplification.

To see this variability between schools, first let's just focus on the first two schools' data and plot the regression line for achievement against SES for each school.

8.2.1 Linear model of school 1224

```
> lm1 <- lm(MathAch ~ SES, data = subset(MathAchieve, School ==
+     1224))
> summary(lm1)

Call:
lm(formula = MathAch ~ SES, data = subset(MathAchieve, School ==
    1224))

Residuals:
    Min      1Q  Median      3Q     Max
-12.849  -6.377  -1.164   6.528  12.491

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.805      1.337   8.081 2.63e-10 ***
SES            2.509      1.765   1.421    0.162
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 7.51 on 45 degrees of freedom
Multiple R-squared: 0.04295,       Adjusted R-squared: 0.02168
F-statistic:  2.02 on 1 and 45 DF,  p-value: 0.1622

> (lm1$coefficients)

(Intercept)         SES
  10.805132    2.508582
```

8.2.2 Linear model of school 1288

```
> lm2 <- lm(MathAch ~ SES, data = subset(MathAchieve, School ==
+     1288))
> summary(lm2)

Call:
lm(formula = MathAch ~ SES, data = subset(MathAchieve, School ==
    1288))
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-15.648  -5.700   1.047   4.420   9.415 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   13.115      1.387   9.456 2.17e-09 ***
SES            3.255      2.080   1.565    0.131    
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 6.819 on 23 degrees of freedom
Multiple R-squared: 0.09628,	Adjusted R-squared: 0.05699 
F-statistic:  2.45 on 1 and 23 DF,  p-value: 0.1312
```

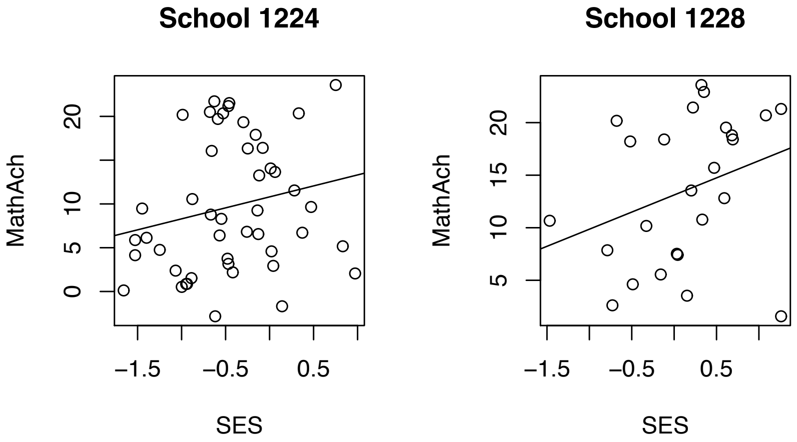
```
> (lm2$coefficients)
```

```
(Intercept)         SES 
  13.114937    3.255449 
```

The commands above show the computations for the linear model, achievement as a function of SES, for each of the two schools. Let's visualize these two fits.

8.2.3 Visualization of the linear models for schools 1224 and 1288

```
> multiplot(1, 2)
> plot(MathAch ~ SES, data = subset(MathAchieve, School == 1224),
+     main = "School 1224")
> abline(lm1$coefficients)
> plot(MathAch ~ SES, data = subset(MathAchieve, School == 1288),
+     main = "School 1228")
> abline(lm2$coefficients)
```



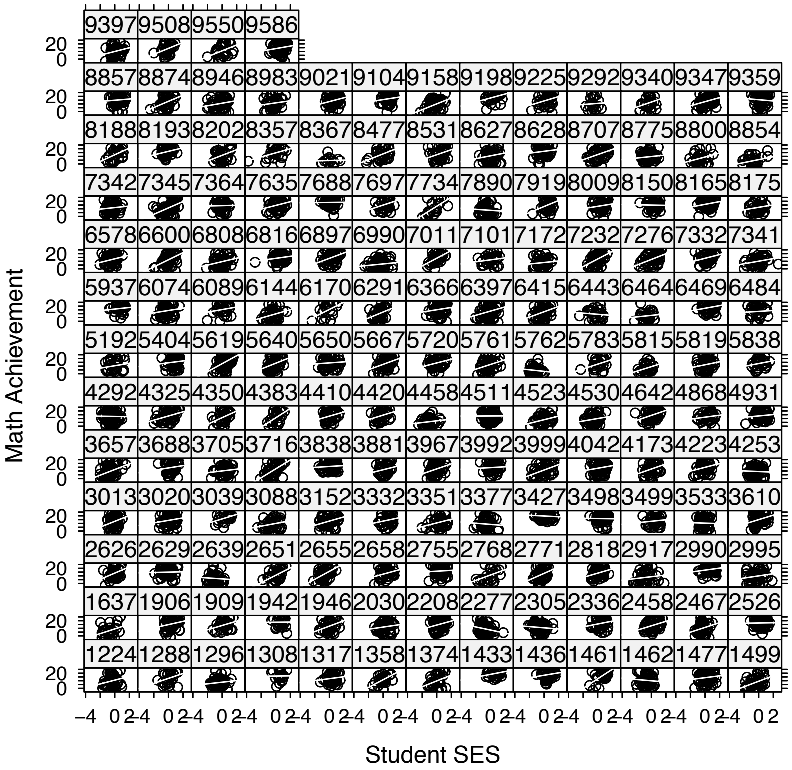
A detail to notice: the x-axis is centered around 0. This is because each SES score is "centered" by subtracting the mean SES score for a school from the raw SES score. The advantage of centering is that it makes the intercept more meaningful (the intercept will now be the mean achievement for the school).

So we have fit two regression lines, one for each of the two schools, and for each school the equation looks like this (taking the centering idea into account in the equation):

$$Y_i = \beta_o + \beta_1(X_i - \bar{X}) + \epsilon_i \tag{8.3}$$

Now, obviously we can fit separate regression lines for *each* of the schools in the dataset. We can visualize these separate fits quite easily:

```
> library(lattice)
> (print(xyplot(MathAch ~ SES | factor(School), MathAchieve, xlab = "Student SES",
+     ylab = "Math Achievement", panel = drawfittedline, scales = scalelist)))
```



8.2.4 Linear model for each school

The command below shows how to compute separate regression lines in R for each school. If you print the result `lme1` you will see that it contains the intercepts and slopes for each school. To save space I print out only the two schools' intercepts and slopes that we just computed above. Compare these intercepts and slopes to what we computed earlier–they're identical.

```
> lme1 <- lmList(MathAch ~ SES | School, MathAchieve)
> lme1 <- lmList(MathAch ~ 1 + SES | School, MathAchieve)
> lme1$"1224"

Call:
lm(formula = formula, data = data)

Coefficients:
(Intercept)          SES
     10.805        2.509

> lme1$"1288"

Call:
lm(formula = formula, data = data)

Coefficients:
```

```
(Intercept)          SES
     13.115        3.255
```

Notice an important point: we can do a t-test on the list of intercepts and slopes to determine if they are significantly different from zero:

```
> t.test(coef(lme1)[1])

        One Sample t-test

data:  coef(lme1)[1]
t = 57.2824, df = 159, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 12.16658 13.03551
sample estimates:
mean of x
 12.60104

> t.test(coef(lme1)[2])

        One Sample t-test

data:  coef(lme1)[2]
t = 17.0747, df = 159, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1.946981 2.456300
sample estimates:
mean of x
 2.201641
```

The separate regression lines for each school j can be characterized as a single system of equations ($\bar{X}_{.j}$ refers to the mean of school j):

$$Y_{ij} = \beta_{oj} + \beta_{1j}(X_{ij} - \bar{X}_{.j}) + \epsilon_{ij} \tag{8.4}$$

We now have a separate intercept and slope for each school: β_{oj}, and β_{1j}. These intercepts and slopes have a variance, and they covary (Covariance: $Cov(X,Y) = \frac{\sum(X-\bar{x})(Y-\bar{x})}{n-1}$). Covariance allows us to characterize how things, well, covary. This means that when one increases, the other could also increase (covariance positive); or when one increases, the other could decrease (negative covariance); or there could be no such relationship (zero covariance).

Let's give the different variances above a name:

- $\text{Var}(\beta_{oj}) = \tau_{00}$
- $\text{Var}(\beta_{1j}) = \tau_{11}$
- $\text{Cov}(\beta_{0j}, \beta_{1j}) = \tau_{01}$

These three τs allow us to compute the population correlation between means and slopes (the code chunk below shows this relationship in R):

$$Cor(\beta_{oj}, \beta_{1j}) = \frac{\tau_{01}}{\sqrt{\tau_{00}\tau_{11}}} \tag{8.5}$$

These correlations are interesting for the following reason. The effectiveness and equity for each school j is described by the pair (β_{0j}, β_{1j}). If the intercept for a school has a high value this means it's an effective school (in terms of math achievement), and if the slope is small then this means that the school is more equitable across SESs.

We can now ask the following informative question: is there a relationship between individual school means (i.e. intercepts) and slopes? Are schools that have higher overall effectiveness also more equitable?

Consider the covariance of (β_{0j}, β_{1j}). If τ_{01} is positive, this means that increasing effectiveness makes schools less equitable. In R, we can ask this question directly. Just for fun, we also compute the covariance mentioned above, and verify the relationship between the various τs.

```
> lme1 <- lmList(MathAch ~ SES | School, MathAchieve)
> intercepts <- coef(lme1)[1]
> slopes <- coef(lme1)[2]
> (cov(intercepts, slopes))
```

```
                  SES
(Intercept) 0.3555241
```

```
> (cov(intercepts, slopes)/sqrt(var(intercepts) * var(slopes)))
```

```
                   SES
(Intercept) 0.07833762
```

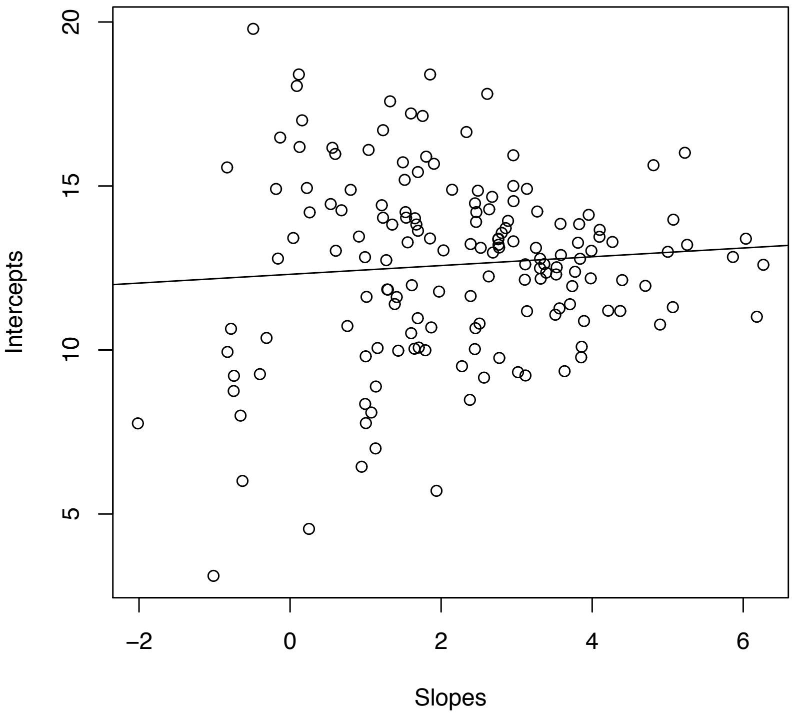
```
> (cor(intercepts, slopes))
```

```
                   SES
(Intercept) 0.07833762
```

It appears that $\tau_{01} = 0.36$. Greater effectiveness of a school means greater inequity across socio-economic statuses.

Let's also take a graphical look at how the intercepts and slopes across schools relate to each other:

```
> intslopes <- data.frame(intercepts, slopes)
> colnames(intslopes) <- c("Intercepts", "Slopes")
> plot(Intercepts ~ Slopes, intslopes)
> lm.intslopes <- lm(Intercepts ~ Slopes, data = intslopes)
> abline(coefficients(lm.intslopes))
```



8.3 Predictors of achievement

It turns out that we also know which school is Catholic and which not. We can pull up a related dataset that provides that information and merge it with the one we have:

```
> MathAchSchool <- read.table("mathachschool.txt")
> colnames(MathAchSchool) <- c("School", "Size", "Sector", "PRACAD",
+     "DISCLIM", "HIMINTY", "MEANSES")
> MathScores <- merge(MathAchieve, MathAchSchool, by = "School")
```

Suppose that we have a hypothesis (two, actually): Catholic schools are more effective and more egalitarian than public schools (See the Bryk and Raudenbush book for details on why this might be so). How can we find out if these two hypotheses are valid?

Basically, we need to be define (a) a model which predicts effectiveness as a function of the school type; (b) a model which predicts equitableness as a function of school type. So we need one equation to predict the intercept β_{oj} and another to predict the slope β_{1j}, and we need a way to specify school type. We can characterize each school j as being Catholic or Public by defining a variable W_j that has value 0 for public school and 1 for Catholic school.

Note that R has a default dummy coding from the data in such cases:

```
> (contrasts(MathScores$Sector))
```

```
         Public
Catholic      0
Public        1
```

The intercept and slope of each school can now be characterized as follows:

$$\beta_{oj} = \gamma_{00} + \gamma_{01} W_j + \epsilon_{oj} \tag{8.6}$$

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + \epsilon_{1j} \tag{8.7}$$

- γ_{00} is the mean achievement for catholic schools.
- γ_{01} is the mean achievement difference between Catholic and Public schools.
- γ_{10} is the average SES achievement slope for public schools.
- γ_{11} is the mean difference in SES-achievement slopes between Catholic and public schools.
- ϵ_{0j} is the effect of school j on mean achievement holding W_j constant.
- ϵ_{1j} is the effect of school j on the SES-achievement slope holding W_j constant.

Now, obviously we cannot estimate the above two linear models in the usual way; in order to do that the slopes and intercepts would have to have been dependent variables that had been *observed* in the data. The intercepts and slopes we have in the above code chunks are *estimated* values, not observed ones. So what to do now? Our goal was to use the above equations to evaluate the hypotheses about effectiveness and equitableness as a function of school-type.

Consider the model we saw a bit earlier:

$$Y_{ij} = \beta_{oj} + \beta_{1j}(X_{ij} - \bar{X}_{.j}) + \epsilon_{ij} \tag{8.8}$$

We can use this equation to assemble one giant predictor equation which shows achievement scores as a function of school type. We can do this by just substituting the equations for the intercepts and slopes.

$$\begin{align}
Y_{ij} =& \beta_{oj} + \beta_{1j}(X_{ij} - \bar{X}_{.j}) + \epsilon_{ij} \tag{8.9}\\
=& \gamma_{00} + \gamma_{01} W_j + \epsilon_{oj} + (\gamma_{10} + \gamma_{11} W_j + \epsilon_{1j})(X_{ij} - \bar{X}_{.j}) + \epsilon_{ij} \tag{8.10}\\
=& \gamma_{00} + \gamma_{01} W_j + \epsilon_{oj} + \gamma_{10}(X_{ij} - \bar{X}_{.j}) + \gamma_{11} W_j (X_{ij} - \bar{X}_{.j}) + \epsilon_{1j}(X_{ij} - \bar{X}_{.j}) + \epsilon_{ij} \tag{8.11}\\
=& \gamma_{00} + \gamma_{01} W_j + \gamma_{10}(X_{ij} - \bar{X}_{.j}) + \gamma_{11} W_j (X_{ij} - \bar{X}_{.j}) + \epsilon_{oj} + \epsilon_{1j}(X_{ij} - \bar{X}_{.j}) + \epsilon_{ij} \tag{8.12}
\end{align}$$

The last line just rearranges the random errors to appear at the end of the equation.

Notice that this is no longer a simple linear model: for that to be true the random errors would have to be iid. The random errors have a much more complex structure: $\epsilon_{oj} + \epsilon_{1j}(X_{ij} - \bar{X}_{.j}) + \epsilon_{ij}$ Therefore ordinary least squares will not help us find parameter estimates here.

8.4 The levels of the complex linear model

The combined model in equation is composed of two parts:

The level-1 model

$$Y_{ij} = \beta_{oj} + \beta_{1j}(X_{ij} - \bar{X_{.j}}) + \epsilon_{ij} \tag{8.13}$$

The level-2 models

$$\beta_{oj} = \gamma_{00} + \gamma_{01} W_j + \epsilon_{oj} \tag{8.14}$$

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + \epsilon_{1j} \tag{8.15}$$

We will call the β parameters in the Level-1 model the Level-1 coefficients, and γ the Level-2 coefficients.

The above model has a single Level-1 predictor (X_{ij}) and a single Level-2 predictor (W_j). Such a model is called a hierarchical linear model or HLM; hierarchical because of its different levels. Any dataset that has a grouped structure has this hierarchical structure. In psycholinguistics, within-subject, repeated measures experiments are a good example. In dialectology, subjects grouped within geographical regions are another example. And so on.

Now let's look at how this kind of a model is computed in R:

```
> lme1.fm <- lmer(MathAch ~ SES + Sector + (1 + SES | School),
+     MathScores)
> summary(lme1.fm)

Linear mixed model fit by REML
Formula: MathAch ~ SES + Sector + (1 + SES | School)
   Data: MathScores
   AIC   BIC logLik deviance REMLdev
 46616 46664 -23301    46597   46602
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 School   (Intercept)  3.96385 1.99094
          SES          0.43431 0.65902  0.550
 Residual             36.80088 6.06637
Number of obs: 7185, groups: School, 160

Fixed effects:
             Estimate Std. Error t value
(Intercept)   14.0138     0.2604   53.82
SES            2.3853     0.1179   20.24
SectorPublic  -2.5409     0.3445   -7.37

Correlation of Fixed Effects:
            (Intr) SES
SES          0.098
SectorPublc -0.741  0.079
```

We can extract the fixed effect coefficients:

```
> fixef(lme1.fm)

 (Intercept)          SES SectorPublic
   14.013800     2.385342    -2.540925
```

The estimates for the random effects are shown below:

```
Random effects:
 Groups   Name Variance Std.Dev. Corr
 School         3.96373 1.99091
                0.43453 0.65919  0.549
 Residual      36.80079 6.06637
```

What these mean:

- Var(School)=3.96373 (This is the τ_{00} we saw earlier for β_{0j}
- 0.43453 is the variance of the slopes (τ_{11}).
- Var(ϵ_{ij})= 36.80079
- Cor(β_{0j}, β_{1j})=0.549

Appendix: random variables

Recall our rain example. Consider now only four instances of four drops of rain. What is the probability of there being 0 . . . 4 Right-stone hits?

X, the number of R-stone hits	0	1	2	3	4
Probability of R-stone hits	?	?	?	?	?

X above is referred to as RANDOM VARIABLE, and is defined as below:

Definition:
A random variable is a real value function defined on a sample space. I.e., X(e) = some real value, e an event in the sample space.

A more general representation of the four-drop scenario:

X	x_1	x_2	x_3	. . .	x_k
Probability	$f(x_1)$	$f(x_2)$	$f(x_3)$	. . .	$f(x_k)$

The $f(x_1) \dots f(x_k)$ is the **probability distribution** (constrast this with the **frequency distribution** that we've been plotting in past simulations).

Question: $\sum_{i=1}^{k} f(x_i) =???$

.1 The role of the probability distribution in statistical inference

Suppose there are two products A and B, and we want to know if there is a greater preference for one or another. We start with the assumption (the null hypothesis) that both are equally preferred. If this were so, in a random survey of customers, the theoretical probability of one or the other product being preferred is 0.5. So, we can create the **a priori** probability distribution when, say, 4 customers make a choice (call this one observation):

X, the number of A product preferences	0	1	2	3	4
Probability	1/16	4/16	6/16	4/16	1/16

Suppose now that we run an experiment using four randomly selected customers, and we get all 4 customers choosing A. The probability of this happening is 0.0625. This can either mean:

- The preference for A is not the same as the preference for B.
- The preferences for A and B are identical, but an unlikely event occurred.

.2 Expectation

Notice a funny thing: we can calculate the mean of a bunch of numbers $x_1, \dots, x_k$ in two ways:

Let's compute the mean of 0,2,2,1,2,3,0,1,2,1

1. Just use the usual formula: $\frac{0+2+2+1+2+3+0+1+2+1}{10} = 1.4$
2. Count the relative frequency of occurrence of each number, and multiply by that number:

   $0 \times \frac{2}{10} + 2 \times \frac{4}{10} + 1 \times \frac{3}{10} + 3 \times \frac{1}{10} = 1.4 = \sum_{i=1}^{k} x_i \times \text{RelativeFrequency}(x_i)$

That was a computation from a **sample**. Now think of the binomial situation (Heads or Tails). Here, X=0,1. Suppose we want to know the "mean" given the prior probability of a heads $p = 0.5$. Here's how we can compute the **population** mean:

$$\mu_X = \sum_{i=1}^{k} (\text{Value} \times \text{Probability}) \tag{16}$$

This population mean is called the EXPECTATION.

Definition of expectation E(X):

$E(X) = \sum_{i=1}^{k} x_i f(x_i)$

To understand the origin of the term "expectation", think of the situation where you were gambling in a casino with a coin, and for each heads you get 50 Euro-cents (this is equivalent to I-don't-know how many US dollars at the time of writing), but you have to pay a playing fee of c Euros for each throw. Then, assuming that the coin is fair, your expected gain is $0 \times 0.5 + 0.5 \times 0.5 = 0.25$ by the above definition. If the casino charges you 1 Euro, the expected gain of the casino in the long run is 50 cents per game.

Note the similarity with the sample mean we just computed using the same formula above, but also note that $\mu = E(X)$ is the population mean here, and is computed from the theoretical probability distribution (which depends on our assumptions about a particular situation like a coin toss), not from any sample.

Consider again the product A versus product B situation.

X, the number of A product preferences	0	1	2	3	4
Probability	1/16	4/16	6/16	4/16	1/16

Suppose for some reason we want to know the Expectation of a function of a random variable, e.g., $g(X) = (X-2)^2$. There are two ways to do this:

Method 1: Compute distinct values of $(X-2)^2$ and then compute the probabilities of each of these values. This function of the original random variable X is itself a random variable Y now – *but the probabilities associated with Y's values is a function of X's probabilities.* Then apply the definition of Expectation to Y.

- When X=0, $(X-2)^2 = 4$. Probability of X=0: 1/16
- When X=1, $(X-2)^2 = 1$ Probability of X=1: 4/16
- When X=2, $(X-2)^2 = 0$ Probability of X=2: 6/16
- When X=3, $(X-2)^2 = 1$ Probability of X=3: 4/16

- When X=4, $(X-2)^2 = 4$ Probability of X=4: 1/16

$Y = (X-2)^2$	0	1	4
$p(Y = y_i)$	3/8	4/8	1/8
$y_i \times p(y_i)$	0	4/8	4/8

Method 2: Compute g(X) and then multiply each with the $f(x_i)$:

X	0	1	2	3	4
$f(x_i)$	1/16	4/16	6/16	4/16	1/16
$(x_i-2)^2$	4	1	0	1	4
$(x_i-2)^2 f(x_i)$	4/16	4/16	0	4/16	4/16

Notice that the expection is the same when computed with the two methods: $\sum = y_i \times f(y_i) = \sum (x_i - 2)^2 f(x_i) = 1$.

The expectation computed by either of the same methods is always going to yield the same result. Reason: In method one we are doing computations like $g(x) \times (f(x_1) + f(x_3))$ while in method 2 we are doing computations like $g(x) \times f(x_1) + g(x) \times f(x_3)$. These will always yield the same result.

This motivates the definition of the expectation of a function $g(X)$ of a random variable (so: a function of a function – remember that the random variable is really a function).

Definition:
$E(g(X)) = \sum g(x_i) f(x_i)$

.3 Properties of Expectation

(i) $E(a) = a$

(ii) $E(bX) = b \times E(X)$

(iii) $E(X + a) = E(X) + a$

(iv) $E(a + bX) = a + b \times E(X)$
Proof:

$$\begin{aligned} E(a+bX) &= \sum (a + bx_i) f(x_i) \ldots \text{see above definition of E(g(X))} \quad (17) \\ &= \sum a f(x_i) + \sum b x_i f(x_i) \quad (18) \\ &= a \sum f(x_i) + b \sum x_i f(x_i) \quad (19) \\ &= a \times 1 + bE(X) \ldots \text{because} \sum f(x_i) = 1 \quad (20) \\ &= a + bE(X) \quad (21) \end{aligned}$$

(v) $E(a + bX + cX^2) = a + b \times E(X) + c \times E(X^2)$
Proof: see homework assignment on page 150

.4 Variance

We have worked out so far that $\mu = E(X)$. In order to characterize spread about a mean value, we can use deviations from the mean: $X - \mu$. But this will necessarily give us the answer 0 (see page 3 for why if you've forgotten).

Suppose $X = x_1, \dots, x_k$. Then:

$$\begin{aligned} E(X-\mu) &= \sum (x_i - \mu) f(x_i) && (22) \\ &= 0 && (23) \end{aligned}$$

So, as before, we square the deviations, and take that as the measure of spread, and call this, as before, Variance:

$$\begin{aligned} Var(X) &= E((X-\mu)^2) && (24) \\ &= E(X^2 - 2\mu X + \mu^2) && (25) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \dots \text{from property (v) above} && (26) \\ &= E(X^2) - 2\mu^2 + \mu^2 && (27) \\ &= E(X^2) - \mu^2 && (28) \end{aligned}$$

And if we scale it down to the dimensions of the mean (as before), we get the standard deviation of the population:

$$sd(X) = \sqrt{Var(X)} = \sqrt{E(X^2) - \mu^2} = \sigma_X \tag{29}$$

.5 Important properties of variance

(i) Var(X+a)=Var(X)

(ii) $Var(bX) = b^2 Var(X)$
Proof:

$$\begin{aligned} Var(bX) &= E((bX)^2) - (E(bX))^2 && (30) \\ &= E(b^2X^2) - (E(bX))^2 && (31) \\ &= b^2E(X^2) - (E(bX))^2 \dots \text{property (ii) of Expectation} && (32) \\ &= b^2E(X^2) - (bE(X))^2 \dots \text{property (v) of Expectation} && (33) \\ &= b^2E(X^2) - b^2E(X)^2 && (34) \\ &= b^2(E(X^2) - E(X)^2) \dots \text{factoring out } b^2 && (35) \\ &= b^2(Var(X)) && (36) \end{aligned}$$

.6 Mean and SD of the binomial distribution

X	0	1
$f(x_i)$	q	p

$$E(X) = 0 \times q + 1 \times p = p \tag{37}$$

$$E(X^2) = 0 \times q + 1 \times p = p \tag{38}$$

$$\begin{aligned} Var(X) &= E(X^2) - \mu^2 &(39)\\ &= p - p^2 &(40)\\ &= p(1-p) &(41)\\ &= pq &(42)\end{aligned}$$

The above is for one observation. For $n > 1$ observations: $X = X_1 + \cdots + X_n$. It follows that (assuming independence of each observation):

$$E(X) = E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n) = np \tag{43}$$

Similarly,

$$Var(X) = E(X_1 + \cdots + X_n) = npq \tag{44}$$

.7 Sample versus population means and variances

Assume that the population mean: μ_X and population variance: σ_X We know that $\bar{X} = \frac{X_1+\cdots+X_n}{n}$. From the properties of Expectation and Variance discussed above, we can deduce the following two facts:

$$\begin{aligned} E(\bar{X}) &= \frac{1}{n}E(X_1 + \cdots + X_n) &(45)\\ &= \frac{1}{n}(E(X_1) + \cdots + E(X_n)) &(46)\\ &= \frac{1}{n}(\mu + \cdots + \mu) &(47)\\ &= \frac{1}{n}(n \times \mu) &(48)\\ &= \mu &(49)\end{aligned}$$

The above is the sampling distribution of the sampling mean (the distribution of the mean values when we repeatedly sample from a population).

$$
\begin{align}
Var(\bar{X}) &= Var(\frac{1}{n}Var(X_1 + \dots X_n)) \tag{50}\\
&= \frac{1}{n^2}(Var(X_1) + \cdots + Var(X_n)) \tag{51}\\
&= \frac{1}{n^2}(\sigma^2 + \cdots + \sigma^2) \tag{52}\\
&= \frac{1}{n^2}(n \times \sigma^2) \tag{53}\\
&= \frac{\sigma^2}{n} \tag{54}
\end{align}
$$

In other words:

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}} \tag{55}$$

This is the standard deviation of the sampling distribution of the sampling means. The next chapter will look at this distribution and its properties in detail, and this will lead us to our first hypothesis test – the t-test.

.8 Exercise

Prove the following statement:
$E(a + bX + cX^2) = a + b \times E(X) + c \times E(X^2)$

.9 Brief aside: Random errors are your friends

A sample STATISTIC (like the SAMPLE COUNT – e.g., the number of R-stone hits, the number of red balls in a sample) is a totally unreliable estimator of the *exact* population parameter. **A key insight**: If you average over many samples, the mean of the sampling distribution will be *close* to the true parameter. The best mathematicians failed to understand this.

A statistic i is *always* wrong by some error amount ϵ_i. So, if Θ is the (population) parameter, and $\hat{\Theta}_i$ an estimate from a sample X_i:

$$\hat{\Theta}_i = \Theta + \epsilon_i \tag{56}$$

$\hat{\Theta}_i$ is called a POINT ESTIMATOR. If you take the mean of the point estimator, you get:

$$
\begin{align}
\hat{\Theta}_i &= \Theta + \epsilon_i \tag{57}\\
Mean(\hat{\Theta}_i) &= Mean(\Theta + \epsilon_i) \tag{58}\\
&= Mean(\Theta) + Mean(\epsilon_i) \tag{59}
\end{align}
$$

Although averaging does not *seem* to get rid of the error . . . :

$$Mean(\hat{\Theta}_i) = Mean(\Theta) + Mean(\epsilon_i) \tag{60}$$

. . . actually it tends to.

Probability theorists failed to see this, even after the normal distribution had been discovered – more or less by accident – by Abraham de Moivre (“the gambler consultant”) in 1733.

Some historical context: Euler was trying to estimate orbital parameters for Jupiter and Saturn, and reasoned that since the observations were erroneous, more observations would only muddy the waters further. His (and others') key misunderstanding was the assumption that all errors are equally likely. In fact, the normal distribution shows that the worse the errors, the less likely it is to occur. Interestingly, early empiricists (Hipparchus, 2nd c. BC; Tycho Brahe, 16th c.) had already noticed this in their data, but had no theoretical justification for it.

This insight is why the normal distribution is important. Recall that our simulations earlier suggest that the worst errors (deviations from the true population parameter) are extremely unlikely (if the error is random). We saw that the normal distribution tells us that the mean of the observations is erroneous but *points to* the actual population parameter – random error is not fatal, but actually reveals something useful.

.10 Unbiased estimators

Any statistic that allows us to "zero in" accurately on a target parameter is called an UNBIASED ESTIMATOR.

More formally:

> A statistic used to estimate a parameter is unbiased iff the mean of its sampling distribution is equal to the true value of the parameter being estimated.

This fact is what makes statistical inference possible.

.11 Summing up

Here are some shortcuts we derived in this chapter (page 148): To compute mean and deviation of a sample count X that has binomial distribution B(n,p):

$$\mu_X = n \times p \tag{61}$$

$$\sigma_X = \sqrt{n \times p(1-p)} \tag{62}$$

Suppose we have a population of 1000 students. 600 male, 400 female. We take one random sample of 40 students. How many females are there?

```
> females <- rbinom(40, 1, 0.4)
> females

 [1] 0 1 0 1 1 0 0 0 0 0 1 1 1 1 1 0 1 0 1 1 1 0 0 0 0 1 1 0 1 1 0 0 0 0 0 0 0 0
[39] 0 0

> sum(females)

[1] 16
```

We know that the 95% CI is about $2 \times \sigma_x$. Let's write a function to compute 95 percent CIs for a sample x.

```
> populationsize <- 1000
> samplesize <- 40
> p <- 0.4
> compute95CIpopulation <- function(populationsize, samplesize,
+     p) {
+     females <- rbinom(samplesize, 1, 0.4)
+     samplesumfemales <- sum(females)
+     sdsample <- sqrt(samplesize * p * (1 - p))
+     sample95CI <- c(samplesumfemales - 2 * sdsample, samplesumfemales +
+         2 * sdsample)
+     population95CI <- populationsize/samplesize * sample95CI
+     print(population95CI)
+ }
```

A 95% CI means: if we repeatedly take samples of a given size, 95% of the time the population mean will lie within it.

Recall that we *know* the population mean here: 400. So, just for fun, let's sample it a couple of time to see what happens. Occasionally (about 5% of the time) we should get an interval which does not contain the population mean. (Note, if you run this code yourself, the results will vary. If that is not clear, you need to start more or less at the beginning of these notes.)

```
> compute95CIpopulation(1000, 40, 0.4)

[1] 320.0807 629.9193

> compute95CIpopulation(1000, 40, 0.4)

[1] 245.0807 554.9193

> compute95CIpopulation(1000, 40, 0.4)

[1] 245.0807 554.9193

> compute95CIpopulation(1000, 40, 0.4)

[1] 245.0807 554.9193

> compute95CIpopulation(1000, 40, 0.4)

[1] 170.0807 479.9193

> compute95CIpopulation(1000, 40, 0.4)

[1] 220.0807 529.9193

> compute95CIpopulation(1000, 40, 0.4)

[1] 245.0807 554.9193

> compute95CIpopulation(1000, 40, 0.4)

[1] 195.0807 504.9193

> compute95CIpopulation(1000, 40, 0.4)

[1] 395.0807 704.9193

> compute95CIpopulation(1000, 40, 0.4)

[1] 245.0807 554.9193
```

```
> compute95CIpopulation(1000, 40, 0.4)

[1] 145.0807 454.9193

> compute95CIpopulation(1000, 40, 0.4)

[1] 245.0807 554.9193

> compute95CIpopulation(1000, 40, 0.4)

[1] 345.0807 654.9193

> compute95CIpopulation(1000, 40, 0.4)

[1] 245.0807 554.9193
```

References

Berger, J. L., & Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. Statistical Science, 11(4), 283–302.

Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge, UK: Cambridge University Press.

Hauck, W. W., & Anderson, S. (1996, November). [bioequivalence trials, intersection-union tests and equivalence confidence sets]: Comment. Statistical Science, 11(4), 303.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. The American Statistician, 55(1), 19-24.

Leisch, F. (2002). Sweave: Dynamic generation of statistical reports using literate data analysis. In W. Härdle & B. Rönz (Eds.), Compstat 2002 — proceedings in computational statistics (pp. 575–580). Physica Verlag, Heidelberg. (ISBN 3-7908-1517-9)

Maxwell, S. E., & Delaney, H. D. (2000). Designing experiments and analyzing data. Mahwah, New Jersey: Lawrence Erlbaum Associates.

McBride, G. B. (1999). Equivalence tests can enhance environmental science and management. Australian and New Zealand Journal of Statistics, 41(1), 19-29.

Ray, W. J. (2000). Methods: Toward a science of behavior and experience. Belmont, CA: Wadsworth/Thomson Learning.

Rosen, K. H. (1994). Discrete mathematics and its applications (Third Edition ed.). New York: Mc-Graw Hill, Inc.

Stegner, B. L., Bostrom, A. G., & Greenfield, T. K. (1996). Equivalence testing for use in psychosocial and services research: An introduction with examples. Evaluation and Program Planning, 19(3), 193-198.

Index