Chapter 7 Making population inferences

Adapted by Abigail Noyce from *Learning Statistics with R*, Danielle Navarro.[13]

The process of induction is the process of assuming the simplest law that can be made to harmonize with our experience. This process, however, has no logical foundation but only a psychological one. It is clear that there are no grounds for believing that the simplest course of events will really happen. It is an hypothesis that the sun will rise tomorrow: and this means that we do not know whether it will rise.
– Ludwig Wittgenstein[14]

To a lot of people, collecting some data and summarizing it is is all there is to statistics: it's about calculating averages, collecting all the numbers, drawing pictures, and putting them all in a report somewhere. Kind of like stamp collecting, but with numbers. However, statistics covers much more than that. In fact, descriptive statistics is one of the smallest parts of statistics, and one of the least powerful. The bigger and more useful part of statistics is that it can let you make inferences about the larger world, beyond just the data you collected.

Inferential statistics provides the tools that we need to answer these sorts of questions, and since these kinds of questions lie at the heart of the scientific enterprise, they take up the lion's share of every introductory course on statistics and research methods. The role of *descriptive* statistics is to concisely summarize what we do know. In contrast, the purpose of *inferential* statistics is to "learn what we do not know from what we do". Inferential statistics are are traditionally divided into two "big ideas": estimation and hypothesis testing, but before we can explore either, we need to first understand the ideas behind sampling.

7.1 Sampling

7.1.1 Samples and populations

In order to make inferences from our data to the broader world, we need to make some fairly general assumptions about the relationship between them. This is where sampling theory comes in. If probability theory is the foundation upon which all statistical theory builds, sampling theory is the frame around which you can build the rest of the house. Sampling theory plays a huge role in specifying the assumptions upon which your statistical inferences rely. And in order to talk about "making inferences" the way statisticians think about it, we need to be a bit more explicit about what it is that we're drawing inferences from (the sample) and what it is that we're drawing inferences about (the population).

In almost every situation of interest, what we have available to us as researchers is a *sample* of data. The data set available to us is finite, and incomplete. We can't possibly get every person in the world to do our experiment; a polling company doesn't have the time or the money to ring up every voter in the country etc.

The sample is a concrete thing. You can open up a data file, and there's the data from your sample. A *population*, on the other hand, is a more abstract idea. It refers to the set of all possible people, or all possible observations, that you want to draw conclusions about, and is generally much bigger than the sample. In an ideal world, every research study would begin with a clearly specified population of interest.

Sometimes it's easy to state the population of interest. For instance, in a political poll the population consists of all registered voters at the time of the study – millions of people. The sample might then be a set of 1000 people who all belong to that population. In a typical psychological experiment, on the other hand, determining the population of interest is a bit more complicated. Suppose I run an experiment using 100 undergraduate students as my participants. My goal, as a cognitive scientist, is to try to learn something about how the mind works. So, which of the following would count as "the population"?

- All of the undergraduate psychology students at the University of Adelaide
- Undergraduate psychology students in general, anywhere in the world
- Australians currently living

- Australians of similar ages to my sample
- Anyone currently alive
- Any human being, past, present or future
- Any biological organism with a sufficient degree of intelligence operating in a terrestrial environment
- Any intelligent being

Each of these defines a real group of entities, all of which might be of interest to me as a cognitive scientist. There is no hard and fast rule about defining the population of interest, or determining how broadly your statistical estimates can be generalized. It will depend on which of the characteristics of your sample are necessary for your results. For example, we generally don't expect eye color to be related to perceptual sensitivity for pitch, and so even if my sample consisted only of brown-eyed humans, I would be comfortable generalizing to a population that included all eye colors. On the other hand, if all of my participants are music majors, that probably is related, and I should be cautious generalizing to a population without that specific background.

7.1.2 Simple random samples

The relationship between a sample and a population depends on the procedure by which the sample was selected. This procedure is referred to as a sampling method, and it is important to understand why it matters.

To keep things simple, let's imagine that we have a bag containing 10 chips. Each chip has a unique letter printed on it, so we can distinguish between the 10 chips. The chips come in two colours, black and white. This set of chips is the population of interest, and it is depicted graphically on the left of Figure 7.1. As you can see from looking at the picture, there are 4 black chips and 6 white chips, but of course in real life we wouldn't know that unless we looked in the bag. Now imagine you run the following "experiment": you shake up the bag, close your eyes, and pull out 4 chips without putting any of them back into the bag. If you wanted, you could then put all the chips back in the bag and repeat the experiment, as depicted on the right hand side of Figure 7.1. Each time you get different results, but the procedure is identical in each case. The

fact that the same procedure can lead to different results each time, we refer to it as a *random process*.[15] However, because we shook the bag before pulling any chips out, it seems reasonable to think that every chip has the same chance of being selected.

A procedure in which every member of the population has the same chance of being selected is called a *simple random sample*. The fact that we did not put the chips back in the bag after pulling them out means that you can't observe the same thing twice, and in such cases the observations are said to have been sampled *without replacement*. From any of the samples shown, we can conclude that there are definitely both black and white chips in the bag, even if we're still uncertain about their relative proportions.

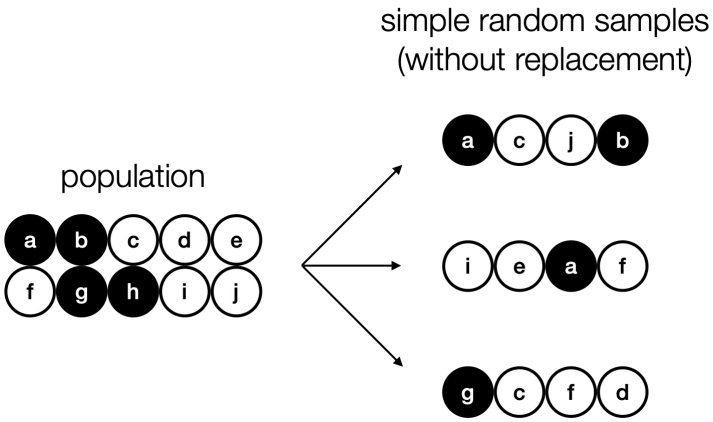

Figure 7.1: Simple random sampling without replacement from a finite population.

Consider an alternative way in which the experiment could have been run. Suppose that the researcher had opened the bag, and decided to pull out four black chips. This biased sampling scheme is depicted in Figure 7.2. What do these samples tell you about the contents of the bag? If you know that the sampling scheme is biased to select black chips, then a sample that consists of only black chips doesn't tell you very much about the population! For this reason, statisticians really like it when a data set can be considered a simple random sample, because it makes the data analysis much easier.

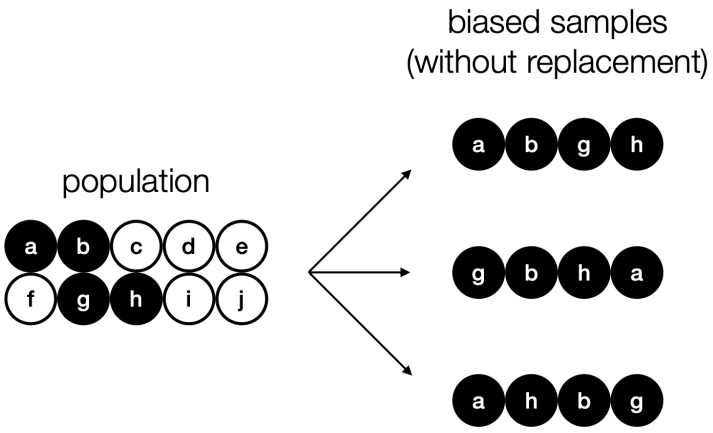

Figure 7.2: Biased sampling without replacement from a finite population.

A third procedure is worth mentioning. This time around we close our eyes, shake the bag, and pull out a chip. This time, however, we record the observation and then put the chip back in the bag. Again we close our eyes, shake the bag, and pull out a chip. We then repeat this procedure until we have 4 chips. Data sets generated in this way are still simple random samples, but because we put the chips back in the bag immediately after drawing them it is referred to as a sample *with replacement*. The difference between this situation and the first one is that it is possible to observe the same chip multiple times, as illustrated in Figure @ref(fig:7.3.

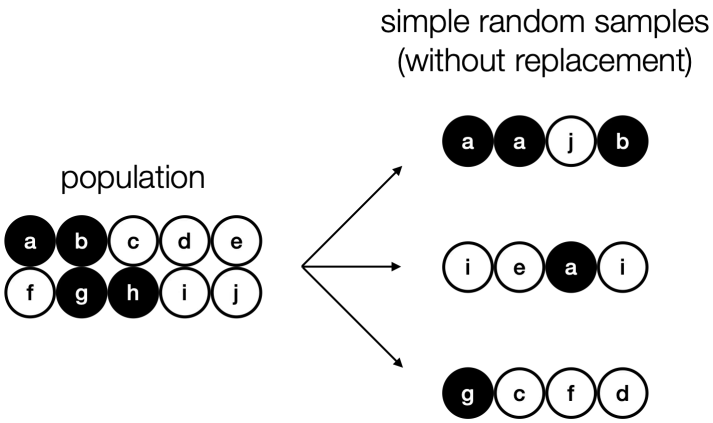

Figure 7.3: Simple random sampling with replacement from a finite population.

Most psychology experiments use sampling without replacement, because the same person is not allowed to participate in the experiment twice. While the theoretical underpinnings of inferential statistics use sampling *with* replacement, the difference between the two is too small to matter as long as the population is larger than about 10 entities. The difference between simple random samples and biased samples, on the other hand, is not such an easy thing to dismiss.

7.1.2.1 Most samples are not simple random samples

As you can see from the list of possible populations above, it is almost impossible to obtain a simple random sample from most populations of interest. When I run experiments, I'd consider it a minor miracle if my participants turned out to be a random sampling of the undergraduate psychology students at my university, even though this is by far the narrowest population that I might want to generalise to. A thorough discussion of other types of sampling schemes is beyond the scope of this book, but to give you a sense of what's out there I'll list a few of the more important ones:

- *Stratified sampling.* Suppose your population is (or can be) divided into several different subpopulations, or *strata*. Perhaps you're running a study at several different sites, for example. Instead of trying to sample randomly from the population as a whole, you instead try to collect a separate random sample from each of the strata. Stratified sampling is sometimes easier to do than simple random sampling, especially when the population is already divided into the distinct strata. It can also be more efficient that simple random sampling, especially when some of the subpopulations are rare. For instance, when studying schizophrenia it would be much better to divide the population into two[16] strata (schizophrenic and not-schizophrenic), and then sample an equal number of people from each group. If you selected people randomly, you would get so few schizophrenic people in the sample that your study would be useless. This specific kind of of stratified sampling is referred to as oversampling because it makes a deliberate attempt to over-represent rare groups.

- *Convenience sampling.* The samples are chosen in a way that is convenient to the researcher, and not selected at random from the population of interest. A common example in psychology are studies that rely on undergraduate psychology students. These samples are generally non-random in two respects: firstly, reliance on undergraduate psychology students automatically means that your data are restricted to a single subpopulation. Secondly, the students usually get to pick which studies they participate in, so the sample is a self selected subset of psychology students not a randomly selected subset. In real life, most studies are convenience samples of one form or another. This is sometimes a severe limitation, but not always.

So real world data collection tends not to involve nice simple random samples. Does that matter? A little thought should make it clear to you that it can matter if your data are not a simple random sample: just think about the difference between Figures 7.1 and 7.2. However, it's not quite as bad as it sounds. For instance, when using a stratified sampling technique you actually *know* what the bias is because you created it deliberately, often to *increase* the effectiveness of your study, and there are statistical techniques that you can use to adjust for the biases you've introduced (not covered in this course!). So in those situations it's not a problem.

More generally though, it's important to remember that random sampling is a means to an end, not the end in itself. Let's assume you've relied on a convenience sample, and as such you can assume it's biased. This is only a problem if it causes you to draw the wrong conclusions. That is, we don't need the sample to be randomly generated in *every* respect: we only need it to be *random with respect to the psychologically-relevant phenomenon of interest*. Suppose I'm doing

a study looking at working memory capacity. In study 1, I actually have the ability to sample randomly from all human beings currently alive, with one exception: I can only sample people born on a Monday. In study 2, I am able to sample randomly from the Australian population. I want to generalise my results to the population of all living humans. Which study is better? The answer, obviously, is study 1. Why? Because we have no reason to think that being "born on a Monday" has any interesting relationship to working memory capacity. In contrast, I can think of several reasons why "being Australian" might matter. Australia is a wealthy, industrialised country with a very well-developed education system. People growing up in that system will have had life experiences much more similar to the experiences of the people who designed the tests for working memory capacity. This shared experience might easily translate into similar beliefs about how to "take a test", a shared assumption about how psychological experimentation works, and so on. These things might actually matter. For instance, "test taking" style might have taught the Australian participants how to direct their attention exclusively on fairly abstract test materials relative to people that haven't grown up in a similar environment; leading to a misleading picture of what working memory capacity is.

There are two points hidden in this discussion. Firstly, when designing your own studies, it's important to think about what population you care about, and try hard to sample in a way that is appropriate to that population. In practice, you're usually forced to put up with a "sample of convenience" (e.g., psychology lecturers sample psychology students because that's the least expensive way to collect data, and our coffers aren't exactly overflowing with gold), but if so you should at least spend some time thinking about what the dangers of this practice might be.

Secondly, if you're going to criticise someone else's study because they've used a sample of convenience rather than laboriously sampling randomly from the entire human population, at least offer a *specific theory as to how this might have distorted the results*. Remember, everyone in science is aware of this issue, and does what they can to alleviate it. Merely pointing out that "the study only included people from group BLAH" is entirely unhelpful, and borders on being insulting to the researchers, who are of course aware of the issue. They just don't happen to be in possession of the infinite supply of time and money required to construct the perfect sample.

7.1.3 Sampling distributions and the central limit theorem

Setting aside the thorny methodological issues associated with obtaining a random sample, let's consider a slightly different issue. Up to this point we have been talking about populations the way a scientist might. To a psychologist, a population might be a group of people. To an

ecologist, a population might be a group of bears. In most cases the populations that scientists care about are concrete things that actually exist in the real world. Statisticians, however, are a funny lot. On the one hand, they are interested in real world data and real science in the same way that scientists are. On the other hand, they also operate in the realm of pure abstraction in the way that mathematicians do. As a consequence, statistical theory tends to be a bit abstract in how a population is defined. In much the same way that psychological researchers operationalise our abstract theoretical ideas in terms of concrete measurements, statisticians operationalise the concept of a "population" in terms of mathematical objects that they know how to work with. Usually, these objects are probability distributions.

For example, let's say we're talking about IQ scores. To a psychologist, the population of interest is a group of actual humans who have IQ scores. A statistician "simplifies" this by operationally defining the population as the probability distribution depicted in Figure 7.4. IQ tests are designed so that the average IQ is 100, the standard deviation of IQ scores is 15, and the distribution of IQ scores is normal. These values are referred to as the *population parameters* because they are characteristics of the entire population. That is, we say that the population mean is 100, and the population standard deviation is 15.

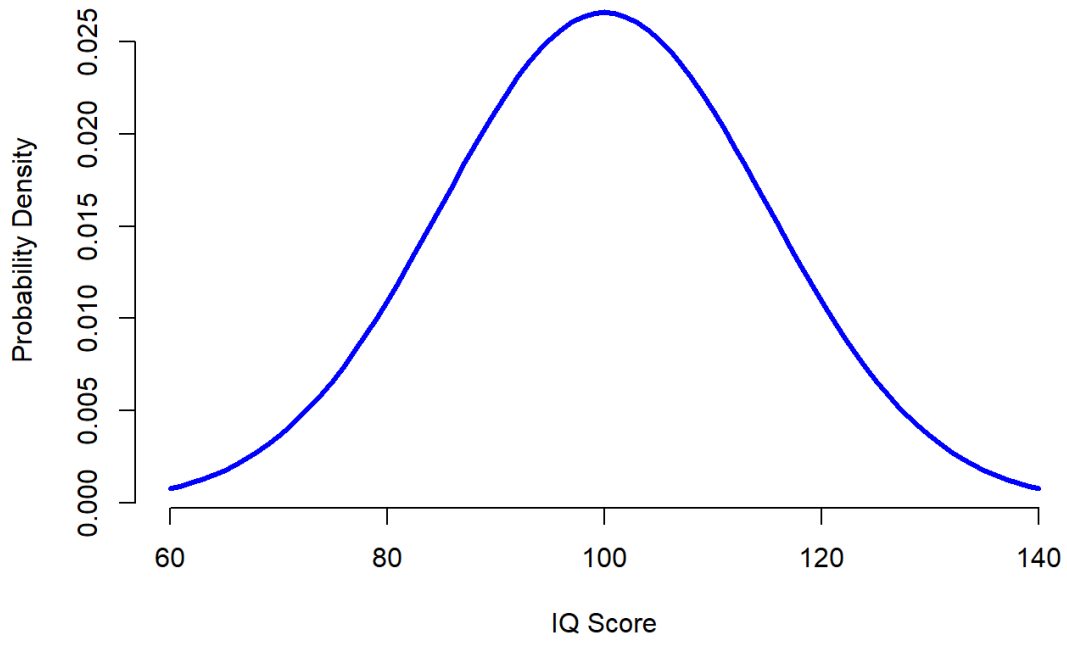

Figure 7.4: The population distribution of IQ scores.

Now suppose I run an experiment. I select 100 people at random and administer an IQ test, giving me a simple random sample from the population. My sample would consist of a collection of numbers like this:

```
iq_sample <- rnorm(n = 100,
      mean = 100,
      sd = 15)

glimpse(iq_sample)
```

```
##  num [1:100] 85.4 100.6 78.3 77.2 108.5 ...
```

If I plot this sample as a histogram, I get something like the one shown in Figure 7.5.

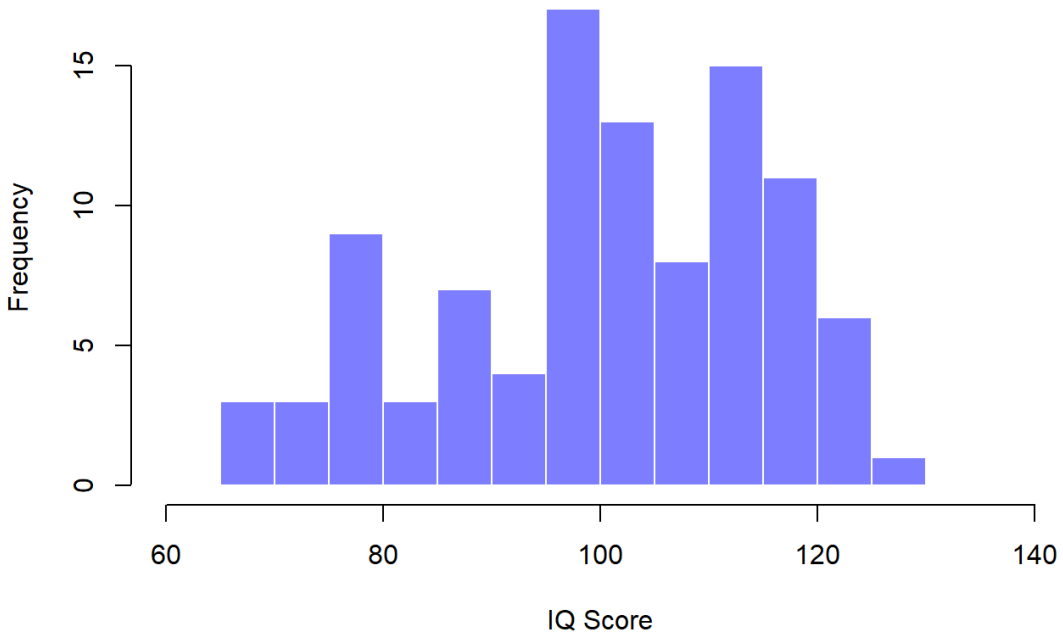

Figure 7.5: A sample of 100 observations drawn from the population of IQ scores.

As you can see, the histogram is roughly the right shape, but it's a very crude approximation to the true population distribution shown in Figure 7.4. When I calculate the mean of my sample, I get a number that is fairly close to the population mean 100 but not identical. In this case, it turns out that the people in my sample have a mean IQ of 98.6, and the standard deviation of

their IQ scores is 14.6. These results are somewhat encouraging: the sample mean is a pretty reasonable approximation to the true mean. In many scientific studies that level of precision is perfectly acceptable, but in other situations you need to be a lot more precise. If we want our sample statistics to be much closer to the population parameters, what can we do about it?

7.1.3.1 The law of large numbers

The obvious answer is to collect more data. Suppose that we ran a much larger experiment, this time measuring the IQs of 10,000 people. The histogram of this much larger sample is shown in Figure 7.6. Even a moment's inspections makes clear that the larger sample is a much better approximation to the true population distribution than the smaller one. This is reflected in the sample statistics: the mean IQ for the larger sample turns out to be 99.9, and the standard deviation is 15.1. These values are now very close to the true population.

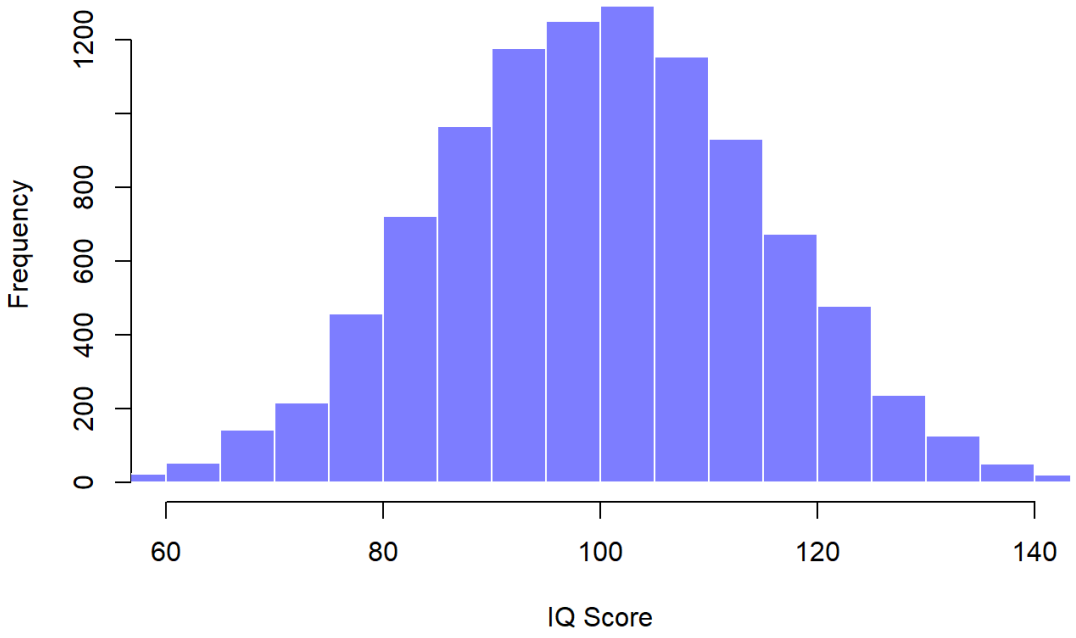

Figure 7.6: A sample of 10,000 observations drawn from the population of IQ scores.

I feel a bit silly saying this, because it's so bloody obvious that it shouldn't need to be said, but the thing I want you to take away from this is that large samples generally give you better information. In fact, it's such an obvious point that when Jacob Bernoulli – one of the founders of probability theory – formalised this idea back in 1713, he was kind of a jerk about it:

> *For even the most stupid of men, by some instinct of nature, by himself and without any instruction (which is a remarkable thing), is convinced that the more observations have been made, the less danger there is of wandering from one's goal.*

The question is, why is this so? Not surprisingly, this intuition that we all share turns out to be correct, and statisticians refer to it as the *law of large numbers*. The law of large numbers is a mathematical law that applies to many different sample statistics, but the simplest way to think about it is as a law about sample means. The law of large numbers states is that as the sample gets larger, the sample mean tends to get closer to the true population mean. Or, to say it a little bit more precisely, as the sample size "approaches" infinity (written as $N \rightarrow \infty$) the sample mean approaches the population mean ($\bar{X} \rightarrow \mu$).[17]

We won't step through a proof of this, but it's one of the most important tools for statistical theory. The law of large numbers justifies our belief that collecting more and more data will eventually lead us to the truth. For any particular data set, the sample statistics that we calculate from it will be wrong, but the law of large numbers tells us that if we keep collecting more data those sample statistics will tend to get closer and closer to the true population parameters.

The law of large numbers is a very powerful tool, but it's not going to be good enough to answer all our questions. All it gives us is a "long run guarantee". In the long run, if we were somehow able to collect an infinite amount of data, then the law of large numbers guarantees that our sample statistics will be correct. In real life, however, nobody gets to collect infinite data, and the law of large numbers is cold comfort when my actual data set has a sample size of N=100. In real life, then, we need to understand the behaviour of statistics calculated from more modest data sets!

7.1.3.2 Sampling distributions

With this in mind, let's abandon the idea that our studies will have sample sizes of 10,000, and consider a very modest experiment indeed. This time around we'll sample N=5 people and measure their IQ scores.

```
IQ.1 <- round( rnorm(n=5, mean=100, sd=15 ))
IQ.1
```

```
## [1] 76 93 96 90 98
```

The mean IQ in this sample turns out to be exactly 90.6. Now imagine that I decided to replicate the experiment. That is, I repeat the procedure as closely as possible: I randomly sample 5 new people and measure their IQ. Again, R allows me to simulate the results of this procedure:

```
IQ.2 <- round( rnorm(n=5, mean=100, sd=15 ))
IQ.2
```

```
## [1] 110  89 126  83  98
```

This time around, the mean IQ in my sample is 101.2. If I repeat the experiment 10 times I obtain the results shown in Table 7.1, and as you can see the sample mean varies from one replication to the next.

Table 7.1: Ten replications of the IQ experiment, each with a sample size of $N = 5$.

	Person1	Person2	Person3	Person4	Person5	SampleMean
Replication1	76	93	96	90	98	90.6
Replication2	110	89	126	83	98	101.2
Replication3	95	98	117	96	108	102.8
Replication4	104	78	109	96	110	99.4
Replication5	92	97	93	72	99	90.6
Replication6	96	116	108	107	92	103.8
Replication7	108	92	114	115	63	98.4
Replication8	98	93	101	117	87	99.2
Replication9	92	106	80	78	73	85.8
Replication10	101	114	85	76	117	98.6

Now suppose that I decided to keep going in this fashion, replicating this “five IQ scores” experiment over and over again. Every time I replicate the experiment I write down the sample mean. Over time, I’d be amassing a new data set, in which every experiment generates a single data point. The first 10 observations from my data set are the sample means listed in Table 7.1.

What if I continued like this for 10 replications and then drew a histogram? We would end up with a distribution of sample means. This distribution has a special name in statistics: it's called the ***sampling distribution of the mean***.[18]

Sampling distributions are another important theoretical idea in statistics, and they're crucial for understanding the behaviour of small samples. For instance, when I ran the very first "five IQ scores" experiment, the sample mean turned out to be 95. What the sampling distribution in Figure 10.5 tells us, though, is that the "five IQ scores" experiment is not very accurate. If I repeat the experiment, the sampling distribution tells me that I can expect to see a sample mean anywhere between 80 and 120.

7.1.3.3 The central limit theorem

In this section, we'll build on your understanding of the sampling distribution of the mean and look at how it changes as a function of sample size. Intuitively, you already know part of the answer. If you replicate a small experiment and recalculate the mean you'll get a very different answer; the sampling distribution is quite wide. If you replicate a large experiment and recalculate the sample mean you'll probably get almost the same answer you got last time; the sampling distribution is very narrow. This behavior is illustrated in Figure 7.7.

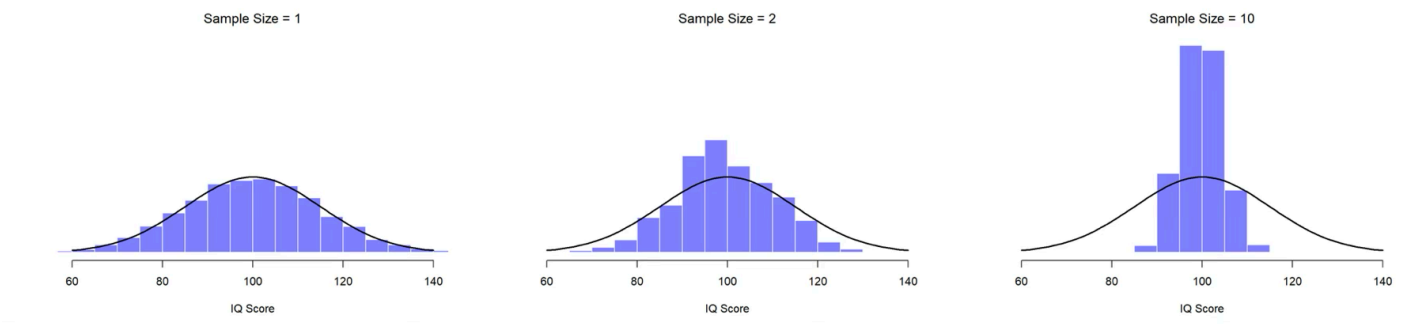

Figure 7.7: Sampling distributions at different sample sizes. When $N = 1$, each data set contains only a single observation and the mean of each sample is just one person's IQ score. The sampling distribution of the mean is then identical to the population distribution of IQ scores (plotted in black). When we raise the sample size to $N = 2$, the mean of any one sample tends to be closer to the population mean, and so the sampling distribution is a bit narrower than the population distribution. By the time we raise the sample size to $N = 10$, the distribution of sample means tend to be fairly tightly clustered around the true population mean.

We can quantify this effect by calculating the standard deviation of the sampling distribution, which is referred to as the ***standard error***. The standard error of a statistic is often denoted SE, and since we're usually interested in the standard error of the sample *mean*, we often use the

acronym SEM. As you can see just by looking at Figure 7.7, as the sample size N increases, the SEM decreases.

However, the central limit theorem is even stronger than this. All of the examples up to this point have been based on averaging IQ scores, and because IQ scores are roughly normally distributed, we've assumed that the population distribution is normal. But what if that isn't true? What happens to the sampling distribution of the mean when the population distribution is not normally distributed? Remarkably, no matter what shape the population distribution has, as N increases, the sampling distribution of the sample mean begins to follow a normal distribution.

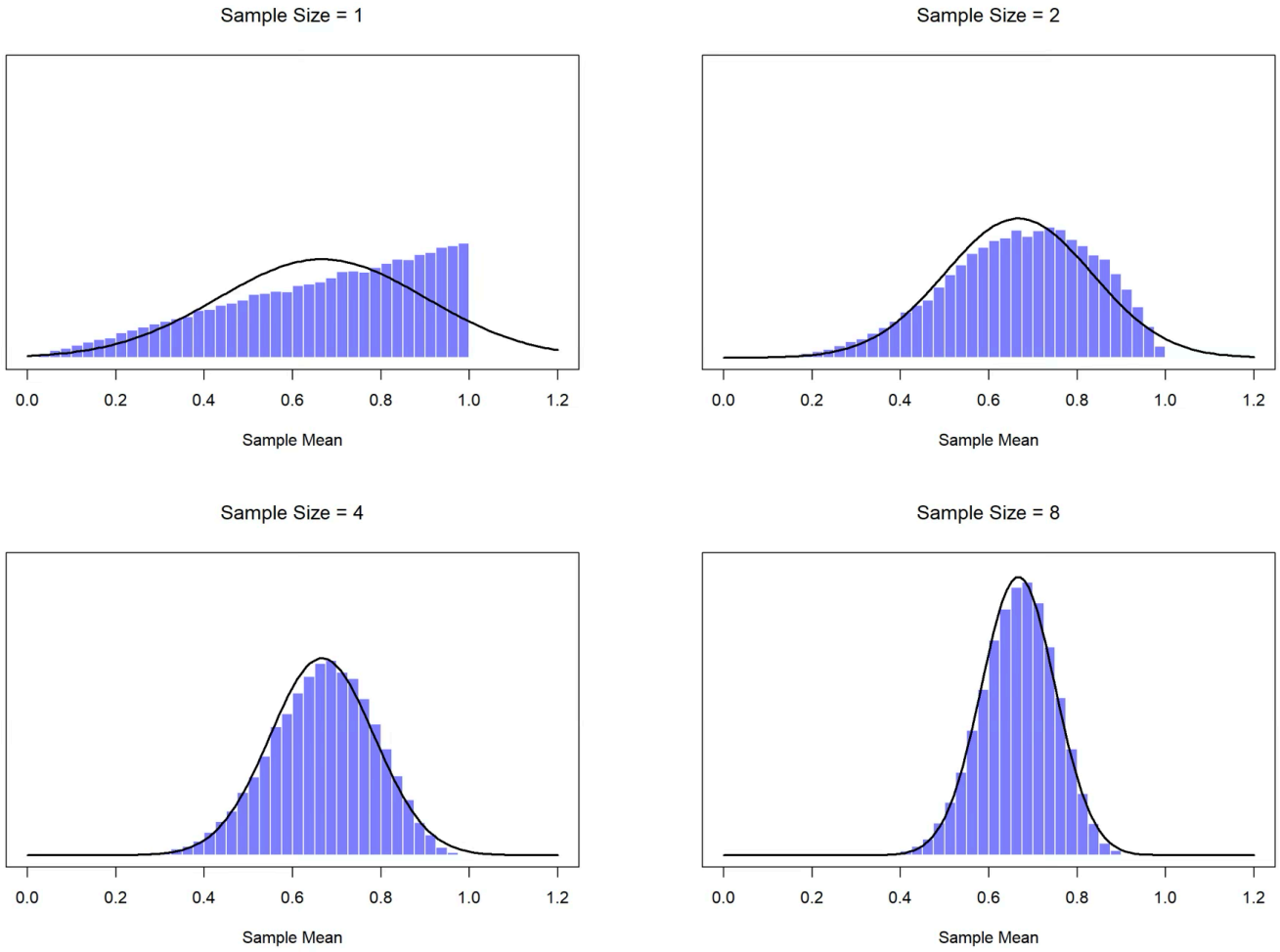

Figure 7.8: A non-normal population distribution and the sampling distributions of the sample mean at 4 values of N. When $N = 1$, the sampling distribution is just the population distribution (purple histogram), and the closest normal distribution (black curve) is a poor fit. As N increases to 2, 4, and 8, the sampling distribution becomes more symmetric and a better fit to its closest normal distribution.

Figure 7.8 shows how the sampling distribution of the mean approaches normality, even for a very non-normal population distribution. By $N = 8$ - which is not a particularly large sample! - the deviation from the normal distribution is barely visible. In other words, as long as your sample size isn't tiny, the sampling distribution of the mean will be approximately normal no matter what your population distribution looks like!

On the basis of these figures, it seems like we have evidence for all of the following claims about the sampling distribution of the mean:

- The mean of the sampling distribution is the same as the mean of the population.
- The standard deviation of the sampling distribution (i.e., the standard error) gets smaller as the sample size increases.
- The shape of the sampling distribution becomes normal as the sample size increases.

As it happens, not only are all of these statements true, there is a very famous theorem in statistics that proves all three of them, known as the ***central limit theorem***. Among other things, the central limit theorem tells us that if the population distribution has mean μ and standard deviation σ,[19] then the sampling distribution of the mean also has mean μ, and the standard error of the mean is $\mathrm{SEM} = \frac{\sigma}{\sqrt{N}}$. Because the denominator increases with the sample size, the SEM decreases.[20]

This result is useful for all sorts of things. It tells us why large experiments are more reliable than small ones, and because it gives us an explicit formula for the standard error it tells us *how much* more reliable a large experiment is. It tells us why the normal distribution is, well, *normal*. In real experiments, many of the things that we want to measure are actually averages of lots of different quantities (e.g., arguably, "general" intelligence as measured by IQ is an average of a large number of "specific" skills and abilities), and when that happens, the averaged quantity should follow a normal distribution. Because of this mathematical law, the normal distribution pops up over and over again in real data.

7.2 Estimating population parameters

In all the IQ examples in the previous sections, we actually knew the population parameters ahead of time. As every undergraduate gets taught in their very first lecture on the measurement of intelligence, IQ scores are defined to have mean 100 and standard deviation 15. However, this

is a bit of a lie. How do we know that IQ scores have a true population mean of 100? Well, we know this because the people who designed the tests have administered them to very large samples, and have then "rigged" the scoring rules so that their sample has mean 100.[21]

More often, we are trying to make inferences about a measurement where the population parameters are not known in advance. If we are designing a new experimental paradigm, there are no norms describing people's performance. Even if we are measuring something like IQ, it's not clear that the population we are sampling from will be a good fit to the sample that was used to norm the scoring. We're going to have to estimate the population parameters from a sample of data. How do we do this?

7.2.1 Estimating the population mean

Suppose we go to our small town and 100 of the locals are kind enough to sit through an IQ test. The average IQ score among these people turns out to be $\bar{X} = 98.5$. So what is the true mean IQ for the entire population of the town? Obviously, we don't know the answer to that question. It could be 97.2, but it could also be 103.5 . Our sampling isn't exhaustive so we cannot give a definitive answer. Nevertheless if I was forced at gunpoint to give a "best guess" I'd have to say 98.5 . That's the essence of statistical estimation: giving a best guess.

In this example, estimating the unknown poulation parameter is straightforward. I calculate the sample mean, and I use that as my estimate of the population mean. It's pretty simple, and in the next section I'll explain the statistical justification for this intuitive answer. However, for the moment what I want to do is make sure you recognise that the *sample statistic* and the *estimate of the population parameter* are conceptually different things. A sample statistic is a description of your data, whereas the estimate is a guess about the population. With that in mind, statisticians often different notation to refer to them. For instance, if true population mean is denoted μ then we would use $\hat{\mu}$ to refer to our estimate of the population mean. In contrast, the sample mean is denoted $\bar{X}$ or sometimes m. Table 7.2 lays these out for you.

Table 7.2: Symbols, names, and knowability for sample and population mean.

Symbol	What.is.it	Do.we.know.what.it.is
$\bar{X}$	Sample mean	Yes - calculated from the raw data
m	Sample mean (other notation)	Yes - calculated from the raw data
μ	Population mean	Almost never known for sure
$\hat{\mu}$	Estimate of the population mean	Yes - same as the sample mean

In simple random samples, the estimate of the population mean is identical to the sample mean; we say that the mean is an *unbiased estimator.*

7.2.2 Estimating the population standard deviation

So far, estimation seems pretty simple, and you might be wondering why I forced you to read through all that stuff about sampling theory. In the case of the mean, our estimate of the population parameter (i.e. $\hat{\mu}$) turned out to identical to the corresponding sample statistic (i.e. $\bar{X}$). However, that's not always true. To see this, let's have a think about how to construct an ***estimate of the population standard deviation***, which we'll denote $\hat{\sigma}$. What shall we use as our estimate in this case? Your first thought might be that we could do the same thing we did when estimating the mean, and just use the sample statistic as our estimate. That's almost the right thing to do, but not quite.

Here's why. Suppose I have a sample that contains a single observation. For this example, it helps to consider a sample where you have no intutions at all about what the true population values might be, so let's use something completely fictitious. Suppose the observation in question measures the *cromulence* of my shoes. It turns out that my shoes have a cromulence of 20. So here's my sample: 20.

This is a perfectly legitimate sample, even if it does have a sample size of $N = 1$. It has a sample mean of 20, and because every observation in this sample is equal to the sample mean (obviously!) it has a sample standard deviation of 0. As a description of the sample this seems quite right: the sample contains a single observation and therefore there is no variation observed within the sample. A sample standard deviation of $s = 0$ is the right answer here. But as an estimate of the population standard deviation, it feels completely insane, right? Admittedly, you and I don't know anything at all about what "cromulence" is, but we know something about data:

the only reason that we don't see any variability in the sample is that the sample is too small to display any variation! So, if you have a sample size of N = 1 , it feels like the right answer is just to say "no idea at all".

Notice that you *don't* have the same intuition when it comes to the sample mean and the population mean. If forced to make a best guess about the population mean, it doesn't feel completely insane to guess that the population mean is 20. Sure, you probably wouldn't feel very confident in that guess, because you have only the one observation to work with, but it's still the best guess you can make.

Let's extend this example a little. Suppose I now make a second observation. My data set now has $N = 2$ observations of the cromulence of shoes, and the complete sample now looks like this: [20, 22]

This time around, our sample is *just* large enough for us to be able to observe some variability: two observations is the bare minimum number needed for any variability to be observed! For our new data set, the sample mean is $\bar{X} = 21$, and the sample standard deviation is $s = 1$. What intuitions do we have about the population? Again, as far as the population mean goes, the best guess we can possibly make is the sample mean: if forced to guess, we'd probably guess that the population mean cromulence is 21. What about the standard deviation? This is a little more complicated. The sample standard deviation is only based on two observations, and if you're at all like me you probably have the intuition that, with only two observations, we haven't given the population "enough of a chance" to reveal its true variability to us. It's not just that we suspect that the estimate is *wrong*: after all, with only two observations we expect it to be wrong to some degree. The worry is that the error is *systematic*. Specifically, we suspect that the sample standard deviation is likely to be smaller than the population standard deviation.

This intuition feels right, but it would be nice to demonstrate this somehow. There are in fact mathematical proofs that confirm this intuition, but unless you have the right mathematical background they don't help very much. Instead, what I'll do is use R to simulate the results of some experiments. With that in mind, let's return to our IQ studies. Suppose the true population mean IQ is 100 and the standard deviation is 15. I can use the `rnorm()` function to generate the the results of an experiment in which I measure $N = 2$ IQ scores, and calculate the sample standard deviation. If I do this over and over again, and plot a histogram of these sample standard deviations, what I have is the *sampling distribution of the standard deviation*. I've plotted this distribution in Figure 10.11. Even though the true population standard deviation is 15, the average of the *sample* standard deviations is only 8.5. Notice that this is a very different

result to what we found in Figure 10.8 when we plotted the sampling distribution of the mean. If you look at that sampling distribution, what you see is that the population mean is 100, and the average of the sample means is also 100.

Now let's extend the simulation. Instead of restricting ourselves to the situation where we have a sample size of $N = 2$, let's repeat the exercise for sample sizes from 1 to 10. If we plot the average sample mean and average sample standard deviation as a function of sample size, you get the results shown in Figure 10.12. On the left hand side (panel a), I've plotted the average sample mean and on the right hand side (panel b), I've plotted the average standard deviation. The two plots are quite different: the mean is an unbiased estimator, meaning that *on average*, the sample mean is equal to the population mean. The plot on the right is quite different: on average, the sample standard deviation s is *smaller* than the population standard deviation σ. It is a ***biased estimator***. In other words, if we want to make a "best guess" $\hat{\sigma}$ about the value of the population standard deviation σ, we should make sure our guess is a little bit larger than the sample standard deviation s.

The fix to this systematic bias turns out to be very simple. Here's how it works. Before tackling the standard deviation, let's look at the variance. If you recall from our discussion of descriptive statistics, the sample variance is defined to be the average of the squared deviations from the sample mean.

$$s^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_N - \bar{X})^2}{N}$$

The sample variance s^2 is a biased estimator of the population variance σ^2. But as it turns out, we only need to make a tiny tweak to transform this into an unbiased estimator. All we have to do is divide by $N - 1$ rather than by N.

$$s^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_N - \bar{X})^2}{N - 1}$$

This is an unbiased estimator of the population variance σσ. Moreover, this finally answers the question we raised earlier. Why did R give us slightly different answers when we used the `var()` function? Because the `var()` function calculates σ^2, not s^2. A similar story applies for the standard deviation.

One final point: in practice, a lot of people tend to refer to $\hat{\sigma}$ (i.e., the formula where we divide by $N - 1$) as the *sample* standard deviation. Technically, this is incorrect: the *sample* standard deviation should be equal to s (i.e., the formula where we divide by N). These aren't the same

thing, either conceptually or numerically. One is a property of the sample, the other is an estimated characteristic of the population. However, in almost every real life application, what we actually care about is the estimate of the population parameter, and so people always report $\hat{\sigma}$ rather than s. This is the right number to report, of course, it's that people tend to get a little bit imprecise about terminology when they write it up, because "sample standard deviation" is shorter than "estimated population standard deviation". It's no big deal, and in practice I do the same thing everyone else does. Nevertheless, it's important to keep the two *concepts* separate: it's never a good idea to confuse "known properties of your sample" with "guesses about the population from which it came". The moment you start thinking that s and $\hat{\sigma}$ are the same thing, you start doing exactly that. Table 7.3 might help keep things clear.

Table 7.3: Symbols, names, and knowability for sample and population standard deviation and variance.

Symbol	What.is.it	Do.we.know.what.it.is
s	Sample standard deviation	Yes - calculated from the raw data
σ	Population standard deviation	Almost never known for sure
$\hat{\sigma}$	Estimate of the population standard deviation	Yes - but not the same as the sample standard deviation
s^2	Sample variance	Yes - calculated from the raw data
σ^2	Population variance	Almost never known for sure
$\hat{\sigma}^2$	Estimate of the population variance	Yes - but not the same as the sample variance

7.2.3 Confidence intervals

Up to this point in this chapter, I've outlined the basics of sampling theory which statisticians rely on to make guesses about population parameters on the basis of a sample of data. As this discussion illustrates, one of the reasons we need all this sampling theory is that every data set leaves us with a some of uncertainty, so our estimates are never going to be perfectly accurate. The thing that has been missing from this discussion is an attempt to *quantify* the amount of uncertainty that attaches to our estimate. It's not enough to be able guess that, say, the mean IQ of undergraduate psychology students is 115 (yes, I just made that number up). We also want to

be able to say something that expresses the degree of certainty that we have in our guess. For example, it would be nice to be able to say that there is a 95% chance that the true mean lies between 109 and 121. The name for this is a ***confidence interval*** for the mean.

Armed with an understanding of sampling distributions, constructing a confidence interval for the mean is actually pretty easy. Here's how it works. Suppose the true population mean is μ and the standard deviation is σ. I've just finished running my study that has N participants, and the mean IQ among those participants is $\bar{X}$. We know from our discussion of the central limit theorem (Section 7.1.3.3) that the sampling distribution of the mean is approximately normal. We also know from our discussion of standard deviations (Section 6.3.4) that there is a 95% chance that a normally-distributed quantity will fall within two standard deviations of the true mean. (More exactly, 95% of a normally-distributed quantity between 1.96 standard deviations of the mean.) Finally, recall that the standard deviation of the sampling distribution of the mean is referred to as the standard error of the mean, SEM. When we put all these pieces together, we can see that there is a 95% probability that the sample mean $\bar{X}$ that we have actually observed lies within 1.96 standard errors of the population mean. That is, there's a 95% probability that this inequality is true:

$$\mu - (1.96 \times \text{SEM}) \leq \bar{X} \leq \mu + (1.96 \times \text{SEM})$$

However, that's not answering the question that we're actually interested in. What we *want* is to have this work the other way around: we want to know what we should believe about the population parameters, given that we have observed a particular sample. A little bit of algebra shows that these are equivalent, so there is also a 95% probability that this inequality is true:

$$\bar{X} - (1.96 \times \text{SEM}) \leq \mu \leq \bar{X} + (1.96 \times \text{SEM})$$

That is, this range of values has a 95% probability of containing the true population mean μ. This range is a ***95% confidence interval***, denoted CI_{25}. As long as N is large enough for us to believe that our sampling distribution of the mean is normal, we can use this approach to define our 95% confidence interval.

There's nothing particular special about the numbers 1.96, other than that they're the quantiles of the normal distribution that bound 95% of the data. If we wanted a 70% confidence interval, we could find the 15th and 85th quantiles and use those instead, computing CI_{70}.

7.3 Hypothesis testing

In its most abstract form, hypothesis testing really a very simple idea: the researcher has some theory about the world, and wants to determine whether or not the data actually support that theory. However, the details are messy, and most people find the theory of hypothesis testing to be the most frustrating part of statistics. The structure of the chapter is as follows. Firstly, I'll describe how hypothesis testing works, in a fair amount of detail, using a simple running example to show you how a hypothesis test is "built". I'll try to avoid being too dogmatic while doing so, and focus instead on the underlying logic of the testing procedure.[22] Afterwards, I'll spend a bit of time talking about the various dogmas, rules and heresies that surround the theory of hypothesis testing.

7.3.1 A menagerie of hypotheses

To pick a rather silly example, let's consider a study seeking evidence of extrasensory perception (ESP). My first study is a simple one, in which I seek to test whether clairvoyance exists. Each participant sits down at a table, and is shown a card by an experimenter. The card is black on one side and white on the other. The experimenter takes the card away, and places it on a table in an adjacent room. The card is placed black side up or white side up completely at random, with the randomisation occurring only after the experimenter has left the room with the participant. A second experimenter comes in and asks the participant which side of the card is now facing upwards. It's purely a one-shot experiment. Each person sees only one card, and gives only one answer; and at no stage is the participant actually in contact with someone who knows the right answer. My data set, therefore, is very simple. I have asked the question of N people, and some number X of these people have given the correct response. To make things concrete, let's suppose that I have tested $N = 100$ people, and $X = 62$ of these got the answer right… a surprisingly large number, sure, but is it large enough for me to feel safe in claiming I've found evidence for ESP? This is the situation where hypothesis testing comes in useful. However, before we talk about how to *test* hypotheses, we need to be clear about what we mean by hypotheses.

The first distinction that you need to keep clear in your mind is between research hypotheses and statistical hypotheses. In my ESP study, my overall scientific goal is to demonstrate that clairvoyance exists. In this situation, I have a clear research goal: I am hoping to discover evidence for ESP. In other situations I might actually be a lot more neutral than that, so I might

say that my research goal is to determine whether or not clairvoyance exists. Regardless of how I want to portray myself, the basic point that I'm trying to convey here is that a research hypothesis involves making a substantive, testable scientific claim… if you are a psychologist, then your research hypotheses are fundamentally about psychological constructs. Any of the following would count as research hypotheses:

- *Listening to music reduces your ability to pay attention to other things.* This is a claim about the causal relationship between two psychologically meaningful concepts (listening to music and paying attention to things), so it's a perfectly reasonable research hypothesis.
- *Intelligence is related to personality.* Like the last one, this is a relational claim about two psychological constructs (intelligence and personality), but the claim is weaker: correlational not causal.
- *Intelligence is speed of information processing.* This hypothesis has a quite different character: it's not actually a relational claim at all. It's an ontological claim about the fundamental character of intelligence (and I'm pretty sure it's wrong). It's worth expanding on this one actually: It's usually easier to think about how to construct experiments to test research hypotheses of the form "does X affect Y?" than it is to address claims like "what is X?" And in practice, what usually happens is that you find ways of testing relational claims that follow from your ontological ones. For instance, if I believe that intelligence is* speed of information processing in the brain, my experiments will often involve looking for relationships between measures of intelligence and measures of speed. As a consequence, most everyday research questions do tend to be relational in nature, but they're almost always motivated by deeper ontological questions about the state of nature.

Notice that in practice, my research hypotheses could overlap a lot. My ultimate goal in the ESP experiment might be to test an ontological claim like "ESP exists", but I might operationally restrict myself to a narrower hypothesis like "Some people can "see" objects in a clairvoyant fashion". That said, there are some things that really don't count as proper research hypotheses in any meaningful sense:

- *Love is a battlefield*. This is too vague to be testable. While it's okay for a research hypothesis to have a degree of vagueness to it, it has to be possible to operationalise your theoretical ideas. Maybe I'm just not creative enough to see it, but I can't see how this can be converted into any concrete research design. If that's true, then this isn't a scientific research hypothesis, it's a pop song. That doesn't mean it's not interesting – a lot of deep

questions that humans have fall into this category. Maybe one day science will be able to construct testable theories of love, or to test to see if God exists, and so on; but right now we can't, and I wouldn't bet on ever seeing a satisfying scientific approach to either.

- *The first rule of tautology club is the first rule of tautology club*. This is not a substantive claim of any kind. It's true by definition. No conceivable state of nature could possibly be inconsistent with this claim. As such, we say that this is an unfalsifiable hypothesis, and as such it is outside the domain of science. Whatever else you do in science, your claims must have the possibility of being wrong.

- *More people in my experiment will say "yes" than "no"*. This one fails as a research hypothesis because it's a claim about the data set, not about the psychology (unless of course your actual research question is whether people have some kind of "yes" bias!). As we'll see shortly, this hypothesis is starting to sound more like a statistical hypothesis than a research hypothesis.

As you can see, research hypotheses can be somewhat messy at times; and ultimately they are *scientific* claims. ***Statistical hypotheses*** are neither of these two things. Statistical hypotheses must be mathematically precise, and they must correspond to specific claims about the characteristics of the data generating mechanism (i.e., the "population"). Even so, the intent is that statistical hypotheses bear a clear relationship to the substantive research hypotheses that you care about!

For instance, in my ESP study my research hypothesis is that some people are able to see through walls or whatever. What I want to do is to "map" this onto a statement about how the data were generated. What might that statement be? The sample statistic I've computed is $P(\text{correct})$, and I want to make inferences about the true-but-unknown probability with which people in the population are able to answer the question correctly. Let's use the Greek letter θ (theta) to refer to this probability. Here are four different statistical hypotheses:

- If ESP doesn't exist and if my experiment is well designed, then my participants are just guessing. So I should expect them to get it right half of the time and so my statistical hypothesis is that the true probability of choosing correctly is $\theta = 0.5$.

- Alternatively, suppose ESP does exist and participants can see the card. If that's true, people will perform better than chance. The statistical hypotheis would be that $\theta > 0.5$.

- A third possibility is that ESP does exist, but the colors are all reversed and people don't realise it (okay, that's wacky, but you never know…). If that's how it works then you'd expect people's performance to be *below* chance. This would correspond to a statistical hypothesis

that $\theta < 0.5$.

- Finally, suppose ESP exists, but I have no idea whether people are seeing the right colour or the wrong one. In that case, the only claim I could make about the data would be that the probability of making the correct answer is not equal to 0.5. This corresponds to the statistical hypothesis that $\theta \neq 0.5$.

While some of these seem more plausible than others, all are legitimate examples of a statistical hypothesis. They are statements about a population parameter and are meaningfully related to my experiment.

What this discussion makes clear, I hope, is that when preparing to construct a statistical hypothesis test, the researcher actually has two quite distinct hypotheses to consider. They start from a research hypothesis (a claim about psychology), and this corresponds to a statistical hypothesis (a claim about some population parameter). The key thing to recognize is this: *a statistical hypothesis test is a test of the statistical hypothesis, not the research hypothesis*. If your study is badly designed, then the link between your research hypothesis and your statistical hypothesis is broken. To give a silly example, suppose that my ESP study was conducted in a situation where the participant can actually see the card reflected in a window; if that happens, I would be able to find very strong evidence that $\theta \neq 0.5$ but this would tell us nothing about whether "ESP exists".

7.3.2 Null hypotheses and alternative hypotheses

So far, so good. I have a research hypothesis that corresponds to what I want to believe about the world, and I can map it onto a statistical hypothesis that corresponds to what I want to believe about how the data were generated. It's at this point that things get somewhat counterintuitive for a lot of people. Because what I'm about to do is invent a new statistical hypothesis (the "null" hypothesis, H_0) that corresponds to the exact opposite of what I hope is true, and then focus exclusively on that, almost to the neglect of the thing I'm actually interested in (which is now called the "alternative" hypothesis, H_1). In our ESP example, the null hypothesis is that $\theta = 0.5$, since that's what we'd expect if ESP *didn't* exist. My hope as a researcher, of course, is that ESP is totally real, and so the *alternative* to this null hypothesis is $\theta \neq 0.5$. In essence, what we're doing here is dividing up the possible values of θ into two groups: those values that I really hope aren't true (the null), and those values that I'd be happy with if they turn

out to be right (the alternative). Having done so, the important thing to recognize is that the goal of a hypothesis test is *not* to show that the alternative hypothesis is (probably) true; the goal is to show that the null hypothesis is (probably) false. Most people find this pretty weird.

The best way to think about it, in my experience, is to imagine that a hypothesis test is a criminal trial[23]: *the trial of the null hypothesis*. The null hypothesis is the defendant, the researcher is the prosecutor, and the statistical test itself is the judge. Just like a criminal trial, there is a presumption of innocence: the null hypothesis is *deemed* to be true unless you, the researcher, can prove beyond a reasonable doubt that it is false. You are free to design your experiment however you like (within reason, obviously!), and your goal when doing so is to maximise the chance that the data will yield a conviction… for the crime of being false. The catch is that the statistical test sets the rules of the trial, and those rules are designed to protect the null hypothesis – specifically to ensure that if the null hypothesis is actually true, the chances of a false conviction are guaranteed to be low. This is pretty important: after all, the null hypothesis doesn't get a lawyer. And given that the researcher is trying desperately to prove it to be false, *someone* has to protect it.

Finally, just like in a criminal trial, the final decision is a yes/no answer. The defendant is either convicted or freed; the null hypothesis is either rejected or retained.[24]

7.3.3 Two types of errors

Before going into details about how a statistical test is constructed, it's useful to understand the philosophy behind it. I hinted at it when pointing out the similarity between a null hypothesis test and a criminal trial, but I should now be explicit. Ideally, we would like to construct our test so that we never make any errors. Unfortunately, since the world is messy, this is never possible. Sometimes you're just really unlucky: for instance, suppose you flip a coin 10 times in a row and it comes up heads all 10 times. That feels like very strong evidence that the coin is biased (and it is!), but of course there's a 1 in 1024 chance that this would happen even if the coin was totally fair. In other words, in real life we always have to accept that there's a chance that we did the wrong thing. As a consequence, the goal behind statistical hypothesis testing is not to eliminate errors, but to minimise them.

At this point, we need to be a bit more precise about what we mean by "errors". Because the null hypothesis can be either true or false in the real world, and our test is a yes/no decision about the null hypothesis, there are exactly four possible outcomes:

1. The null hypothesis is **true** in the population, and our statistical test **retains** H_0 (correct decision).
2. The null hypothesis is **true** in the population, but our statistical test **rejects** H_0 (false positive, type I error).
3. The null hypothesis is **false** in the population, but our statistical test **retains** H_0 (false negative, type II error).
4. The null hypothesis is **false** in the population, and our statistical test **rejects** H_0 (correct decision).

Two of these are correct, but there are two different types of errors that we can make. If we reject a null hypothesis that is actually true, then we have made a type I error. On the other hand, if we retain the null hypothesis when it is in fact false, then we have made a type II error.

Remember how I said that statistical testing was kind of like a criminal trial? Well, I meant it. A criminal trial requires that you establish "beyond a reasonable doubt" that the defendant did it. All of the evidentiary rules are (in theory, at least) designed to ensure that there's (almost) no chance of wrongfully convicting an innocent defendant. The trial is designed to protect the rights of a defendant: as the English jurist William Blackstone famously said, it is "better that ten guilty persons escape than that one innocent suffer." In other words, a criminal trial doesn't treat the two types of error in the same way~… punishing the innocent is deemed to be much worse than letting the guilty go free. A statistical test is pretty much the same: the single most important design principle of the test is to control the probability of a type I error (false positives), to keep it below some fixed level. This probability, which is denoted α, is called the *significance level* of the test. I'll say it again, because it is so central to the whole set-up… a hypothesis test is said to have significance level α if the type I error (false positive) rate is no larger than α.

So, what about the type II error rate? Well, we'd also like to keep those under control too, and we denote this probability by β. However, it's much more common to refer to the ***power*** of the test, which is the probability with which we reject a null hypothesis when it really is false, which is $1 - \beta$. To help keep this straight, here's that list of outcomes again, but with the relevant probabilities included:

1. The null hypothesis is **true** in the population, and our statistical test **retains** H_0 (correct decision, $P = 1 - \alpha$).
2. The null hypothesis is **true** in the population, but our statistical test **rejects** H_0 (false positive, type I error, $P = \alpha$).

3. The null hypothesis is **false** in the population, but our statistical test **retains** H_0 (false negative, type II error, $P = \beta$).
4. The null hypothesis is **false** in the population, and our statistical test **rejects** H_0 (correct decision, $P = 1 - \beta$).

A "powerful" hypothesis test is one that has a small value of β, while still keeping α fixed at some (small) desired level, usually 0.5, 0.1, or 0.001. There's a critical asymmetry here: the tests are designed to *ensure* that the α level is kept small, but there's no corresponding guarantee regarding β. We'd certainly *like* the false negative error rate to be small, and we try to design tests that keep it small, but this is very much secondary to the overwhelming need to control the false positive rate. As Blackstone might have said if he were a statistician, it is "better to retain 10 false null hypotheses than to reject a single true one". To be honest, I don't know that I agree that this makes sense in every situation, but that's neither here nor there. It's how the tests are built.

7.3.4 Test statistics and sampling distributions

At this point we need to start talking specifics about how a hypothesis test is constructed. To that end, let's return to the ESP example. Let's ignore the actual data that we obtained, for the moment, and think about the structure of the experiment. Regardless of what the actual numbers are, the form of the data is that X out of N people correctly identified the colour of the hidden card. Moreover, let's suppose for the moment that the null hypothesis really is true: ESP doesn't exist, and the true probability that anyone picks the correct colour is exactly $\theta = 0.5$. What would we expect the data to look like? Well, obviously, we'd expect the proportion of people who make the correct response to be pretty close to 50%. Or, to phrase this in more mathematical terms, we'd say that $\frac{X}{N}$ is approximately 0.5. Of course, as we saw when discussing sampling distributions, this fraction probably wouldn't be exactly 0.5. On the other hand, if $X = 99$ of our participants got the question right, then we'd feel pretty confident that the null hypothesis is wrong. Similarly, if only $X = 3$ people got the answer right, we'd be similarly confident that the null was wrong.

Let's be a little more technical about this: we have a quantity X that we can calculate by looking at our data. After looking at the value of X, we make a decision about whether to believe that the null hypothesis is correct, or to reject the null hypothesis in favour of the alternative. The name for this thing that we calculate to guide our choices is a ***test statistic***.

Having chosen a test statistic, the next step is to state precisely which values of the test statistic would cause is to reject the null hypothesis, and which values would cause us to keep it. In order to do so, we need to determine what the sampling distribution of the test statistic would be if the null hypothesis were actually true (we talked about sampling distributions earlier in Section 7.1.3.2). Why do we need this? Because this distribution tells us exactly what values of X our null hypothesis would lead us to expect. And therefore, we can use this distribution as a tool for assessing how closely the null hypothesis agrees with our data.

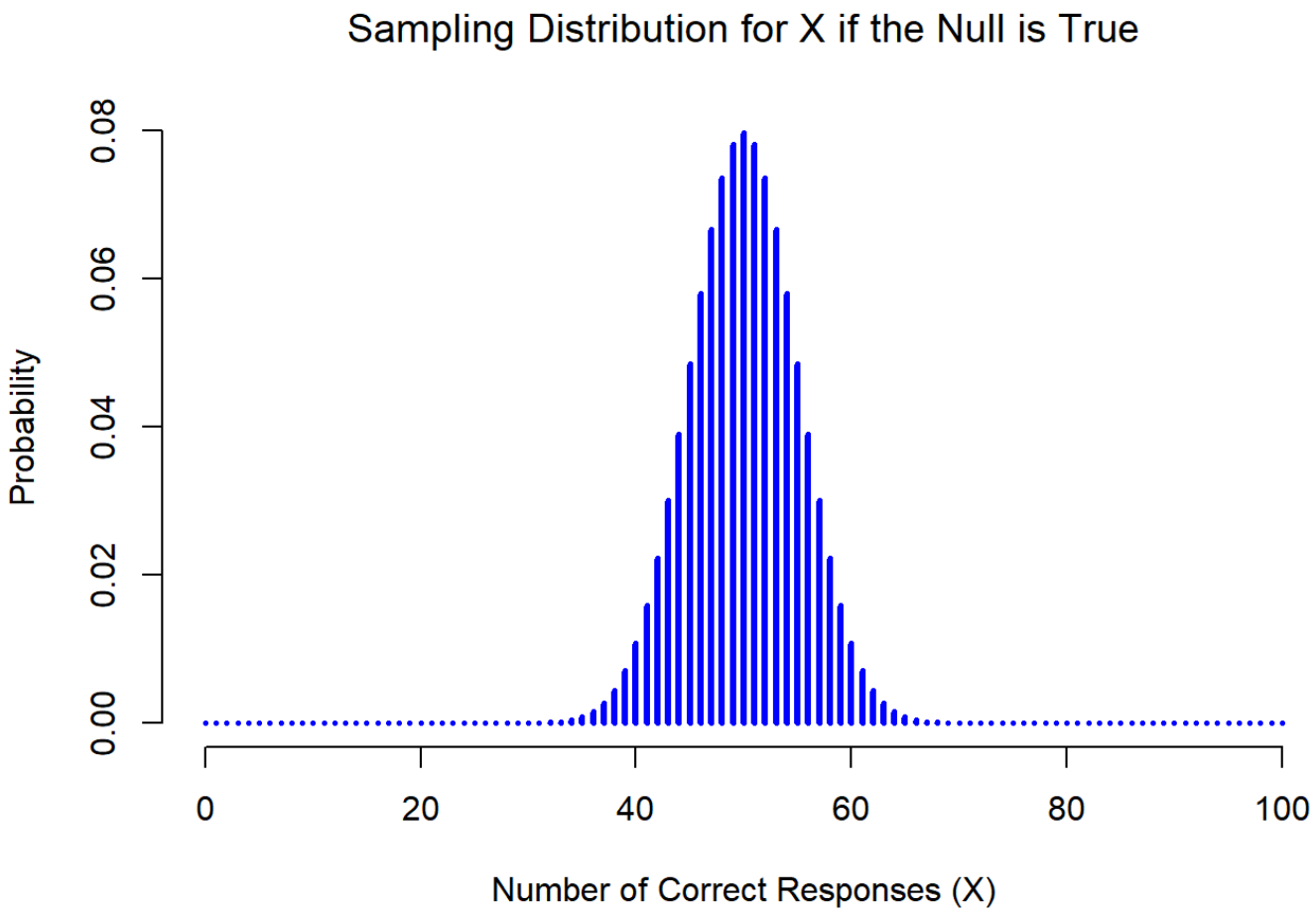

Figure 7.9: The sampling distribution for our test statistic X when the null hypothesis is true. For our ESP scenario, this is a binomial distribution. Not surprisingly, since the null hypothesis says that the probability of a correct response is $\theta = 0.5$, the sampling distribution says that the most likely value is 50 (out of 100) correct responses. Most of the probability mass lies between 40 and 60.

How do we actually determine the sampling distribution of the test statistic? For a lot of hypothesis tests this step is actually quite complicated. However, sometimes it's very easy. And, fortunately for us, our ESP example provides us with one of the easiest cases. Our population parameter θ is just the overall probability that people respond correctly when asked the

question, and our test statistic X is the count of the number of people who did so, out of a sample size of N. As you may remember from your statistics class, this situation is exactly what the binomial distribution describes, and the null hypothesis predicts that the sampled value of X will be binomially distributed. This sampling distribution is plotted in Figure 7.9. No surprises really: the null hypothesis says that $X = 50$ is the most likely outcome, and it says that we're almost certain to see somewhere between 40 and 60 correct responses.

7.3.5 Making decisions

We're almost there! We've constructed a test statistic X, and we're pretty confident that, if X is close to $0.5 \times N$, then we should retain the null, and if not, we should reject it. The question that remains is this: exactly which values of the test statistic should we associate with the null hypothesis, and which exactly values go with the alternative hypothesis? In my ESP study, for example, I've observed a value of $X = 62$. What decision should I make? Should I choose to believe the null hypothesis, or the alternative hypothesis?

To answer this question, we need to introduce the concept of a ***critical region*** for the test statistic X. The critical region of the test corresponds to those values of X that would lead us to reject null hypothesis (which is why the critical region is also sometimes called the rejection region). How do we find this critical region? Well, let's consider what we know:

- We know the sampling distribution of X if the null hypothesis is true (Figure 7.9).
- We know that X should be very big or very small for us to be confident in rejecting the null hypothesis.
- In order to hold $\alpha = 0.05$, the critical region must encompass 5% of the sampling distribution.

As it turns out, those three things uniquely solve the problem: our critical region consists of the most *extreme values*, known as the ***tails*** of the distribution (Figure 7.10). For this distribution, if we want $\alpha = 0.5$, then our critical regions correspond to $X \leq 40$ and $X \geq 60$. That is, if the number of people saying "true" is between 41 and 59, then we should retain the null hypothesis. If the number is between 0 to 40 or between 60 to 100, then we should reject the null hypothesis.

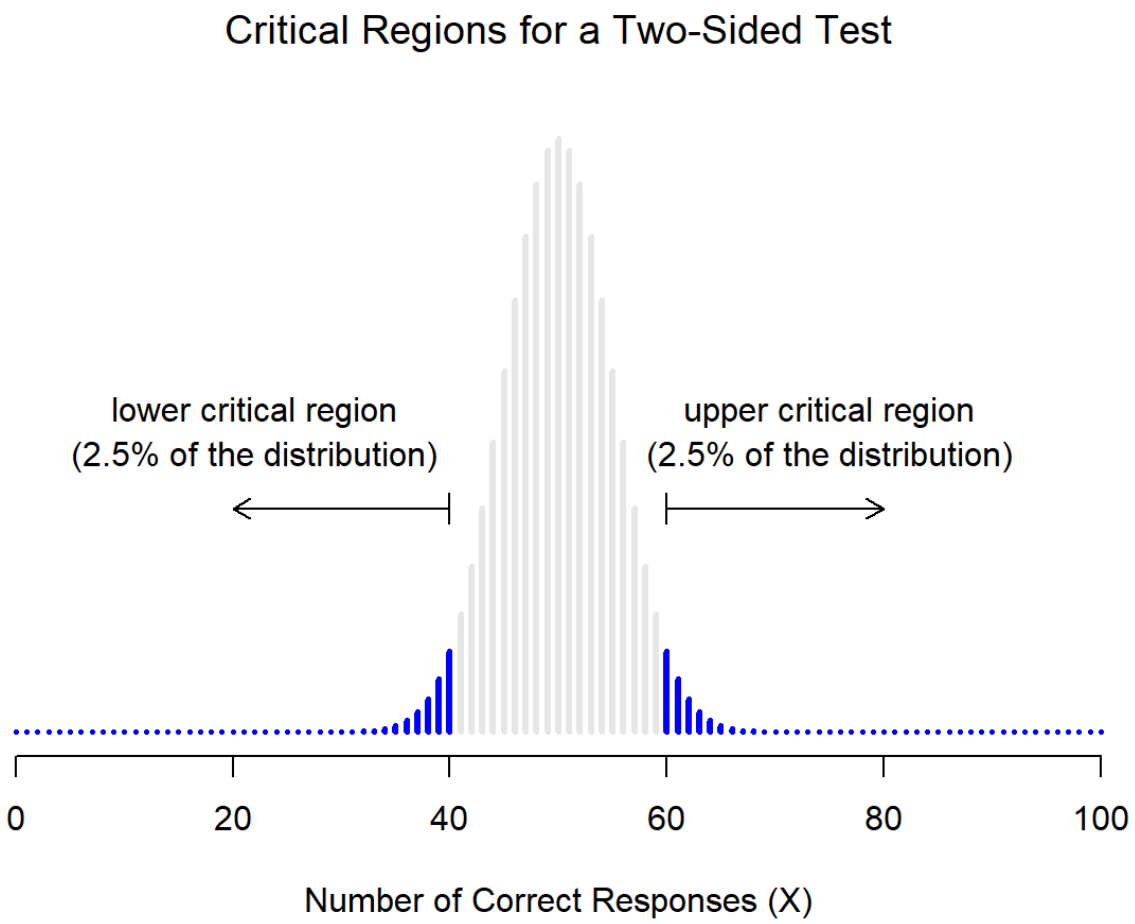

Figure 7.10: The critical region associated with a hypothesis test for the ESP study, with significance level $\alpha = 0.05$. The plot itself shows the sampling distribution of X under the null hypothesis: the grey bars correspond to those values of X for which we would retain the null hypothesis. The purple bars show the critical region: those values of X for which we would reject the null. Because the alternative hypothesis is two sided (i.e., allows both $\theta < 0.5$ and $\theta > .5$), the critical region has a portion in each tail of the distribution. So that the overall α level is 0.5, we need each portion of the critical region to encompasses 2.5% of the sampling distribution.

At this point, our hypothesis test is essentially complete: we (1) choose an α level (e.g., $\alpha = .05$, (2) come up with some test statistic (e.g., X) that does a good job (in some meaningful sense) of comparing H_0 to H_1, (3) figure out the sampling distribution of the test statistic on the assumption that the null hypothesis is true (in this case, binomial) and then (4) identify the critical regions that produces an appropriate α level (0-40 and 60-100). All that we have to do now is calculate the value of the test statistic for the real data ($X = 62$) and then compare it to the critical regions to make our decision. Since 62 falls into the critical region from 60-100, we reject the null hypothesis. Or, to phrase it slightly differently, we say that the test has produced a ***significant*** result.

7.3.5.1 About statistical "significance"

A very brief digression is in order at this point, regarding the word "significant". The concept of statistical significance is actually a very simple one, but has a very unfortunate name. If the data allow us to reject the null hypothesis, we say that "the result is *statistically significant*", which is often shortened to "the result is significant". This terminology is rather old, and dates back to a time when "significant" just meant something like "indicated", rather than its modern meaning, which is much closer to "important". As a result, a lot of modern readers get very confused when they start learning statistics, because they think that a "significant result" must be an important one. It doesn't mean that at all. All that "statistically significant" means is that the data allowed us to reject a null hypothesis. Whether or not the result is actually important in the real world is a very different question, and depends on all sorts of other things.

7.3.6 The p value of a test

In one sense, our hypothesis test is complete; we've constructed a test statistic, figured out its sampling distribution if the null hypothesis is true, and then constructed the critical region for the test. Nevertheless, I've actually omitted the most important number of all: the$p$ value. It is to this topic that we now turn.

There are two somewhat different ways of interpreting a p value, one proposed by Sir Ronald Fisher and the other by Jerzy Neyman. Both versions are legitimate, though they reflect very different ways of thinking about hypothesis tests. Most introductory textbooks tend to give Fisher's version only, but I think that's a bit of a shame. To my mind, Neyman's version is cleaner, and actually better reflects the logic of the null hypothesis test. You might disagree though, so I've included both. I'll start with Neyman's version…

One problem with the hypothesis testing procedure that I've described is that it makes no distinction at all between a result this "barely significant" and those that are "highly significant". For instance, in my ESP study the data I obtained only just fell inside the critical region - so I did get a significant effect, but was a pretty near thing. In contrast, suppose that I'd run a study in which $X = 97$ out of my $N = 100$ participants got the answer right. This would obviously be significant too, but my a much larger margin; there's really no ambiguity about this at all. The procedure that I described makes no distinction between the two. If I adopt the standard convention of allowing $\alpha = .05$ as my acceptable Type I error rate, then both of these are significant results.

This is where reporting an exact p value can come in handy. It turns out that our ESP data has $p = .021$. That is, if we set $\alpha = .022$ or higher, these data would lead us to reject the null hypothesis, but if we set $\alpha = 0.020$ or lower, we would have to retain it. That is, a test's p is *the false positive (Type I error) rate that you must be willing to tolerate if you want to reject the null hypothesis*.

If it turns out that p describes a false positive rate that you find intolerable, then you must retain the null. If you're comfortable with a false positive rate equal to p, then it's okay to reject the null hypothesis on the basis of these data. In effect, p is a summary of all the possible hypothesis tests that you could have run, taken across all possible α values.

The second definition of the p value comes from Sir Ronald Fisher, and it's actually this one that you tend to see in most introductory statistics textbooks. Notice how, when I constructed the critical regions, they corresponded to the tails of the sampling distribution? That's not a coincidence; almost all good statistical tests[25] have this characteristic. This is because the critical region should correspond to those values of the test statistic that are least likely to be observed if the null hypothesis is true. Then, we can define the p as the probability that we would observe a test statistic that is at least as extreme as the one that we did get. If the data are extremely implausible according to the null hypothesis, then the null hypothesis is probably wrong.

7.3.6.1 A common mistake

Unfortunately, there is a third explanation that people sometimes give, especially when they're first learning statistics, and it is **absolutely and completely wrong**. This mistaken approach is to define p as the probability that the null hypothesis is true. It's an intuitively appealing way to think, but even under Bayesian statistics (a set of tools for assigning probabilities to hypotheses), this interpretation is incompatible with the underlying calculations. Never do it.

13. Navarro is Australian, and her original text uses British/Commonwealth spellings and idioms. I (Noyce) am American, and use US ones. Apologies for the occasionally-jarring combination; editing for a consistent set of usages is on the list for a future edition of these notes.↩

14. The quote comes from Wittgenstein's (1922) text, *Tractatus Logico-Philosphicus*.↩

15. The proper mathematical definition of randomness is extraordinarily technical, and way beyond the scope of this book. We'll be non-technical here and say that a process has an element of randomness to it whenever it is possible to repeat the process and get different answers each time.↩

16. Nothing in life is that simple: there's not an obvious division of people into binary categories like "schizophrenic" and "not schizophrenic". But this isn't a clinical psychology course, so please forgive me a few simplifications here and there.↩

17. Technically, the law of large numbers pertains to any sample statistic that can be described as an average of independent quantities. That's certainly true for the sample mean. However, it's also possible to write many other sample statistics as averages of one form or another. The variance of a sample, for instance, can be rewritten as a kind of average and so is subject to the law of large numbers. The minimum value of a sample, however, cannot be written as an average of anything and is therefore not governed by the law of large numbers.↩

18. One thing to keep in mind when thinking about sampling distributions is that *any* sample statistic you might care to calculate has a sampling distribution. For example, suppose that each time I replicated the "five IQ scores" experiment I wrote down the largest IQ score in the experiment. This would give me a very different sampling distribution, the *sampling distribution of the maximum.* Not surprisingly, if you pick 5 people at random and then find the person with the highest IQ score, they're going to have an above average IQ. Most of the time you'll end up with someone whose IQ is measured in the 100 to 140 range.↩

19. Remember, population parameters are indicated with Greek letters.↩

20. The mean is not the only statistic that obeys the central limit theorem; there's a whole class of them. They're called U-statistics.↩

21. That's not a bad thing of course: it's an important part of designing a psychological measurement. However, it's important to keep in mind that this theoretical mean of 100 only attaches to the population that the test designers used to design the tests. Good test designers will actually go to some lengths to provide "test norms" that can apply to lots of different populations (e.g., different age groups, nationalities etc)..↩

22. A technical note. The description below differs subtly from the standard description given in a lot of introductory texts. The orthodox theory of null hypothesis testing emerged from the work of Sir Ronald Fisher and Jerzy Neyman in the early 20th century; but Fisher and

Neyman actually had very different views about how it should work. The standard treatment of hypothesis testing that most texts use is a hybrid of the two approaches. The treatment here is a little more Neyman-style than the orthodox view, especially as regards the meaning of the p value.↩

23. This analogy only works if you're from an adversarial legal system like UK/US/Australia. As I understand these things, the French inquisitorial system is quite different.↩

24. An aside regarding the language you use to talk about hypothesis testing. Firstly, one thing you really want to avoid is the word "prove": a statistical test really doesn't *prove* that a hypothesis is true or false. Proof implies certainty, and as the saying goes, statistics means never having to say you're certain. On that point almost everyone would agree. However, beyond that there's a fair amount of confusion. Some people argue that you're only allowed to make statements like "rejected the null", "failed to reject the null", or possibly "retained the null". According to this line of thinking, you can't say things like "accept the alternative" or "accept the null". Personally I think this is too strong: in my opinion, this conflates null hypothesis testing with Karl Popper's falsificationist view of the scientific process. While there are similarities between falsificationism and null hypothesis testing, they aren't equivalent. However, while I personally think it's fine to talk about accepting a hypothesis (on the proviso that "acceptance" doesn't actually mean that it's necessarily true, especially in the case of the null hypothesis), many people will disagree. And more to the point, you should be aware that this particular weirdness exists, so that you're not caught unawares by it when writing up your own results.↩

25. "Good" in the sense of minimizing the false negative rate, β.↩