Chapter 8

Inferential statistics

This chapter takes statistics one step further, from descriptive – which we dealt with in chapter 6 – to inferential statistics. Inferential statistics are necessary whenever we want to *infer* something from a sample, that is, declare the findings from that sample applicable to a larger set of units, a population. Whenever we use inferential statistics, we do this because there is uncertainty. We have studied a subset of the total universe we are interested in. We want to conclude something about that total universe, but as we have sampled only a fraction of it, we cannot be sure that we know exactly what is going on in that larger universe. So our statements about the larger universe are always uncertain.

How to deal with that uncertainty is what inferential statistics is about. In the following, we will introduce the basics of statistical testing. In doing so, we discuss hypotheses, testing, distributions and the power of tests, dealing with fairly abstract theoretical as well as practical issues. We end with a short paragraph that discusses some basic issues in estimation.

As we will hopefully be able to show, the rationale behind statistical testing is not difficult or ungraspable for the mathematically insecure. In fact, statistical reasoning is quite logical and of the kind that we apply in day-to-day situations. While we will try to keep the formulas and calculations to a minimum, we do need to show some formulas in order to explain the reasoning that is formalized in the procedures and formulas, and to show that they indeed function as we would expect them to. We encourage the reader to take these parts in his or her stride, and if necessary to jump over the formulas if they distract from the main thrust. While it is not necessary to understand or work with these formulas when using statistical testing, it is important that you do understand what testing is about, what it means when you say something is significant, what a confidence interval says, and what the power of a test is and how that relates to sample size. We will highlight these key take-aways in the questions at the end of the chapter.

8.1 Statistical testing

Thus far, whenever we mentioned inferential statistics we did so pretty vaguely and in an off-hand manner. It is now time to become more precise. In the following sections we will explain – again starting with logics and employing ordinary wording as much as possible – how inferences are made in statistics, and how uncertainty is quantified using statistical techniques.

Let us assume that we have a representative sample. Given that the sample is representative, we may after calculating various properties of the sample, such as the mean, median, variance, association between two variables and the like, declare those sample results applicable to the population. We could find that civil law judges have higher IQs than criminal law judges, or that female victims sue for larger amounts of compensation for pains and damages than male victims. However, even though we just stated that such inferences are defensible, we obviously are not *certain* about such generalizations.

Could our finding not simply be attributable to chance, a random result, coincidence? Because our sample was drawn randomly, we may consider it to reflect the population, but even so, uncertainty does remain. After all, we have not observed each and every population member, and there is and remains always a chance that our findings, such as differences between civil and criminal law judges, between female and male victims, hold for the sample but that patterns are quite different in the population...

In order to deal with this uncertainty, we generally use statistical tests. There are very many different kinds of tests. The basic rationale of these tests is however always the same. And this rationale is not difficult, as it follows the kind of reasoning each of us applies in everyday life. We will give examples of that below.

8.1.1 Day-to-day statistical reasoning

Suppose that you go out to dinner at a nice new little restaurant around the corner from your house. The meal tastes wonderful, you're having stuffed calamari. That night, however, you wake up feeling all queasy, and after a while, you have to throw up. The food clearly didn't agree with you. Now, there's stomach flu going around, so you simply assume that that must have been it, and leave it there. A few months later, you invite a friend for a meal at the same place. Again, the food is wonderful, you're having a salad and beef stew. That night, to your surprise the same thing happens as last time, and when you meet your friend the next day, you find out that she has also been hanging over the loo getting rid of the restaurant food. Now you draw your conclusion: this is a bad restaurant. The first and the second time you went out for dinner at the restaurant under the assumption that they would be serving good food. Now you cross out that assumption, you reject it, and conclude that it is a bad restaurant. A very logical, intuitively reasonable conclusion.

Suppose, in a second example, that you are walking along one of the grotty shopping streets of Amsterdam, the Kalverstraat. About halfway, you spot a man, dressed in a suit, who tells you: 'I have ten dice. If I toss the dice and if for each of the ten dice, six spots come up, you pay me 100 euros. Anything else I toss, I pay you 100 euros'. You start calculating fast. The chance that this man will toss ten times six spots

is really very small, actually it is $(\frac{1}{6})^{10}$ so the chance that you will make 100 euros is actually almost 1, namely: $1-(\frac{1}{6})^{10}$! You take the bet.

What happens next? The man tosses ten times six and you have to pay him 100 euros. What do you infer from this? Now most people respond here by saying that the man is a cheater, but that answer is incorrect. What you conclude would be that the dice are loaded, or 'iced'; they've been tampered with. In your calculations you did not reason from the assumption that the man was honest, but that the dice had equal chances of 1, 2, 3, etc. spots coming up. However, the chances of this outcome if the dice are fair are so incredibly extremely small that you now conclude that you cannot reasonably stick to your initial assumptions and keep on believing that they are fair. Whether that is because the man has tampered with them or knew they were loaded is irrelevant.

Are you *certain* of your conclusion that the dice are loaded? 'No', is the right answer here, you are not 100% certain. You have not conducted a physical investigation of the dice, you have not inspected their weight per side on a weighing scale. You have only observed their behaviour, from which you infer something about their properties. Actually there is a theoretical possibility that someone would toss ten times six spots in one toss – but the chances of that happening are really too small. It is really 'too coincidental'. And so you conclude: the chance that this toss is the result of chance is so terribly small that I dare conclude that the dice are not fair.

If we translate this example into statistics-speak, we would say that you accepted the man's bet working from the so-called *null hypothesis* or H_0:

$$H_0 : \text{the dice are fair.}$$

After the experiment, where the man tossed six spots for each of the ten dice, you reject the null hypothesis and instead accept the alternative hypothesis H_A, also referred to as H_1:

$$H_1 : \text{the dice are loaded.}$$

As said, you are not *certain* that the alternative hypothesis is true. However, given the null hypothesis, the dice-tossing results are so unlikely that you dare to switch to the alternative hypothesis.

If the null hypothesis is true, you draw the wrong conclusion. The chance that in that case you do so is however very small, namely:

$$(\frac{1}{6})^{10} = 0.00000001654,$$

or a little over one-and-a-half-millionths of a percent.

Statistical reasoning runs along exactly these lines, relatively common reasoning sets that all of us employ. In that sense, statistics is not difficult and intuitively understandable. We will now discuss statistics more formally.

8.1.2 More formalized statistical reasoning

In the examples above, we discussed two situations. In each, we had first made an assumption about the state of the world: we assumed that a restaurant served good food, and we assumed that ten dice were fair. Next we witnessed occurrences that were fairly unlikely if our assumption were true. We therefore each time concluded that the assumption was untrue, and opted for the reverse scenario: the restaurant does not serve good food, the dice are loaded.

We noted that in each instance we were not entirely certain of our conclusion. We had not subjected the cooking pots of the restaurant to bacterial tests, we had not fine-weighed the dice. We had a number of observations, a sample, and we drew a conclusion from what we saw. This means that there is a chance that we make the wrong decision. We may decide that the restaurant serves bad food, while it actually is a really clean place and we were simply both times unlucky to have fallen ill right when we went there to eat. The same goes for the dice.

We work from an assumption on the state of the world, our H_0 or null hypothesis. This null hypothesis is generally formulated as 'nothing out of the ordinary', such as 'there is no association between X and Y', or 'the IQ of law students is not different from that of psychology students', or 'the dice are fair'. Next, the researcher gathers some data. Then, it is calculated how likely it would be to find these data if the null hypothesis were true. If the results are very unlikely, the researcher decides to abandon the null hypothesis and instead assumes that the alternative hypothesis is true.

Now, if the likelihood of finding the results if the null hypothesis were true – as we say *under* the null hypothesis – is very small, and smaller than a certain threshold value, we say that the finding is *significant*. Formally speaking, the risk of making the wrong decision – or more precisely, the chance of accepting the H_1 when H_0 is true – is called α. This mistake is called a *Type I error*. But, for rejection, how unlikely then under the null hypothesis should the results be? Is 1% enough, is one millionth of a percent enough? Would a 10% chance of finding these results of still do to switch from H_0 to H_1? How small should α be?

In essence, this depends on the researcher. What this chance namely amounts to is the chance to reject H_0 while H_0 is true, and thus the chance to be wrong. It therefore depends on the researcher what risk s/he is willing to take, that is, whether s/he finds a certain α an acceptable risk. As a convention, researchers generally set their α at 5% or 0.05. This rule of thumb means that we find it in general acceptable, as researchers in academia, to abandon our initial assumption 'nothing much is going on' and instead assume that 'something special is going on', for instance that there is an association between X and Y, or that law students are indeed smarter than psychology students. We generally accept a 5% or a one-in-20 chance of 'crying wolf' when in fact nothing is out of the ordinary. This 5% (called 'α') is called the *significance level*.

This is however a convention, and other criteria play a role as well in deciding what an appropriate significance level would be. Suppose for instance that we are not talking about restaurant food or about dice and 100 euros, but about the decision whether a sex offender in a treatment institution for compulsive sex offending should be allowed to go on unsupervised weekend leave. Here the null hypothesis is:

H_0 : the sex offender is still dangerous

and the alternative hypothesis is:

H_1 : the sex offender is not dangerous anymore.

Suppose that some risk taxation instrument indicates that the sex offender is not dangerous anymore. Obviously, such risk taxation instruments are not infallible, and suppose therefore that the chance that while the offender is still dangerous the taxation states that he is not were 5%. Would we find that an acceptable risk here too? Would we find it acceptable that out of 20 sex offenders thus released on weekend leave, one would still be dangerous and thus perhaps commit a new sex offence? Given the seriousness of the consequences, we might decide here that we must scale down the chance of making a wrong decision, perhaps to 1% or even 0.1%.

Setting the significance level that low has its disadvantages, however. If we set α smaller, this implies that the likelihood that H_1 is accepted becomes correspondingly smaller. Eventually, with decreasing α (implying we have ever smaller chances to wrongly reject H_0) we might simply in practice never reject H_0 anymore. This means that we will then never decide that H_1 is true. Obviously, if H_1 were true, our test with its very, very small α will hardly ever make us state that H_1 is true, and will thus not allow us to detect that H_1 is true. Formulated differently, if we are afraid to make one kind of mistake, we may in fact be increasing the chance that we make a different kind of mistake.

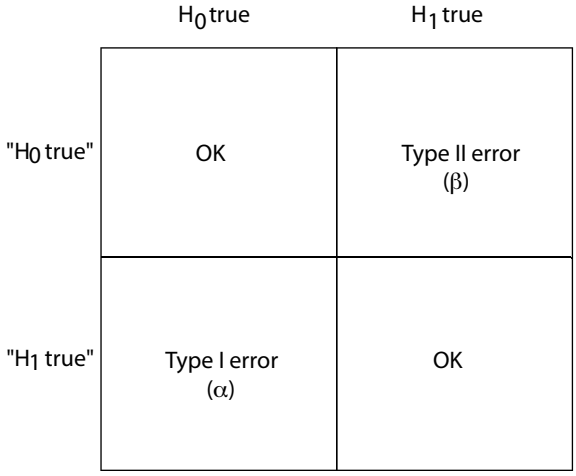
Above, we said that α is also referred to as the 'Type I error'. This suggests that there might also be another type of error. This is exactly the kind of mistake we just discussed: the mistake of not stating that H_1 is true when it is true. This type of error is named the *Type II error*. While the Type I error is the error of stating that something special is going on when everything is as it is supposed to be, failing to note that something out of the ordinary is going on is a Type II error. The probability of making this mistake is labelled β.

The possible outcomes of a statistical testing procedure are summarized in Figure 8.1. Two possible states can be true: either H_0 is true, or H_1 is true. There are two possible decisions we can take: either we decide that H_0 is true, or we decide that H_1 is true. This gives a matrix with four possible situations. See Figure 8.1.

If H_0 is true, and we decide so, we are doing things right: we take the right decision. If however H_0 is true but we state that H_1 is true, we make a Type I error. Moving to the top right of the matrix we have the situation where we also make a mistake: here H_1 is true, but we fail to acknowledge this – this is a Type II error. In the bottom left of the matrix we are doing things right again: H_1 is true, and we acknowledge this.

Note that in all the text above, we have always made *conditional* statements. We have said that *if* such-and-such is the case, *then* the chances are.... Thus, we have never said that the chances of making the wrong decision are x% or y%. We can in fact never say that, because we do not know whether H_1 or H_0 is true. Thus we only say that *if* H_0 is true, *then* ... Given that we are only making conditional statements, we cannot only give the chances of making mistakes (α and β) but also the chances of making the right decision. If namely H_0 is true, and if the chance of making the wrong decision (Type I

Figure 8.1: Overview of testing



error) is α, it follows that the chance of making the right decision is $1–\alpha$. Similarly, if H_1 is true, and the chance of making the wrong decision (Type II error) is β, it follows that the chance of making the right decision is $1–\beta$.

Because the hypothetical situations H_0 and H_1 are constructed so differently, with H_0 the 'normal situation' and H_1 the exceptional case where something really special is going on, the Type I and Type II errors are also fundamentally different mistakes, and of fundamentally different importance. While not a perfectly applicable analogue, the example of a fire-alarm serves the purpose of explaining why.

Suppose that my house has a fire-alarm. If that fire-alarm also rings when someone is frying eggplants in my kitchen, it is set too lightly, it is too sensitive. Such false alarms (where I have to run upstairs and slam the alarm with a newspaper to silence it) are annoying – they are the analogue of the Type I error: much ado about nothing. Clearly this is something we do not want. The fire-alarm can, however, make a different kind of mistake too. If its sensitivity is too low, there is the possibility that it will not detect a fire. This is the analogue of a Type II error: not detecting when something out of the ordinary is going on. Clearly this type of mistake is not just annoying, it is unacceptable and in fact potentially deadly.

The example shows the different nature of the two types of errors. They are not simply each other's mirror image; they are of a qualitatively different nature. Ideally one would want the chance for Type I error as well as the chance for Type II error to be as small as possible. Obviously, they are linked. If I level up the sensitivity of the fire-alarm, it will detect a fire sooner, but also more often give false alarms. If I decrease the sensitivity, I will have fewer false alarms, but the chance will be increased that a real fire is missed. The situation in testing is just like this: if I lower α, this increases the β, if I allow α to be bigger, this generally decreases β.

Before we move to the finale of this section, we zoom in on a particular area of the matrix in Figure 8.1. The lower right corner denotes the situation where something out of the ordinary is going on (H_1 is true) and where the test is able to detect that. The chance that this happens if H_1 is true, is $1–\beta$. In terms of the fire-alarm, this is the chance that the fire alarm starts screeching when there is a fire. In statistics this chance, $1-\beta$, is generally referred to as the *power* of the test. Ideally, we would want the power of any test, just like our fire-alarm, to be really good. However, while we can ourselves decide how high we want to set the α, this is not generally possible for β and thus also not for $1–\beta$.

In general, the power of a test is reduced as samples become smaller, and is increased as samples become larger. This is only intuitively logical, because when our samples become larger we inspect a larger part of the population and therefore must be able to better discern what is going on. For more details we refer to statistical textbooks, such as Hinkle et al. (2003).

All this sounds – hopefully – pretty straightforward. But is testing always possible? Can things go wrong when we test? Yes, they can. Firstly, it may be the case that the sample we are working with is simply too small to generate meaningful results. When samples are too small, almost no difference or association will turn up as being *significant*, i.e. no test will decide that H_1 is true instead of H_0. Tests are then simply unable to detect what is going on, they have too few observations to confidently draw conclusions, and will then always opt for 'nothing out of the ordinary'. Small samples are therefore of little use, statistically speaking, and samples of about 30 are generally regarded as at the absolute lower limit. From sample sizes of about 100 onwards, statistical tests can be carried out with more confidence and stronger conclusions drawn – for reasons to be detailed below.

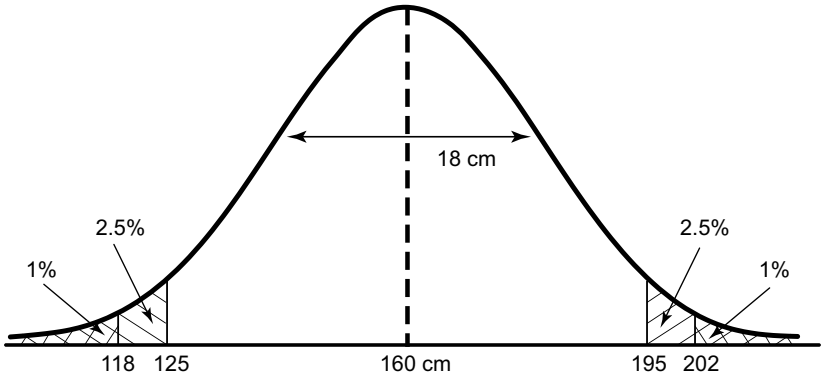
Secondly, testing may be hampered if the data are very noisy. Any association or difference (see also above section 5.4) in a sense 'disappears' in the noise. Tests then cannot detect differences anymore; these are masked by the large variability in the data. Lastly, it may be the case that the assumptions of the testing procedures have been violated. By the latter we mean for instance that a statistical test assumes that the data are normally distributed, and generates significance levels based on that assumption – but the data are differently distributed, rendering the tests invalid. Such issues are detailed in statistics textbooks and are beyond the scope of this book.

8.2 Distributions and tests

In the above examples, we were mostly pretty vague. We spoke of 'unlikely' and 'too coincidental' mostly, and only in the example with the dice did we give exact chances for the occurrences under the null hypothesis. We did also announce that we would become more precise later on, and this is the paragraph where we will do so. Again we start with an example.

One crisp autumn morning, the sun barely risen and the air damp and foggy, I step out of my kitchen door to collect the windfall from my apple tree, when suddenly I spot in my back garden a huge dead heron. Yuck. It lies sprawled across the grass. Even though my garden is not large, this poor beast appears to almost stretch the entire

Figure 8.2: Normal distribution of wing span of *Ardea cinerea*



breadth of the garden and it does look like a lot for a heron! I am convinced it is a huge beast, it must be one of the largest of its kind, I think, that has crashed in my garden.

I measure the dead heron's wing span. It is 2 metres and 3 centimetres. That really seems like a lot. This heron must belong to the top 1% largest of its kind! I look up *Ardea cinerea* (the Latin name of the blue heron): it appears that the wing span of this type of heron is 'normally distributed' with an average wing span of 1.60 metres and a standard deviation of 18 cm. What does that tell me?

Figure 8.2 graphically depicts this information: it is a picture of a normal distribution with a mean of 160 cm and a standard deviation of 18 cm.

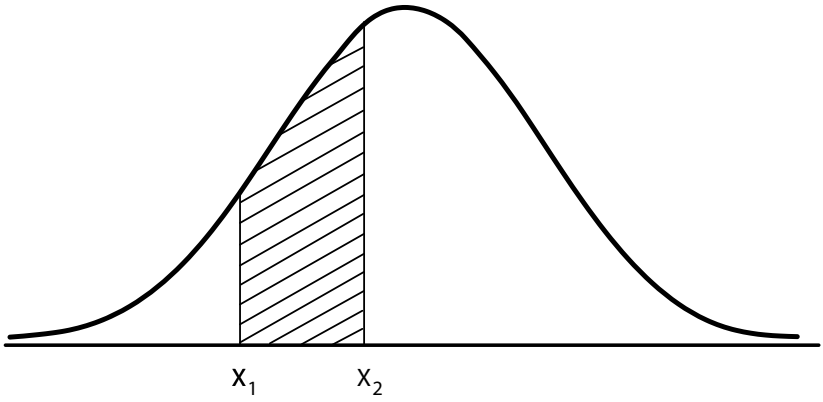
This figure illustrates firstly that a wing span of 160 cm is the most frequent: the peak of the distribution is at 160 cm. This is because the y-axis reflects the number of herons for the various wing spans X. With most herons having a wing span of 1.60 metres, as we move to smaller and larger wing spans, we see that the numbers tail off. Fewer and fewer herons are encountered as we move to more extreme values. In the picture, it is also shown that, on average, wing span deviates 18 cm from this average of 160 cm. The standard deviation is 18 cm. We see also that the distribution of wing span is symmetrical: the distribution has exactly the same shape on the left- and right-hand side of the average. What, however, does it mean when it is said that wing span is 'normally distributed'?

The normal distribution is a distribution that is often encountered in the natural world. Many properties of living beings are normally distributed: height, intelligence, whale teeth size, etc. All these properties have in common that they cluster around a mean. The normal distribution is also referred to as the Gaussian distribution, or more popularly as the bell curve[1].

[1]The normal distribution can be described by its *probability density function*

$$p(\mathrm{x}) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(\frac{-(\mathrm{x}-\mu)^2}{\sigma^2}\right)$$

Figure 8.3: Area under the curve



This probability density function is bell-shaped (hence the 'bell curve'), being more or less attenuated depending on the standard deviation. For now, it is enough to know that the height of the curve (the value on the y-axis) reflects the number of cases with a certain value X on the variable, the so-called *density*; the density is a mathematical concept best understood as – but not equal to – a frequency-like measure. The density has however nice properties that make it possible to translate the properties of the curve directly into statements on chances and probabilities. The probability that a variable assumes values within a certain range, say from X_1 to X_2, is equal to the *area under the curve* of the probability density function between X_1 to X_2. See Figure 8.3, where this has been indicated.

This property is very useful as it enables us to make lots of inferences. Firstly, because the distribution is symmetrical, we know that the chance that a randomly selected heron's wing span is more than 160 cm is equal to 0.5. For the same reason, the chance that a randomly selected heron's wing span is less than 160 cm equals 0.5 as well. Because chances are always between 0 and 1, and the sum of all possible outcomes equals 1, the total area under the curve must be equal to 1, and the mean neatly chunks that area into two – because the distribution is symmetrical, necessarily equal – parts. Note that the chance that a heron's wing span equals exactly 160 cm is not defined: chances are defined as areas under the curve, and the point 160 has no breadth and therefore no area defined under it. This means that we can only determine the likelihood of a *range* of outcomes, say a value being between values 160 and 180, or between $-\infty$ (minus infinity) and 140 cm.

The normal distribution has a number of useful properties. As said, firstly, it is symmetrical. This has the nice advantage that we know that the probability of a heron having a wing span larger than 160 cm is equal to the probability of a heron having a wing span smaller than 160 cm. Secondly and most importantly, because it is possible to compute the probability that a variable assumes values within a certain range A to

B, we can also compute the chance that an observation from a normal distribution is higher than a certain value, or lower than a certain value. This is exactly what I wanted to do for my poor, magnificent, dead heron: I wanted to know what the chances are that a heron has a wing span between 203 cm and ∞.

While it is possible to do these computations oneself, many statistical textbooks, online gadgets and statistical software packages do it for you. Using a statistical textbook (e.g. Hinkle et al., 2003) or surfing on the internet, I can look up a table for the normal distribution and find that – for this normal distribution with mean 160 and standard deviation 18 – 95% of all herons have a wing span between approximately 125 and 195 cm – with only 2.5% above 195 cm and only 2.5% having scores below 125 cm – and 98% having a wing span between 118 and 202 cm – with only 1% having a wing span above 202 cm and only 1% below 118 cm. My poor dead heron, with his wing span of 203 cm, thus indeed belongs to the exclusive 1% of herons with wing spans of 202 cm and more.

8.2.1 Standard scores and testing

If we work with scores such as the wing span of the dead heron discussed just now, the values of the variables we are working with are often *standardized*. We do this because – while the raw scores are relatable – it is not easy to get a feel for how deviant a score is. We devoted quite a number of paragraphs in the previous section to revealing whether the dead heron was extraordinarily big or not. Formulated differently, working with raw scores it is hard to get a 'feel' for the distribution of the variables. I had no clue whether my dead heron with its wing span of 203 cm was really so extraordinary. If we however turn raw scores into standard scores, we much more easily grasp how far from the mean and therefore how deviant scores are.

Standardization usually is done by subtracting from every score the mean, and then dividing the result by the standard deviation:

$$\frac{X_i - \bar{X}}{s_x}.$$

The resulting values are called standard scores, or standardized values, denoted by z or z_x. Standard scores always have a mean of 0 and a standard deviation of 1. The resulting normal distribution of these scores is now called the *standard normal distribution*.[2]

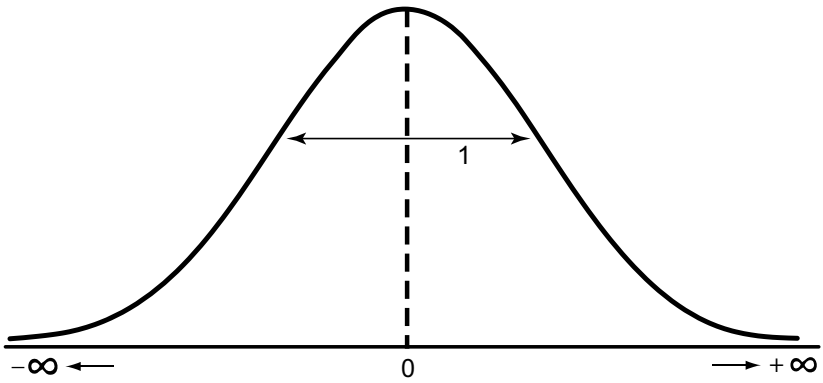
A standard score of 2 already is quite informative: it tells me that on this variable the respondent deviates twice as much as average (namely twice the standard deviation) from the mean. A standard score of -.5 tells me that this respondent is in fact much less 'deviant', the score is only half a standard deviation below the mean. The standard normal distribution is given in Figure 8.4.

In this figure, the mean is at 0, and as scores become lower or higher, they are rarer. In the limiting case (when X approaches either ∞ or -∞), the curve approaches

[2]The probability density function – given that μ equals 0 and σ equals 1 – simplifies to:

$$p(x) = \frac{1}{\sqrt{2\pi}} exp^{-x^2}.$$

Figure 8.4: Standard normal distribution



the x-axis. The curve never reaches the x-axis, it creeps closer and closer. The area under the curve equals 1. The area above 0 equals .5, and the area below 0 equals .5 as well. Tables in any textbook will list the chances of finding observations above certain values. For instance, such tables will tell you that the area beyond 1.99 equals .0233 or 2.33%. This tells you that the chance that a respondent has a standard score higher than 1.99 equals .0233. Because the distribution is symmetrical, the chance that a respondent has a standard score lower than -1.99 is then also 2.33%. The more extreme the score, the less likely it is to find respondents with such scores.

Now how does this translate to testing? Suppose that we have two hypotheses, of which H_0 is:

$$H_0 : \text{Law students are as bright as psychology students,}$$

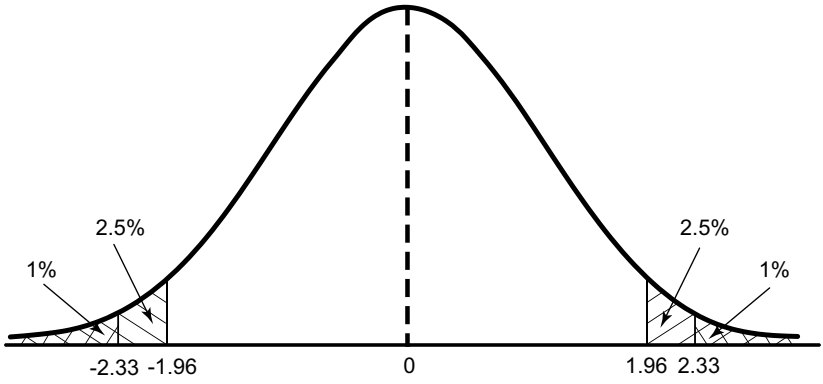
and the alternative hypothesis H_1 is:

$$H_1 : \text{Law students are not as bright as psychology students.}$$

Suppose further that I have measurements of the IQ of a random sample of law students and of the IQ of a random sample of psychology students. On average, the IQ of psychology students is 122. The standard deviation of their IQs is 2. For the law students, I also have measurements of IQ in exactly the same manner. I find that the law sample's IQ is on average 127. Clearly, this is higher. But is this result so striking, so significant, that I dare to decide that the law students have an IQ that differs significantly from the IQ of the psychology students? Could this not simply be a chance result, coincidental? How likely is it that if the law students are actually no different, that by chance I could end up with these results in my sample? This likelihood, this chance, is exactly what the statistical test will tell you.

We go about finding the answer as follows. Our null hypothesis is the hypothesis that states that there is no difference, nothing is out of the ordinary, there is nothing

Figure 8.5: Standard normal distribution with critical standard scores



special going on. So our null hypothesis is that law students stem from the same population as psychology students, namely a population where the mean IQ equals 122, so $\mu_{\text{IQ}} = 122$. We also assume that in the population of law students the standard deviation of IQ is 2, like we measured for the psychology students. So $\sigma_{\text{IQ}} = 2$. The hypotheses are now:

$$\text{H}_0 : \mu_{\text{IQ law students}} = 122,$$

and

$$\text{H}_1 : \mu_{\text{IQ law students}} \neq 122.$$

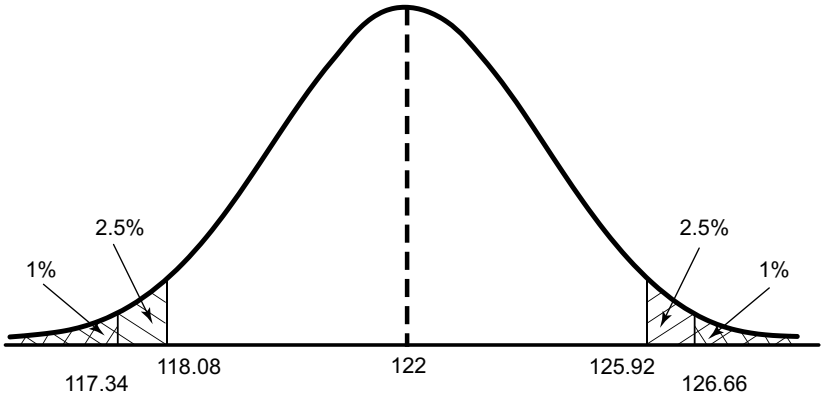
Note that in the hypotheses, which specify the properties of populations, we use Greek letters. Before we do the testing, first the raw scores are rewritten as standard scores:

$$\frac{\text{X}_\text{i} - \mu}{\sigma} = \frac{127 - 122}{2} = 2.5.$$

Now, this value of 2.5 is a z-value, a standard score. This means that we can – assuming that the data we are inspecting are normally distributed – look up what the chances are that a z-value exceeds 2.5. Tables with z-values are in the appendices of any statistical textbook, and can also be found on the internet; see for an application `psych-www.colorado.edu/~mcclella/java/normal/normz.html`. Looking up the z-score, we find that if the z-score were 2.33, the area under the curve beyond 2.33 equals 1%. With the value of z equalling 2.5, the area under the curve therefore sums to a little under 1%. This means that the chance that, if I draw this sample from a population with a mean of 122 and a standard deviation of 2, the chance that I find this result of 127 is a little under 1%. A wee chance. See Figure 8.5.

If we translate this into words, we say that *if* H_0 is true, the chance of finding this result is about 1%. This is pretty unlikely. Not as unlikely as the 0.00000001654 in the dice example, but even so only a one-in-a-hundred chance to find this result if H_0

Figure 8.6: Normal distribution with critical raw values



is true. So, we might decide that H_1 is true. Are we *sure* about this? No, we are not. It is possible to find an IQ of 127 in our sample of law students if they come from a population with a mean of 122. But the chances to find such a result – depending on the sample size – are actually pretty small.

So, to reject the H_0 would be a logical thing to do. There is a chance – if the H_0 is true – of around 1% that we make the wrong decision, but in general that is viewed as an acceptable risk. So, in this example, we would state that this year's law students have significantly higher IQs than psychology students.

Above, we looked at the sample mean and investigated what the chances for this result were under the null hypothesis. However, we can also work the other way round. Then, we decide first on the acceptable risk to make the wrong decision (α), draw the probability density function, and compute the so-called *critical value*: any sample mean above this critical value leads to rejection of the H_0. See Figure 8.6, in which we have depicted the critical values for this example for α's of 5% and 2%. So, if we are willing to accept a risk of making a Type I error of 5%, this means that we reject H_0 if the sample mean is above 125.92, or below 118.08. If we feel that 5% is a risk we dare not take, we set the critical value sharper, rejecting sample means only above 126.66 or below 117.34. These results are intuitively reasonable: if we want to reduce the risk of a Type I error, we reject only from values further away from the mean; we need scores that differ more from the postulated population mean. See Figure 8.6.

This is a very summary description of the procedure of testing. We have discussed only the normal distribution. There are numerous other distributions, of which we name the t-distribution, the F-distribution, the χ^2-distribution. They generate different numbers and are applied slightly differently, but the principles are identical: if the value exceeds a critical value, the H_0 is rejected. There are also distribution-free tests. We will leave such other tests mostly undiscussed, as they are not essential for the purposes of this text. Suffice it to say that the rationale of testing, the choice of hypothesis, the

procedure of rejection or acceptance are all substantively identical: any differences are in the mathematics. See for more information on testing any statistical textbook, for instance Hinkle, Wiersma, & Jurs (2003, chapters 14, 21 or 22).

8.3 Sampling, significance and power

If we summarize what we have described so far, the story goes as follows. In many research situations, we do not have the possibility to investigate all objects of interest. Instead, we investigate a small part of that entire collection, that population. We do that in a clever way. We make sure that the part that we investigate resembles the larger universe it comes from: we pick the sample members by chance. This ensures that any differences between the sample and the population are accidental differences, they are not systematic. In this way, we arrive at an unbiased sample, which is representative.

Obviously, we are not certain about the properties of the population: we have studied only a part. How do we deal with this uncertainty? We do that in a formalized way. This formalized manner of thinking is intuitively logical and reasonable. Because of the particular rules and calculations we follow, it becomes statistical reasoning. What we do is to calculate what the chances of finding our observations are, if a certain assumption about reality were true. If these chances are very small, we switch to the alternative assumption about reality. We then say that the finding is *significant*.

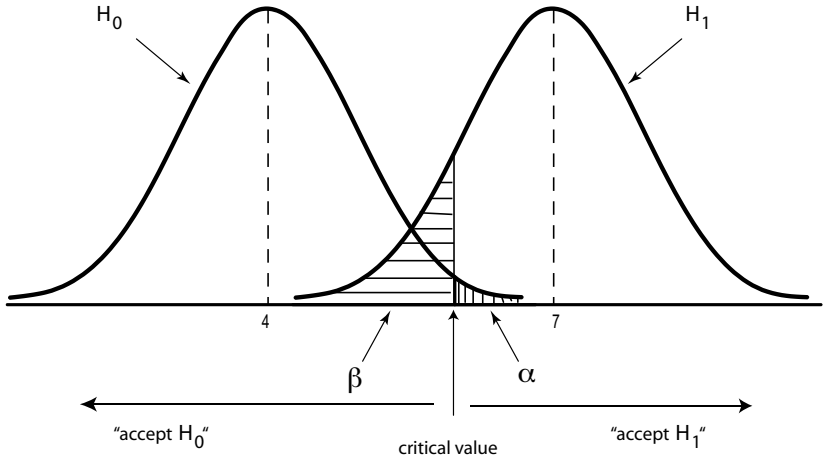
Significant is a term we use in our daily lives as well, it means something like 'important' or 'meaningful'. In statistics, it has this particular meaning: the chances of opting for the alternative scenario if the regular situation were true. In general, we work with significance levels of 5% or 1%, depending on sample size and other considerations. We then say that we 'set α at 5%' or that such-and-such finding is significant 'at the 5% level'.

We want our chances of making a wrong decision to be as small as possible. Deciding that the alternative scenario is true while the regular scenario holds is one type of mistake. As we describe the regular scenario fairly precisely under H_0 (e.g. $H_0 : \mu = 122$), we can make fairly exact calculations as well. If we assume that the data are distributed in a certain way (normally, or F-distributed, or χ^2), or if we know what the chances of a certain number of spots are when we toss dice, we can therefore compute exactly what the chances are that a certain extreme result appears. This means that we can always ourselves determine α. It is for us to decide, and we can actually decide exactly, what risk we are willing to run of making a Type I error.

However, as described earlier, there is a second kind of mistake we could make. We could also on the basis of the empirical results in our sample decide that there is no reason to reject H_0 while in fact H_1 is true. Then we would make a Type II error. The chances for that we call β. Now, while we can set the α, we cannot beforehand design the test in such a way that we have an acceptable β. Why is that so?

The reason that we cannot pre-determine the β the way we can choose our α is because the alternative hypothesis actually comprises a range of options. In other words, while H_0 gives one and only one μ (such as 122 in the example above), under H_1 the μ can assume many different values: 121, 180, 85... So how are we then to compute what the risk is to make a Type II error? We cannot.

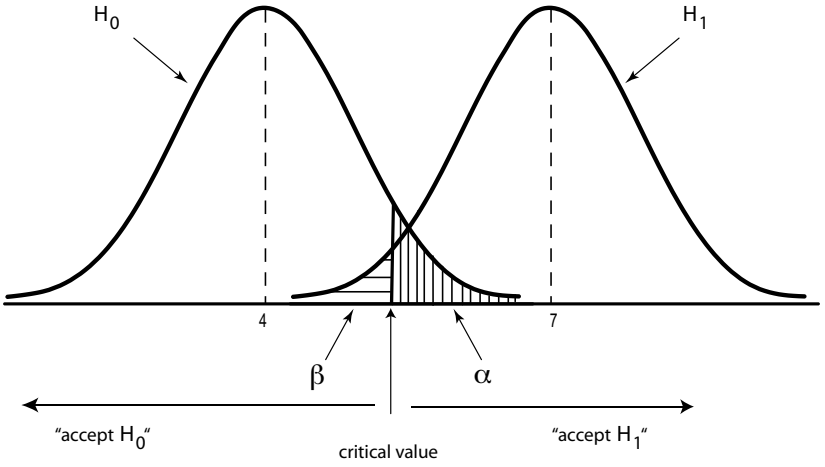
Figure 8.7: Distribution of coping skills under H_0 and H_1 with α=1%



What we can do, however, is pick a certain alternative μ, and compute – for that alternative μ – what the chances are that we would actually miss discovering that H_1 is true. Let us illustrate this with an example. Suppose that we are looking at the coping skills of ex-detainees. While they were detained, their coping skills were 4 on average. We have trained all these detainees to increase their coping skills. A coping skills level of 7 is sufficient for normal societal functioning, while 10 means excellent societal functioning. So, obviously, we would be really pleased if the training we gave our detainees increased their skills to a level of 7: that would mean they did really well. This also means that we would be really unhappy if we were to test the trained ex-detainees, but our test was unable to pick up this relevant difference. Figure 8.7 depicts the testing situation where α has been set at 1% but where in reality – but we don't know this – μ equals 7.

Figure 8.7 shows the α as the area under the H_0 curve beyond the critical value, as an area that has vertical stripes. The area under the H_0 curve to the left of the critical value is the chance that I accept the null hypothesis if H_0 is true. As the total area under the curve is 1, or 100%, and as α is 1%, this chance is therefore 99%. The second curve, on the right-hand side, is the situation when actually H_1 is true. We do not know whether this is the case, but *if* it is the case then this is how the coping skills scores of our ex-detainees would be distributed, around 7, and with a normal distribution. Looking at the curve for H_1, we can see that the chance that we miss discovering that H_1 is true is equal to the area under the H_1 curve to the left of the critical value (the area with horizontal stripes). This is where we end up when we find a not very unlikely result in our sample and therefore see no reason to reject the H_0. If H_1 is true, then the chance that we make the right decision, that we decide in favour of H_1, is equal to the area under the H_1 curve to the right of the critical value: this is the

Figure 8.8: Distribution of coping skills under H_0 and H_1 with α=5%



chance that we decide for H_1 if H_1 is true. This is the power of the test.

The power in this example would be – eyeballing the graph – somewhere in the range of 80–85%, so a four-in-five chance if H_1 were true, that our test would lead to the right conclusion. In general, this is considered a reasonable power. If, however, we decide that that is not enough, that we want to have a higher chance of discovering that H_1 is true if it is, then – all else being equal – we can achieve this by relaxing α. If we decide that 5% is also acceptable as a chance for making a Type I error, this reduces the Type II error, and increases the power. This illustrates that α and β are connected, although not by some simple straightforward formula. In general when one decreases the chance of one type of error, the chance of the other type goes up, and vice versa. See Figure 8.8.

While the chance of a Type I error is determined by the researcher, and generally set at a value between 5% and 1%, the chance of a Type II error, β, and therewith the power 1–β, cannot be pre-set. This is due, as we stated, to the fact that the alternative hypothesis covers many possible values for the property we are studying. What we can do is deduce what the power would be for one particular value of the parameter of interest where the H_1 is true.

We showed how increasing the α decreases the β. So, to increase the power, one has to accept paying the price of a higher chance of the other type of mistake, which is not very attractive. We gain in one area but we simply pay in the other. Is there then no other way to increase the power? For sure there is, and that manner is intuitively understandable. Obviously, if we increase our sample size and investigate a larger chunk of the population, we will have more information, so it can only be that we get better estimates, that we observe more sharply what is going on, what the properties of the population are. Inspecting a larger sample must therefore reduce the chance that

we make mistakes. It would be highly illogical if this were not the case. Below, it will be illustrated using formulas that this does indeed happen: if sample size increases, the chances of errors go down, so that with a given α the power increases. Larger samples have more power.

Several tools are available to aid in choosing the necessary sample size for a certain desired power level. One such tool is G*Power, available as freeware online: the researcher needs to input the test s/he wants to perform (such as a test for the difference between means), chosen α-level, the effect size that the researcher wants to be able to detect and the desired power level. The tool will then return the required minimum sample size. See for quick introductions `stats.oarc.ucla.edu/other/gpower/` and `psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html`; for more background see Faul et al. (2009).

8.3.1 Hypotheses revisited

In the previous section, some readers may have been confused by the following. If we believe that the dead heron we found is extremely large, and if we believe that this year's law students are brighter than psychology students, why then do we have rejection regions both at the left-hand and right-hand side of the graph as in Figures 8.5 and 8.6? Given that in both cases the result in the study is higher than expected under the null hypothesis, why would I need a critical value for the opposite result? The fact that in each of the examples we have areas under the curve both in the left and the right tail of the distribution is a consequence of the fact that we are investigating a so-called *two-sided hypothesis*, in the example of the law students' IQs:

$$H_0 : \mu = 122,$$

and

$$H_1 : \mu \neq 122.$$

If we design the null hypothesis this way, it means that we expect the mean to be 122, and that if it is either very much higher, or very much lower, we reject H_0. In both cases the result is so unlikely that we no longer state the null hypothesis to be true. See Figure 8.9, which sketches the situation.

In this figure, it can be seen that the null hypothesis is rejected when $\bar{X}$, the mean IQ of law students, is either larger than 126.66 or smaller than 117.34. If H_0 is true, and the observations have therefore actually been drawn from a population with mean 122, I have when I test in this way a chance of 1% that I find a sample mean higher than 126.66 and thus a 1% chance of drawing the wrong decision. If I find a sample mean lower than 117.34, then I will also draw the wrong conclusion and the chances of that happening are also 1%. So, in total I accept, in this scheme, a 2% chance of making the wrong decision: 1% if the sample mean is accidentally very low, and another 1% if it happens to be very high. See again Figure 8.9.

Obviously, if I have strong reasons to suspect that the deviation from H_0 is in a certain direction, I do not really need the rejection area on the left-hand side. Then, if I stick to my α of 2%, the critical value from which point onwards I will reject the

Figure 8.9: Test situation for two-sided null hypothesis with critical raw values

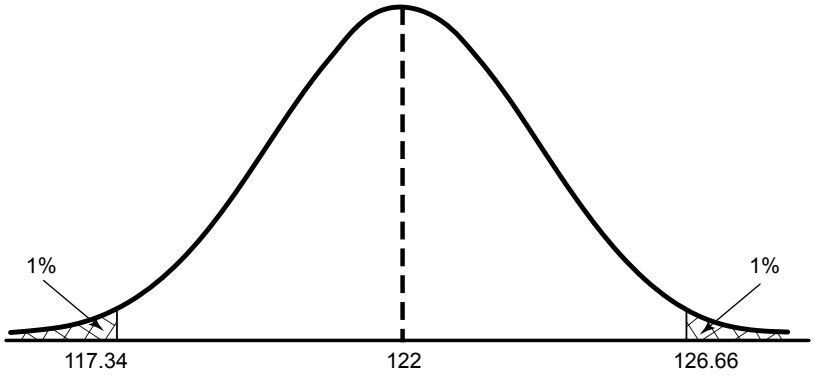
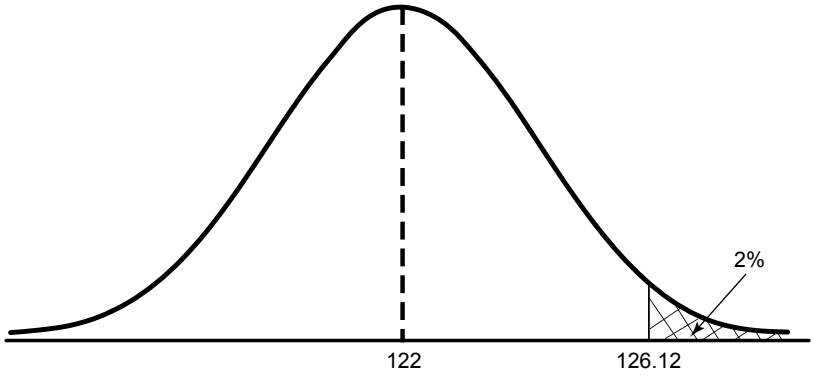


Figure 8.10: Test situation for one-sided null hypothesis with critical raw value



H_0 also changes, see Figure 8.10. The 2% area under the curve that was first divided between the left- and right-hand side of the curve, is now lumped on one side. The corresponding null and alternative hypotheses are:

$$H_0 : \mu \leq 122$$

and

$$H_1 : \mu > 122.$$

This type of hypothesis is called a *one-sided hypothesis*. This means that by formulating the null hypothesis as a one-sided hypothesis, the critical value is lowered. For-

mulated differently, if I had constructed the null hypothesis as a two-sided hypothesis and my sample mean had been 126.40, I would not have been able to reject the null hypothesis: 126.40 is outside the rejection area. I would have been really disappointed as my belief that law students are brighter was so strong and I did not see it confirmed.

If however I construct my null hypothesis as a one-sided hypothesis, and I find for the sample a mean of 126.40, I can suddenly reject the H_0! The α is the same, but in this case I reject and in the other I could not. This reeks of 'lies, damn lies and statistics', some would say, and clearly this offers possibilities for fiddling.

That is why there are clear rules for how hypotheses should be formulated. The first rule is that the hypotheses must be constructed before the data are investigated. They must be constructed beforehand, so that any post-hoc fiddling is precluded. Secondly, there is a clear rule that one-sided hypotheses are formulated and tested only when there are strong theoretical reasons to suspect that any effect will be in one direction only. Obviously, working with one-sided hypotheses has its risks too: if the sample mean suddenly does happen to be very low in the one-sided hypotheses formulated above, H_0 cannot be rejected even though it is obviously very different from what was expected.

Note that the null and alternative hypotheses must always complement each other: together they encompass the entirety of what could possibly be the state of affairs.

In the examples so far, we have always sketched the situation where one is interested in finding the mean μ of the population. It is however also possible that we would be interested to know whether two properties are associated in the population, in which case we could test whether a correlation coefficient differs significantly from zero, or whether an odds ratio differs significantly from 1.

8.4 Theoretical foundations of sampling and testing

In all the texts so far, we have been fairly sloppy, in the sense that we talked about estimation but never defined exactly what we meant by that. Also, we implicitly assumed that the sample mean $\bar{X}$ would be a good estimator of the population mean μ. We talked about normal distributions and we implicitly assumed but never explicitly proved that the sample mean itself would follow a normal distribution. Textbooks and theoretical books have been written about these issues, but much of the detail and theory there are not essential for our purpose which is giving a brief conceptual outline of statistical reasoning and methodological procedures. There is one issue that we announced we would discuss in more depth, however, and that is the issue of sample size. We said that it is intuitively clear that working with larger samples gives better results. But can it be shown that that is so? Yes, it can. What we will use for this is the *central limit theorem*, which states that as the size N of a randomly drawn sample increases, the distribution of the sample mean, regardless of the original distribution (with mean μ and finite variance σ^2) from which the sample was drawn, approximates a normal distribution.

This theorem is central to statistical testing. What it first says is that, regardless of how the original variable was distributed, the sample means of that variable – if I take care to draw large enough samples and draw them randomly – is distributed normally.

So if I would draw 100 samples from a population, each time randomly, the sample means would vary but they would follow a normal distribution. This theorem therefore is the basis for using the normal distribution to test whether a sample mean differs significantly from what would be expected under the null hypothesis.

Secondly, the theorem can be used to show that – something that is again intuitively appealing – the sample mean $\bar{X}$ is an unbiased estimator of the population mean μ.

Thirdly, the theorem can be used to show that as sample size increases, the sample mean becomes a more precise estimator of the population mean. The latter is so because the standard deviation of the sample mean, also called the *standard error of the mean*, equals:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}.$$

Given σ, the standard deviation of the mean of the scores in the population, $\sigma_{\bar{X}}$, therefore becomes smaller as N increases. Intuitively, this can only be so: as I investigate larger proportions of population members, my sample mean can only approach the population mean more and more. The formulas however show how fast this goes: increasing the sample size by a factor 4 decreases the variability of the sample mean by a factor 2. As samples become bigger, the sample mean fluctuates less around the population mean. In the limiting case, as N approaches the population size, the variability approaches zero.

Fourthly, the theorem underscores the importance of random sampling. If we do not draw randomly, then the theorem does not hold, and we miss out on all the associated niceties.

Lastly, the theorem states that as sample size becomes bigger, the sample mean becomes a more precise estimator of the population mean. The nice thing now is that statistical research has shown that already from sample sizes of about 30 the distribution of the sample mean is pretty normal. From a sample size of a little over 120, it is for practical purposes entirely normal. The central limit theorem is one of the reasons why we use the rule of thumb that N, the sample size, should be about 100. In that case, the sample mean is approximately normally distributed, so that we can use the normal distribution for significance testing.

The central limit theorem is, all in all, indeed very central to statistical testing. It tells us that sample means are unbiased estimators of population means, it instructs us to draw random samples, it tells us that large samples are less noisy than smaller samples, and it tells us that we can use the normal distribution to test whether a sample mean differs significantly from a sample mean expected under the null hypothesis. The central limit theorem provides a rule of thumb in the sense that samples of 30 at the minimum are – for simple testing purposes – of sufficient size.

The central limit theorem is formulated assuming that the population variance σ^2 is known. While we assumed that as well in most of our examples above, the population variance is obviously more often unknown than known. In such cases, we use the sample variance as an estimate of the population variance. In that case, however, a sample size of 30 is not sufficient and we have to test using a different distribution, the so-called *Student's t-distribution*. From samples of 100 upwards, the t-distribution

is again practically equivalent to the normal distribution again, and we can use our standard testing procedure.

In summary: given that samples have been randomly drawn, if sample size is above 100, we use the normal distribution to test hypotheses on the population mean. If samples sizes are between 30 and 100, we use the normal distribution if the population variance is known and the t-distribution if it is unknown and we need to estimate the population variance from the sample variance. For samples under 30, testing becomes harder, and in practice the t-distribution is used. Also, researchers then resort to so-called non-parametric tests, statistical tests that do rely on a specific underlying distribution like the normal distribution.

8.5 Estimation and uncertainty

In the above, we have regularly spoken of estimation, of the mean of the population, of the variance of the population, and similarly we might want to estimate the association between variables. Each time when we estimate, we attempt to find the value of a so-called *parameter* in the population. The mean is a parameter, as is the variance or standard deviation. They are properties of the distribution of variables that describe or characterize that distribution. The mean tells you around what value the values cluster and the standard deviation tells you how much values on the variable fluctuate around that mean.

8.5.1 Point estimates versus interval estimates

Above we said, basing ourselves on the central limit theorem, that the sample mean is an unbiased estimator of the population mean. We write the sample mean as an estimator of the population mean as follows:

$$\hat{\mu} = \bar{\mathrm{X}},$$

where the hat above the μ tells us that we have not actually measured μ and that we do not equate μ to the sample mean, but that we employ $\bar{\mathrm{X}}$ as a mere estimate only.

Similarly the sample variance can be used to estimate the population variance:

$$\hat{\sigma}^2 = \mathrm{s}^2,$$

where again the hat tells us that the sample variance is used as an estimator of the population variance.

Now all these estimates are what we call *point estimates*. They tell us what we infer from the sample about the population, what we believe μ in the population to be on the basis of the sample information. However, in some cases we may be more certain about this estimate than in other cases. For instance, when we have investigated only a small sample, we are – with an identical value for that estimate – obviously less certain about that estimate than if we had drawn a large sample. A way to give an indication of that certainty or uncertainty is to present an interval estimate. We will discuss interval estimates or confidence intervals in the next section.

8.5.2 Confidence intervals

Many estimates in quantitative research are reported with a so-called confidence interval. What does this stand for? We illustrate it with a numerical example.

Let us assume for our law students' IQ example above that we found that the sample mean is 122.8. Let us also assume that we rejected the null hypothesis ($H_0 : \mu = 122$), that the standard deviation in the population is 2, and that we had 100 students in our sample. We work with an α of 5% and as we use a two-sided null hypothesis, the testing situation looks like it did in Figure 8.5: we have two areas of rejection, each good for an area of 2.5% under the curve. We have therefore two critical values: the critical values for the z-distribution beyond which we would reject are -1.96 and +1.96.

Our estimate of law students' IQs is the sample mean, which is as we said 122.8. Obviously, as we have drawn a sample, we are not certain that this is the mean IQ in the population of law students, it is an estimate. We can express the certainty with which we report that estimate using a *confidence interval*.

For computing the confidence interval around that estimate, we use the following formula:

$$\text{CI}_{95\%} = \bar{\text{X}} \pm z_{\text{critical value}} \times \sigma_{\bar{\text{X}}},$$

where '$\text{CI}_{95\%}$' stands for the 95% confidence interval. When we plug our data into the formula we get:

$$\text{CI}_{95\%} = 122.8 \pm 1.96 \times \frac{2}{\sqrt{100}},$$

which is equal to

$$\text{CI}_{95\%} = 122.8 \pm 0.392,$$

so that the 95% confidence interval in this example is 122.408–123.192. How do we say in words what this means? A loose, intuitive way of describing the confidence interval is as your margin of uncertainty. A more precise way is to say that we are 95% confident that μ_{IQ} is in the 122.408–123.192 interval.

Note that the value for μ under the null hypothesis that we are testing, 122, is not contained in the confidence interval we just computed. It would be very strange if it were, because we just rejected the H_0 – by which we said that we believe that the μ is *not* equal to 122. So it would be contradictory if now it were contained in our confidence interval!

Confidence intervals are regularly used. For instance, for assessing whether an odds ratio is significant, mostly its confidence interval is computed. If that confidence interval contains 1, we say the odds ratio is not significant, if the confidence interval does not contain 1, we conclude that the odds ratio differs significantly from 1.

As said, the confidence interval gives you a margin of uncertainty: wider confidence intervals mean more uncertainty, narrower confidence intervals mean more certainty. From that it follows that if we were to investigate a larger sample, in which case we would be more certain of our estimate, the confidence interval would have to become

correspondingly narrower. Is that the case? We can simply check this by recalculating the interval for a sample size of, say, 144 instead of 100. The calculations then become:

$$\text{CI}_{95\%} = 122.8 \pm 1.96 \times \frac{2}{\sqrt{144}},$$

so that

$$\text{CI}_{95\%} = 122.8 \pm 0.327,$$

so that the 95% confidence interval in this example is (122.473–123.127), which is indeed narrower than the 95% confidence interval when the sample size was 100. Similarly, it can be computed that as sample size decreases, the confidence interval widens.

These examples have given the 95% confidence interval. It is also possible to compute a 99% confidence interval. As we are in that case talking about more certainty it is logical that with the same data the interval would become wider. If we wanted to be 100% certain, the interval would span $-\infty$ to ∞!

We have given examples where samples of 100 and over were investigated, and where the population variance was known, so that regular z-scores could be used for the critical value. If however, the population variance is unknown, and smallish samples are used, then just like for the testing situation here too for critical value a value from the t-distribution must be used.

Leaving aside the formulas and all the nitty-gritty tech specs of when to use z- and t-values, the important thing here to remember is that the confidence interval reflects statistical precision. Narrow confidence intervals reflect that we are more certain of our estimate, wider confidence intervals reflect the reverse. Confidence intervals narrow and widen with sample size, variability in the data, and the level of confidence we desire, all in a way that is intuitively logical.

8.6 Practical limitations in sampling and testing

Clearly, all the above instructs us that the ideal situation in research is to draw random samples, to ensure that these are big and then we can do some regular standard normal distribution tests, and all will be well. Is that indeed the ideal situation? Yes and no.

Obviously, if one is able to draw a random sample, and thus achieve generalizability, that is better and more desirable than working with non-random samples. However, in many situations, samples simply are non-random, either because there is no sampling frame, or because nonresponse has affected the randomness of the sample, or for some other reason. Often, we simply have to make do with such less than ideal samples. The research situation on the ground simply is not that clear cut and well-organized. The art of research then can no longer be to arrive at a technically perfect situation, but to operate sensibly and pragmatically from a technically suboptimal or defunct situation.

Secondly, large samples – seemingly paradoxically – have their disadvantages too. If I draw a sample of 180,000 persons out of the Dutch population, an approximately 1% sample, that sample size is so huge that any test I perform will turn out to be significant. The test value $z_{s_{\bar{x}}}$ is defined as

$$\frac{\bar{\mathrm{X}} - \mu}{\sigma_{\bar{\mathrm{X}}}},$$

but given that

$$\sigma_{\bar{\mathrm{X}}} = \frac{\sigma}{\sqrt{\mathrm{N}}}$$

the resulting test value z will with such a large N be divided by a very small $\sigma_{\bar{\mathrm{X}}}$, and thus become huge. As can be seen from all the graphs shown before, more extreme values of the test statistic imply that the null hypothesis will be rejected. Thus, with such a large N, any test will turn out to be significant. This is logical, as with such a huge sample the likelihood that any results are attributable to chance is negligible. The flip side is however that testing becomes in a sense meaningless: tests will always be significant. Very tiny differences will be marked as 'significant': a correlation coefficient of 0.01 (that is: a situation where there is practically no association between two variables) will also be defined as not attributable to chance, i.e. significantly different from zero. In these situations, while there is statistical significance, there is no *practical significance*. In that sense, samples can definitely become too big.

So here too there is a trade-off. While a larger sample absolutely does give you a better idea of what is going on in the population, it makes judging how meaningful findings are harder. Small deviations from the situation under the H_0 are marked as 'significant' when they may be too tiny actually to warrant mentioning.

For that reason, the α is often set proportionate to the sample size. A regular α, for ordinary samples of, say, between 100 and 500, is generally set at 5%. Once sample sizes become really big, researchers may revert to a smaller α, say 2.5% or 1%. Conversely, if one has to work with very small samples of well under 100, researchers sometimes relax the α, and may accept 10% as a significance level. It is then the custom, however, not to speak of significance but of a 'trend'.

A second instance when there is reason to use a stricter α is when one wants or needs to perform many tests. If α is 5%, what this means is that out of 100 tests on the same phenomenon, one would expect 5% to show up significant – just accidentally, that is what the reasoning says. So if one performs 100 tests, reporting a number of these as significant doesn't really prove a lot. This is referred to as *chance capitalization*. For that reason, if one performs a large number of tests on phenomena that may be not exactly identical but nevertheless capture the same phenomenon or tap into the same association, it is customary to balance this by also reducing the α. In some contexts, this is referred to as a *Bonferroni correction*.

8.7 Wrapping up

This chapter introduced some technical foundations of statistical testing. We started out by showing how statistical reasoning is in fact quite similar to our day-to-day reasoning. Using the analogy of a fire-alarm, we illustrated how in deriving statements about reality (a population) on the basis of incomplete information (a sample), we may make mistakes: we may think that a fire is burning when it is not, or may miss a fire

that is burning. Statistics help us to deal with, and quantify, the uncertainty that is inherent to sampling. We showed how the chances of mistakes (rejecting H_0 when H_0 is true or, conversely, not accepting H_1 when H_1 is true) are connected, and that only by increasing sample size do we reduce the risk of both. Hypothesis testing and one-sided and two-sided hypotheses were introduced. We briefly outlined the central limit theorem that underscores the importance of random sampling introduced earlier in chapter 3. We discussed confidence intervals as a way to express uncertainty. We ended by contextualizing the former in the sense that we discussed how in practice it is often impossible to achieve truly random sampling and how very large samples may also pose some challenges.

Chapter questions

1. What do we mean when we say that a finding is 'significant'? (section 8.1.2)
2. Describe the null and alternative hypothesis for a study in which we attempt to find out whether mediation settles divorce procedures faster than regular procedures through family court (section 8.1.2)
3. Describe the quadrants in Figure 8.1 for a COVID test: give labels to the rows and columns. Argue what you would like α and β to be (section 8.1.2)
4. Explain why α and β are 'communicating vessels' (section 8.1.2 and section 8.3)
5. What is meant by the power of a statistical test? (section 8.3)
6. What are two ways to increase the power of a test? (section 8.3)
7. Can samples become too large? (section 8.6)
8. What does a confidence interval indicate? (section 8.5.1)