13.1: The one-sample z-test

In this section I'll describe one of the most useless tests in all of statistics: the ***z-test***. Seriously – this test is almost never used in real life. Its only real purpose is that, when teaching statistics, it's a very convenient stepping stone along the way towards the t-test, which is probably the most (over)used tool in all statistics.

13.1.1 inference problem that the test addresses

To introduce the idea behind the z-test, let's use a simple example. A friend of mine, Dr Zeppo, grades his introductory statistics class on a curve. Let's suppose that the average grade in his class is 67.5, and the standard deviation is 9.5. Of his many hundreds of students, it turns out that 20 of them also take psychology classes. Out of curiosity, I find myself wondering: do the psychology students tend to get the same grades as everyone else (i.e., mean 67.5) or do they tend to score higher or lower? He emails me the `zeppo.Rdata` file, which I use to pull up the `grades` of those students,

```
load( "./rbook-master/data/zeppo.Rdata" )
print( grades )
```

```
##  [1] 50 60 60 64 66 66 67 69 70 74 76 76 77 79 79 79 81 82 82 89
```

and calculate the mean:

```
mean( grades )
```

```
## [1] 72.3
```

Hm. It *might* be that the psychology students are scoring a bit higher than normal: that sample mean of $\bar{X}$ = 72.3 is a fair bit higher than the hypothesised population mean of μ=67.5, but on the other hand, a sample size of N=20 isn't all that big. Maybe it's pure chance.

To answer the question, it helps to be able to write down what it is that I think I know. Firstly, I know that the sample mean is $\bar{X}$ =72.3. If I'm willing to assume that the psychology students have the same standard deviation as the rest of the class then I can say that the population standard deviation is σ=9.5. I'll also assume that since Dr Zeppo is grading to a curve, the psychology student grades are normally distributed.

Next, it helps to be clear about what I want to learn from the data. In this case, my research hypothesis relates to the *population* mean μ for the psychology student grades, which is unknown. Specifically, I want to know if μ=67.5 or not. Given that this is what I know, can we devise a hypothesis test to solve our problem? The data, along with the hypothesised distribution from which they are thought to arise, are shown in Figure 13.1. Not entirely obvious what the right answer is, is it? For this, we are going to need some statistics.

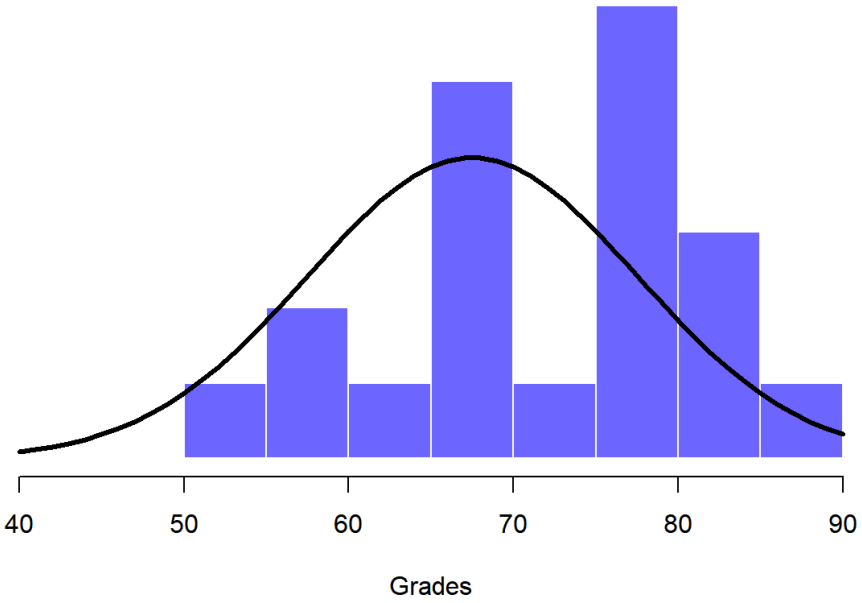

Figure 13.1: The theoretical distribution (solid line) from which the psychology student grades (grey bars) are supposed to have been generated.

13.1.2 Constructing the hypothesis test

The first step in constructing a hypothesis test is to be clear about what the null and alternative hypotheses are. This isn't too hard to do. Our null hypothesis, H_0, is that the true population mean μ for psychology student grades is 67.5%; and our alternative hypothesis is that the population mean *isn't* 67.5%. If we write this in mathematical notation, these hypotheses become,

$$H_0: \mu = 67.5$$

$$H_1: \mu \neq 67.5$$

though to be honest this notation doesn't add much to our understanding of the problem, it's just a compact way of writing down what we're trying to learn from the data. The null hypotheses H_0 and the alternative hypothesis H_1 for our test are both illustrated in Figure 13.2. In addition to providing us with these hypotheses, the scenario outlined above provides us with a fair amount of background knowledge that might be useful. Specifically, there are two special pieces of information that we can add:

1 The psychology grades are normally distributed. 1 The true standard deviation of these scores σ is known to be 9.5.

For the moment, we'll act as if these are absolutely trustworthy facts. In real life, this kind of absolutely trustworthy background knowledge doesn't exist, and so if we want to rely on these facts we'll just have make the *assumption* that these things are true. However, since these assumptions may or may not be warranted, we might need to check them. For now though, we'll keep things simple.

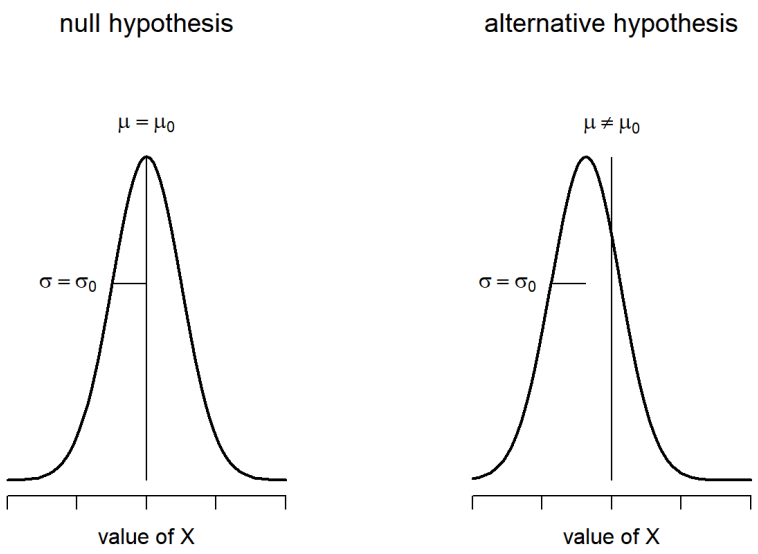

Figure 13.2: Graphical illustration of the null and alternative hypotheses assumed by the one sample z-test (the two sided version, that is). The null and alternative hypotheses both assume that the population distribution is normal, and additionally assumes that the population standard deviation is known (fixed at some value σ_0). The null hypothesis (left) is that the population mean μ is equal to some specified value μ_0. The alternative hypothesis is that the population mean differs from this value, $\mu \neq \mu_0$.

The next step is to figure out what we would be a good choice for a diagnostic test statistic; something that would help us discriminate between H_0 and H_1. Given that the hypotheses all refer to the population mean μ, you'd feel pretty confident that the sample mean $\bar{X}$ would be a pretty useful place to start. What we could do, is look at the difference between the sample mean $\bar{X}$ and the value that the null hypothesis predicts for the population mean. In our example, that would mean we calculate $\bar{X}$ - 67.5. More generally, if we let μ_0 refer to the value that the null hypothesis claims is our population mean, then we'd want to calculate

$$\bar{X} - \mu_0$$

If this quantity equals or is very close to 0, things are looking good for the null hypothesis. If this quantity is a long way away from 0, then it's looking less likely that the null hypothesis is worth retaining. But how far away from zero should it be for us to reject H_0?

To figure that out, we need to be a bit more sneaky, and we'll need to rely on those two pieces of background knowledge that I wrote down previously, namely that the raw data are normally distributed, and we know the value of the population standard deviation σ. If the null hypothesis is actually true, and the true mean is μ_0, then these facts together mean that we know the complete population distribution of the data: a normal distribution with mean μ_0 and standard deviation σ. Adopting the notation from Section 9.5, a statistician might write this as:

$$X \sim \text{Normal}(\mu_0, \sigma^2)$$

Okay, if that's true, then what can we say about the distribution of $\bar{X}$? Well, as we discussed earlier (see Section 10.3.3), the sampling distribution of the mean $\bar{X}$ is also normal, and has mean μ. But the standard deviation of this sampling distribution SE ($\bar{X}$), which is called the *standard error of the mean*, is

$$\text{SE}(\bar{X}) = \frac{\sigma}{\sqrt{N}}$$

In other words, if the null hypothesis is true then the sampling distribution of the mean can be written as follows:

$$\bar{X} \sim \text{Normal}(\mu_0, \text{SE}(\bar{X}))$$

Now comes the trick. What we can do is convert the sample mean $\bar{X}$ into a standard score (Section 5.6). This is conventionally written as z, but for now I'm going to refer to it as $z_{\bar{X}}$. (The reason for using this expanded notation is to help you remember that we're calculating standardised version of a sample mean, *not* a standardised version of a single observation, which is what a z-score usually refers to). When we do so, the z-score for our sample mean is

$$z_{\bar{X}} = \frac{\bar{X} - \mu_0}{SE(\bar{X})}$$

or, equivalently

$$z_{\bar{X}} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{N}}$$

This z-score is our test statistic. The nice thing about using this as our test statistic is that like all z-scores, it has a standard normal distribution:

$$z_{\bar{X}} \sim \text{Normal}(0,1)$$

(again, see Section 5.6 if you've forgotten why this is true). In other words, regardless of what scale the original data are on, the z-statistic iteself always has the same interpretation: it's equal to the number of standard errors that separate the observed sample mean $\bar{X}$ from the population mean μ_0 predicted by the null hypothesis. Better yet, regardless of what the population parameters for the raw scores actually are, the 5% critical regions for z-test are always the same, as illustrated in Figures 13.4 and 13.3. And what this meant, way back in the days where people did all their statistics by hand, is that someone could publish a table like this:

desired α level	two-sided test	one-sided test
.1	1.644854	1.281552
.05	1.959964	1.644854
.01	2.575829	2.326348
.001	3.290527	3.090232

which in turn meant that researchers could calculate their z-statistic by hand, and then look up the critical value in a text book. That was an incredibly handy thing to be able to do back then, but it's kind of unnecessary these days, since it's trivially easy to do it with software like R.

Figure 13.3: Rejection regions for the two-sided z-test

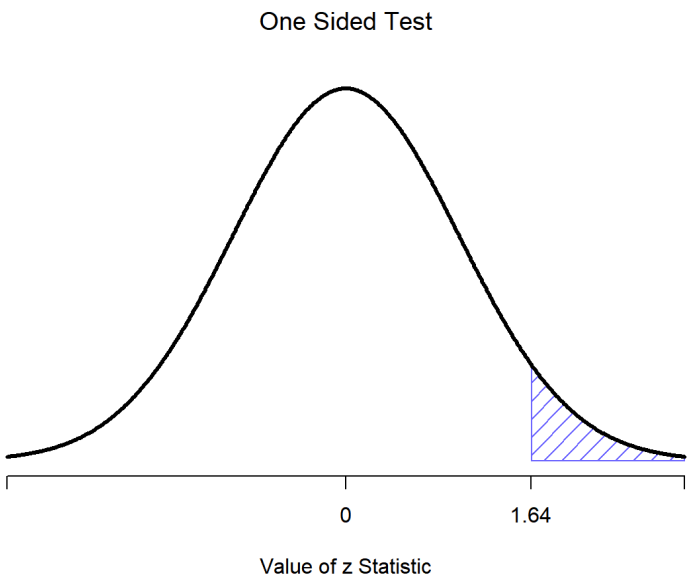

Figure 13.4: Rejection regions for the one-sided z-test

13.1.3 worked example using R

Now, as I mentioned earlier, the z-test is almost never used in practice. It's so rarely used in real life that the basic installation of R doesn't have a built in function for it. However, the test is so incredibly simple that it's really easy to do one manually. Let's go back to the data from Dr Zeppo's class. Having loaded the `grades` data, the first thing I need to do is calculate the sample mean:

```
sample.mean <- mean( grades )
print( sample.mean )
```

```
## [1] 72.3
```

Then, I create variables corresponding to known population standard deviation (σ=9.5), and the value of the population mean that the null hypothesis specifies (μ_0=67.5):

```
mu.null <- 67.5
sd.true <- 9.5
```

Let's also create a variable for the sample size. We could count up the number of observations ourselves, and type `N <- 20` at the command prompt, but counting is tedious and repetitive. Let's get R to do the tedious repetitive bit by using the `length()` function, which tells us how many elements there are in a vector:

```
N <- length( grades )
print( N )
```

```
## [1] 20
```

Next, let's calculate the (true) standard error of the mean:

```
sem.true <- sd.true / sqrt(N)
print(sem.true)
```

```
## [1] 2.124265
```

And finally, we calculate our z-score:

```
z.score <- (sample.mean - mu.null) / sem.true
print( z.score )
```

```
## [1] 2.259606
```

At this point, we would traditionally look up the value 2.26 in our table of critical values. Our original hypothesis was two-sided (we didn't really have any theory about whether psych students would be better or worse at statistics than other students) so our hypothesis test is two-sided (or two-tailed) also. Looking at the little table that I showed earlier, we can see that 2.26 is bigger than the critical value of 1.96 that would be required to be significant at α=.05, but smaller than the value of 2.58 that would be required to be significant at a level of α=.01. Therefore, we can conclude that we have a significant effect, which we might write up by saying something like this:

With a mean grade of 73.2 in the sample of psychology students, and assuming a true population standard deviation of 9.5, we can conclude that the psychology students have significantly different statistics scores to the class average (z=2.26, N=20, p<.05).

However, what if want an exact p-value? Well, back in the day, the tables of critical values were huge, and so you could look up your actual z-value, and find the smallest value of α for which your data would be significant (which, as discussed earlier, is the very definition of a p-value). However, looking things up in books is tedious, and typing things into computers is awesome. So let's do it using R instead. Now, notice that the α level of a z-test (or any other test, for that matter) defines the total area "under the curve" for the critical region, right? That is, if we set α=.05 for a two-sided test, then the critical region is set up such that the area under the curve for the critical region is .05. And, for the z-test, the critical value of 1.96 is chosen that way because the area in the lower tail (i.e., below −1.96) is exactly .025 and the area under the upper tail (i.e., above 1.96) is exactly .025. So, since our observed z-statistic is 2.26, why not calculate the area under the curve below −2.26 or above 2.26? In R we can calculate this using the `pnorm()` function. For the upper tail:

```
upper.area <- pnorm( q = z.score, lower.tail = FALSE )
print( upper.area )
```

```
## [1] 0.01192287
```

The `lower.tail = FALSE` is me telling R to calculate the area under the curve from 2.26 *and upwards*. If I'd told it that `lower.tail = TRUE`, then R would calculate the area from 2.26 *and below*, and it would give me an answer 0.9880771. Alternatively, to calculate the area from −2.26 and below, we get

```
lower.area <- pnorm( q = -z.score, lower.tail = TRUE )
print( lower.area )
```

```
## [1] 0.01192287
```

Thus we get our p-value:

```
p.value <- lower.area + upper.area
print( p.value )
```

```
## [1] 0.02384574
```

13.1.4 Assumptions of the z-test

As I've said before, all statistical tests make assumptions. Some tests make reasonable assumptions, while other tests do not. The test I've just described – the one sample z-test – makes three basic assumptions. These are:

- *Normality*. As usually described, the z-test assumes that the true population distribution is normal.[186] is often pretty reasonable, and not only that, it's an assumption that we can check if we feel worried about it (see Section 13.9).
- *Independence*. The second assumption of the test is that the observations in your data set are not correlated with each other, or related to each other in some funny way. This isn't as easy to check statistically: it relies a bit on good experimetal design. An obvious (and stupid) example of something that violates this assumption is a data set where you "copy" the same observation over and over again in your data file: so you end up with a massive "sample size", consisting of only one genuine observation. More realistically, you have to ask yourself if it's really plausible to imagine that each observation is a completely random sample from the population that you're interested in. In practice, this assumption is never met; but we try our best to design studies that minimise the problems of correlated data.
- *Known standard deviation*. The third assumption of the z-test is that the true standard deviation of the population is known to the researcher. This is just stupid. In no real world data analysis problem do you know the standard deviation σ of some population, but are completely ignorant about the mean μ. In other words, this assumption is *always* wrong.

In view of the stupidity of assuming that σ is known, let's see if we can live without it. This takes us out of the dreary domain of the z-test, and into the magical kingdom of the t-test, with unicorns and fairies and leprechauns, and um…
