A walkthrough guide for calculating t-statistics and p-values in a one-sample t-test

Overview

In week 3 class we encountered t-statistics and p-values. If you aren't familiar with using statistics then this can be one of the toughest weeks. When you first encounter them, t-statistics and p-values are pretty counter-intuitive. As such, I've written this short walkthrough guide which should help you to understand (i) how we calculate these values and (ii) why we use them. A word of warning – this walkthrough is pretty lengthy, and you'll probably need to invest some time in reading it. But, it really should help you to understand t-statistics and p-values (and who can put a price on knowledge?).

Key concepts covered: one-sample t-test, hypothesis testing, samples and sampling distribution, standard error of the mean (SEM), t-statistics, p-values and alpha.

Contents

Authors note: thanks to Jonathan Bright and Scott Hale for feedback and edits to this document.

Walkthrough

Hypothesis Specification

It is standard practice in both the social and natural sciences to use hypotheses to guide data collection and analysis. Using hypotheses gives our findings rigor and credibility: rather than just randomly looking for patterns in the data or cherry-picking results which we think might be noteworthy, we use theory to generate hypotheses which we then test. Typically, we want to understand the relationship between two 'measured phenomena' – that is, we want to know whether there is a relationship between two variables. Two hypotheses are always stipulated: (i) the null hypothesis and (ii) the alternative hypothesis.

(i) The null hypothesis always stays the same. It states that there is no relationship between two measured phenomena.

In the example from class we looked at whether the the income of supporters of a political party ('Party A') differs from the income of the population. In this case the 'measured phenomena' are income and affiliation with Party A. The income of the population was reported as £27,500. The null hypothesis is therefore: 'there is no relationship between income and Party A affiliation. The average income of Party A supporters is £27,500.' This means that we do not expect the income of Party A supporters to differ from the income of the overall population.

(ii) The alternative hypothesis states that there is a relationship between two measured phenomena. The alternative hypothesis is therefore: 'there is a relationship between income and Party A affiliation. The average income of Party A supporters differs from £27,500.'

The way that we have expressed the alternative and null hypotheses means that they are Mutually Exclusive and Collectively Exhaustive (MECE). This means that (i) the two hypotheses do not overlap (i.e., they are mutually exclusive; so, only one of the two can be true) and (ii) the two hypotheses cover every available option (i.e., they are collectively exhaustive; so, one of the two must be true).

A quick note - in this example we want to know whether a specific feature (income) of a particular sub-group (Party A supporters) differs from the same feature for the overall population. As such we use a one-sample test. In week 3, the test we looked at was the one sample "z test", though we didn't actually use this name. In week 4, we looked at the t test for two-samples. Because the one-sample t test is far more common than the one-sample z-test in practice, **the rest of this guide looks solely at the t test rather than the z test**. If we wanted to compare two samples (e.g. Party A supporters with Party B supporters) then we would use a two-sample t-test – which is what we did in week 4.

Samples and Significance Testing

We want to find something out about all Party A supporters. Here, we come across our first roadblock. Usually, collecting data on all members of a group is not possible: it is just too difficult/expensive/time-consuming. Instead, we take a random sample of Party A supporters. By deploying inferential statistics we can then use this sample to infer things about the population of Party A supporters. It is worth noting that if we had data on the whole population (i.e. if we took a census) then we wouldn't need to use inferential statistics, and as such wouldn't be worrying about significance testing.

So, we take a random sample of 100 Party A supporters (the size of the sample is typically denoted by 'n='). We calculate the mean income of the sample (in this case, £28,000), and compare this with the mean income of the whole population (£27,500). In this case, we observe a difference between the means of £500 (hurray!). But, we have a problem: because we took a sample we can't be sure that we are observing a real difference. What we need to know is whether we can trust the difference of £500 we have observed to infer something about Party A supporters in general.

The way that we find this out might seem, at first, a little bit odd. We work out how likely it is that we would observe a sample with a mean that differs by this much or more from the population mean *if the null hypothesis were true*. Accordingly, the question that we ask is: if there were no relationship between Income and Party A affiliation how likely is it that we observe a sample with a mean that differs from the population mean by £500 or more? The reason why this is the right question to ask might not be immediately obvious, but it really is useful information. The intuition behind it is simple: if randomly sampling a mean that differs from the population mean by £500 is actually quite likely then we should be less confident about saying that it represents a real difference between Party A supporters and the general population. To understand why this is the case we now need to think about another concept that we covered in class: the sampling distribution.

The Sampling Distribution

The sampling distribution can be defined as 'the distribution of means of an infinite series of samples drawn from the population'.

If you wanted to generate a sampling distribution you would follow these steps: (1) Decide the size of the sample by setting n. (2) Randomly select a sample from the population. (3) Calculate the mean of the sample. (4) Record this sample mean. (5) Keep doing this for an infinite number of samples (6) Plot all of the recorded sample means. You now have a sampling distribution (this is what we were able to do with the simulator discussed in class).

The sampling distribution has two really useful features. First, the mean of the sampling distribution equals the mean of the population. Second, the sampling distribution has a normal distribution, even when the distribution of the underlying population is skewed (if the sample size is big enough). This second point might seem particularly counter-intuitive, but it is crucial to inferential statistics. It is described formally though the ‘Central Limit Theorem’.

A useful way of thinking about sampling is that we actually select the sample from the sampling distribution, and not from the population. This makes sense – once we decided the sample size, we randomly selected one particular sample from the huge number of different samples (with that n) that we could have been picked. If we approach the issue like this (and bearing in mind the second feature, that the mean of the sampling distribution equals the mean of the population) then we can ask a subtly different question to the one we posed above. Above, we asked: how likely is it that we observe a sample with a mean that differs by £500 from the population mean if the null hypothesis were true? We can now ask: how likely is it that we observe a sample with a mean that differs by at least £500 or more from the *sampling distribution mean* if the null hypothesis were true?

The t-statistic

In effect, what we want to know is how far our randomly selected sample lies from the mean of the sampling distribution. Remember feature 2 of the sampling distribution: the sampling distribution is normally distributed, even if the distribution of the underlying population is skewed. So, because of this, we can express the difference between the sample mean and the population mean in terms of the standard deviation of the sampling distribution (hereafter known as sampling distribution s.d.).

This is really useful because the normal distribution has some key properties which we can use to quantify how likely it is we would observe values which are a certain number of standard deviations away from the mean. So, if the sample mean is one standard deviation away from the sampling distribution mean then we know this is not that rare – 68% of values will be within approximately one standard deviation. The further that the sample mean is away from the sampling distribution mean (measured in standard deviations) then, mathematically, the less and less likely it is that it will occur: 95% of values will be within approximately 2 standard deviations and 99% of values within 3 standard deviations… So, we can use the sampling distribution s.d. to get a sense of how likely it is that we would randomly select the sample that we observed. This is what the t-statistic expresses.

Figure 1 shows the formula for the t-statistic. The numerator is the sample mean less the population mean (in this case £500). The denominator is the standard deviation of the sampling distribution.

$$t - statistic = \frac{Sample\ mean - Sampling\ distribution\ mean}{Sampling\ distribution\ standard\ deviation}$$

Figure 1: t-statistic

So far, so good: t-statistics are easy! But, alas, at this stage we have a problem. Unfortunately, we cannot directly access the sampling distribution and so are unable to directly calculate the sampling distribution s.d.. In the formula above, we are missing the denominator! Looks like we won't be able to calculate the answer (so much for t-tests…). But, handily, the wonderful world of statistics provides us with a solution: the Standard Error of the Mean (SEM).

SEM is an estimator of the standard deviation of the sampling distribution. It uses the very limited information that we have available in the sample to estimate the characteristics of the sampling distribution. The formula below shows how SEM is calculated. The numerator is the standard deviation of the sample (in our case, the lecture slides reported this as £5,000). The denominator is the square root of the sample size (in our case, the sample size is 100, so the square root is 10). The SEM is therefore: £5,000 / 10 = £500.

$$Standard\ Error\ of\ the\ Mean = \frac{Sample\ Standard\ Deviation}{\sqrt{Sample\ Size}}$$

Figure 2: Standard Error of the Mean (SEM)

Now that we are using an estimate of the sampling distribution we need to make one more allowance. When we use SEM we can no longer assume that the sampling distribution is normally distributed. Instead, we have to use the *t*-distribution (this is actually where the name of the t-statistic comes from). Compared to the normal distribution the *t*-distribution has a lower peak than the normal distribution and thicker tails (we describe this as the distribution being 'leptokurtic'). This means that it is more likely that we will observe values that are further away from the sampling distribution mean. Figure 1 below is taken from the lecture slides. It shows a comparison of t distributions with different parameters, and the normal distribution (the x-axis shows standard deviations). The line with the highest peak is the normal distribution.

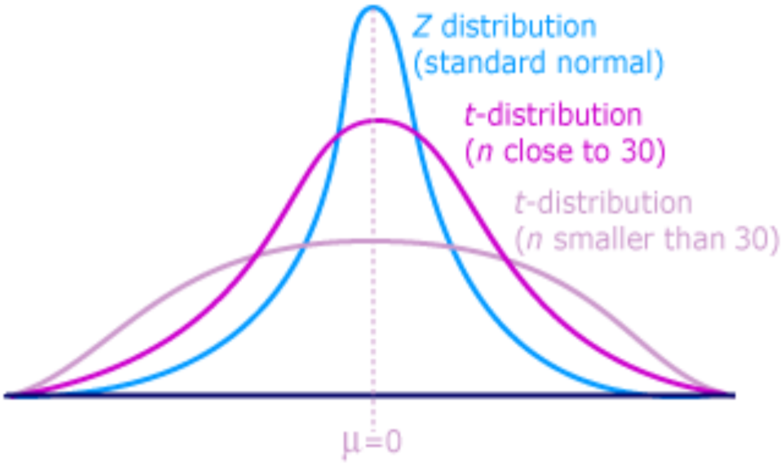

Figure 3: t-distributions and the normal distribution

The larger the sample size the more that the sampling distribution resembles a normal distribution. Once the sample size hits about 30 then the *t*-distribution is a very close approximation of the normal distribution. This makes sense – the larger the size of our sample the more that we can trust it when estimating the standard deviation of the sampling distribution. And, luckily for us, we don't ever have to work any of this out by hand – all statistical software, including R, will calculate the correct *t*-distribution for us.

So, let's go back to the t-statistic formula we encountered above and readjust it to include the SEM. The only difference between figure 4 (below) and figure 1 is the denominator, which has changed from the sampling distribution s.d. to the SEM.

$$t - statistic = \frac{Sample\ mean - Sampling\ distribution\ mean}{Standard\ Error\ of\ the\ Mean}$$

Figure 4: t-statistic with SEM

In our example, the difference between sampling distribution mean and sample mean is £500 and the SEM is £500, so the t-statistic Is 1 (£500/£500 = 1). We can interpret this as saying that the mean of the sample we observed lies 1 standard deviation away from the mean of the sampling distribution.

Let's just take a moment to look at the maths behind the t-statistic in more detail. The denominator of the t-statistic is the sampling distribution s.d.. *Ceteris paribus* the smaller the denominator, the larger the t-statistic is (i.e. if the size of the difference between sample mean and population mean is 10 then the t-statistic is higher when the sampling distribution s.d. is 5 (10/5 = 2) than when the SEM is 20 (10/20 = 0.5)). The intuition behind this is that when the sampling distribution is more spread out (as measured by a higher standard deviation) then we are more likely to observe extreme values. So, if we observe a difference that looks large (e.g. 10) but the standard deviation of the sampling distribution is also large (e.g. 20) then the result is not all that unexpected: it is actually pretty likely that we would randomly select a sample which has a mean that differs by this much from the population mean.

The key takeaway point here is that a large t-statistic means that the sample we have selected lies a long way from the mean of the sampling distribution, and as such is relatively unlikely to have been observed.

P-values

The t-statistic expresses how far the mean of the sample we observed lies from the mean of the sampling distribution. This is really useful information. But, it is not always easy to interpret - we know that a sample mean which lies 1 SEM away from the sampling distribution mean is pretty likely, but just how likely is it? What we want is an easy-to-interpret single value that expresses how confident we are about the data. This is where the p-value comes in.

The p-value expresses the probability of observing a sample with a mean as extreme or more extreme than the one observed given that the null hypothesis is true. The p-value is calculated directly from the t-statistic: *ceteris paribus* a high t-statistic will produce a small p-value. P-values are really easy to interpret. For instance, if we calculate a p-value of 0.1 this means that there is a 0.1 probability that we would observe a sample with a mean at least this extreme given that the null hypothesis is true.

P-values are particularly useful because they can be directly compared across different studies and even different statistical tests. This is one of the reasons why p-values are the most well-used statistical concept in the social sciences. As such, it is really worth getting your head round what the p-value expresses. If you want to find out more about p-values then read this paper: http://journal.frontiersin.org/article/10.3389/fphy.2016.00006/full.

Returning to our Two Hypotheses

The whole way through this exercise we have been in the 'world of the null hypothesis'. That is, we have worked from the assumption that the null hypothesis is true, and as such that the sample we observed was taken from the sampling distribution of the general population. We have been trying to work out how likely it is that we would observe a sample mean as extreme or more than this one *given that the null hypothesis is true*.

We can now work out whether we should be in the world of the null hypothesis or the world of the alternative hypothesis. This is where inferential statistics get interesting and, for the purposes of theory-driven social scientific research, really useful. Let's return to our starting two hypotheses: (i) the null hypothesis - that there is no relationship between Income and Party A affiliation, and (ii) the alternative hypothesis - that there is a relationship between income and Party A affiliation. Is the data in our sample more consistent with the alternative hypothesis or the null hypothesis?

In our example, the t-statistic of 1 produces a p-value of 0.32. This means there is a 0.32 chance we would observe a sample mean at least this extreme given that the null hypothesis is true. So, based on this p-value, should we accept or reject the null hypothesis?

To interpret the p-value we compare it against a pre-chosen value of 'alpha'. Alpha is usually set at 0.05 (though values of 0.01 and 0.001 are also common). Only if the p-value is less than alpha are we able to reject the null hypothesis (and, unfortunately, here 0.32 > 0.05 so we can't). Note that this doesn't mean the null hypothesis is true. It is simply means we have not observed data that greatly contradicts it. Showing you one yellow duck does not prove that all ducks are yellow.

If we report a p-value below alpha (say, 0.03) then this gives us grounds to reject the null hypothesis. What we are effectively saying is that there is such a small chance that this sample was randomly selected from the overall population that we think it probably did *not* actually come from the population. As such, we tentatively accept the alternative hypothesis: that there is a relationship between the sub group we are interested in (in this case, Party A supporters) and the measurement variable we are interested in (in this case, income). This is an inferential 'leap', which *might always be incorrect*. We can never be totally sure that the effect we have reported genuinely exists.

This inferential 'leap' can seem a bit odd. But it is the basis of all frequentist statistics: when the p-value is less than alpha (*if the null hypothesis were true*) we say that the probability of observing a sample with this mean is so low that we are going to reject the null hypothesis and tentatively accept the alternative hypothesis instead.

Concluding Remarks

Well done for making it to the end of this walkthrough! I know it's been quite a long read, but you should now have a good idea of why we do significance testing, how the t-statistic is calculated, why we use p-values and how to interpret them. If you have any questions about what has been written here (or if you identify any errors), then please email me at bertievidgen@gmail.com – and feel free to share this document with anyone who you think might find it useful.