Null Hypothesis Significance Testing

If you'd like to find out more about p-values, please read this review piece myself and Prof Yasseri wrote a while ago:

> Vidgen and Yasseri 2016), 'p-values: misunderstood and misused', *Frontiers in Physics*.
>
> Available at: https://www.frontiersin.org/articles/10.3389/fphy.2016.00006/full

Sources

Excellent sources for any issues you might have:

- Minitab
- Stattrek.com
- Stack overflow
- Statistcshowto
- Simple.wikipedia (simple is great because everything is written out in intuitive, straightforward English with very little jargon)

1 | NHST overview

Hypothesis testing

Across both social and natural sciences, null hypothesis testing is the basis of most scientific inquiry. We can use it to ask many different types of questions – but whatever questions we are asking, there are two things we have to specify:

1. The null hypothesis. This states that there is no difference between the groups we are studying. So, if we are studying how many hours of games are played each week by users of different consoles, the null hypothesis states: *there is no difference in how many hours of games are played by users of different consoles*.
2. The alternative hypothesis. This states that there is a difference between two groups. So, for our example: *there is a difference in how many hours of games are played by users of different consoles*.

Populations and samples

Ideally, we would test our hypotheses on the entire population – but that is usually totally infeasible. For instance, if we wanted to study whether users of Twitter are happier than users of Facebook we would be looking at populations which consist of billions of users. So, we take a random sample of users instead – in this example, one sample from users of Facebook and one sample from users of Twitter. This is a good way of getting round the fact that studying entire populations is expensive, time-consuming and difficult. But it raises a pretty obvious problem: how can we be sure that the difference we observe between the samples *actually exists in the population*? This is the basic problem that inferential statistics is trying to solve. We want to know whether we can *trust* any difference we have observed.

One-sample vs. two-sample tests

The first thing to work out is whether you are conducting a one-sample or two-sample test as these two tests produce fundamentally different types of information. Often, we just want to compare a specific group of users with the population in general. For instance, we could find out: do Conservative party members have higher incomes than the UK population? We call this a one-sample test because we only draw one sample. But in other cases, we want to compare two groups – so we might ask: do Conservative party members have higher incomes than Labour party members? We call this a two-sample test because we draw two samples

(one for Conservatives and one for Labour). Most of the time, in social science we conduct two-sample tests.

One-tailed vs. two-tailed tests

Whether the test has one or two 'tails' (sometimes also called 'one-sided' or 'two-sided') is really important for how we assess significance. The key difference between one- and two-tailed tests is: are we testing for a difference between means in *just one direction* (i.e. X is greater than Y) or are we testing for both directions (i.e. X can be greater than Y and Y can be greater than X). If we are testing for just one direction that means we have a strong ingoing assumption about the difference between the two groups So, if we really strongly think that Facebook users are more likely to be happy than Twitter users – and have good previous research/theoretical justifications to back it up – we could use a one-tailed test. The risk with this is that if our ingoing assumption is wrong and Twitter users are happier than Facebook users, we would *not* pick it up.

Usually, we want to look at both directions so we use a two-tailed test. However, because we are now looking at both directions we have to make a correction. Instead of using the value of alpha as it is stated (e.g. 0.05), we have to cut it in half – so we have 0.025 significance threshold at one end of the test statistic distribution and 0.025 significance threshold at the other.

- The 'upper tail' is where the true mean of the first sample is *greater* than either the population or the other sample we are comparing it with.
- The 'lower tail' is where the true mean of the first sample is *lesser* than either the population or the other sample we are comparing it with.

NOTE: annoyingly, the language used in statistics is all quite similar: but whether there are one/two samples or one/two tails *are different things!* The number of samples tells us what sort of comparison we are making (sample vs. sample or sample vs. population), and the number of tails tells us whether we are testing for both directions of the relationship or just one.

2 | Testing significance

Testing for significance is – on paper – pretty straight forward. Here are the steps:

1. Write out the null and alternative hypotheses.
2. Decide the level of alpha (the 'significance threshold') for rejecting the null.
3. Calculate the test statistic.
4. From the test statistic, calculate the p-value.
5. Compare the p-value with alpha.
6. Reject or accept the null hypothesis.

Easy! At least it is once you understand all of the steps and jargon…

Test statistics

The test statistic is a value that we calculate from our sample(s). It is the basis of how we assess whether to accept or reject the null hypothesis. In the OII Statistics class, we deal with two test statistics: the t-test and the z-test. For this guide, we are going to focus on the t-test. That said, note that we can use the z-test when:

- The sample size is large.
- The standard deviation of the population is known.

T-tests

Remember that what we want to know is whether an observed difference between the means of two samples represents an actual difference in the population. To make this assessment, we need to know a couple of things:

1. The size of the difference between the two means (duh).
2. The standard deviation of the sampling distribution (yikes. A bit more complicated).

Sampling distribution

The sampling distribution is the distribution of means calculated for an infinitely large number of samples taken from a population. So, let's say that we are studying the number of hours spent online by Facebook users. We could take a sample of 100 people, record how many hours they spend online and then calculate the mean. If we kept doing this then we would end up with a big list of means for samples of size 100. Whilst some of the sample means would be extremely high or low, most would cluster around the middle. These means have a distribution which we can plot and analyse – *this is the sampling distribution!* A nice feature of the sampling distribution is that its mean is equal to the mean of the population.

The example we've just given is for a single population (suitable for a one-sample t-test). If you wanted to generate a sampling distribution for a two-sample test you would follow these steps:

1. Decide the size of the samples (e.g. 100 instances)
2. Randomly select one sample from Population 1 and one sample from Population 2.
3. Calculate the means of the two samples.
4. Calculate the difference between the two means and record it.
5. Keep doing this for an infinite number of samples.
6. Plot all of the recorded differences between the two means. Et voila, you now have a sampling distribution for the difference between the means.

The sampling distribution is super useful because it helps us to understand whether the difference that we observed between two samples is likely to occur by chance – which is crucial for answering our starting problem of: *how can we be sure that the difference we observe between the samples actually exists in the population?*

Standard deviation of the sampling distribution

Measuring the standard deviation of the sampling distribution lets us account for the fact that in some sampling distributions it is very likely that we will observe big differences between sample means and in others it is very unlikely (in practice, it just depends upon the nature of the data). To calculate the standard deviation of the sampling distribution of the difference between the two means we use a nice property of the *variance sum law*. This states that:

> *"the variance of the sampling distribution of the difference between means is equal to the variance of the sampling distribution of the mean for Population 1 plus the variance of the sampling distribution of the mean for Population 2."* (http://onlinestatbook.com/2/sampling_distributions/samplingdist_diff_means.html)

What the variance sum law means in practice is that we can just add the variance of the two sampling distributions together to get the variance of the difference between their means. Remember also that standard deviation is just the square root of variance. So we can add the variance of the two sampling distributions together, then square root them to get the standard deviation (see Figure 2 below). Easy.

But, here comes the rub – in nearly all empirical research we can't directly observe the sampling distribution as we are never going to actually go out and take an infinite number of samples from a population! This leaves us in a pickle as without the sampling distribution we can't calculate its standard deviation …Fortunately, instead we can *approximate* it using the standard error (SE).

Standard Error

We calculate the standard error of a sampling distribution for one sample mean using the only information we have available: (1) the standard deviations of the sample and (2) the size of the sample. Standard error of a sampling distribution for one sample mean is given as:

$$Standard\ Error\ (1\ sample) = \sqrt{\frac{Variance\ of\ sample}{Size\ of\ sample}}$$

Figure 1: Standard Error of the sampling distribution for one sample mean

Using the wonders of the variance sum law (as described above) we can now calculate the standard error of the sampling distribution for the difference between *two* means (sorry about the #WordSalad that is going on here). It is given as:

$$Standard\ Error\ (2\ samples) = \sqrt{\frac{Variance\ of\ sample\ 1}{Size\ of\ sample\ 1} + \frac{Variance\ of\ sample\ 2}{Size\ of\ sample\ 2}}$$

Figure 2: Standard Error of the sampling distribution for the difference between two means

When the Standard Error is high it means that we expect to see big variations: because both the two samples are themselves very dispersed it is more likely that we would see a big difference *between* the two samples. Grasping this logic is crucial to understanding how significance testing works.

The intuition behind this is well-illustrated by the diagram below. The difference in means is the same in both cases (zoom in on the diagrams to see the scales!) But, intuitively, we should trust the plot on the right more because the two distributions have better separation the standard error is very low. In contrast, in the plot on the left the values are very *dispersed* and so overlap to a considerable extent (the standard error is higher). Thus, we can have more trust that the observed difference between the two sample means really exists with the right hand plot compared to the left hand one.

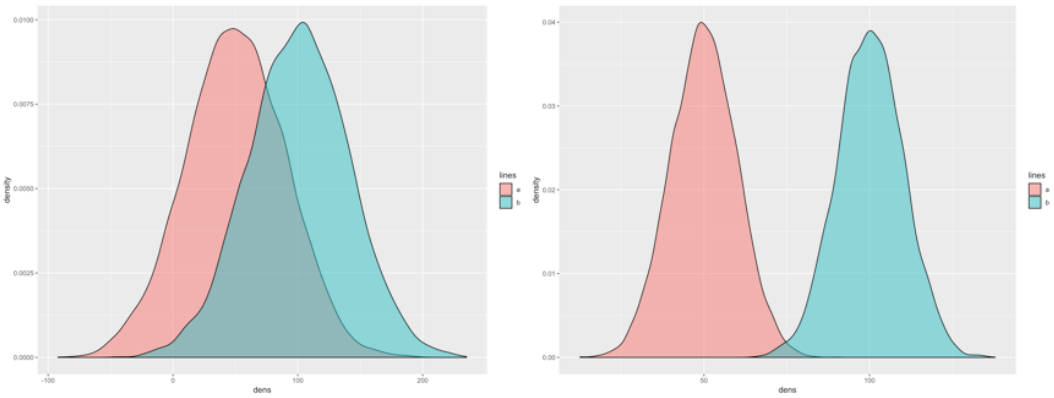

Figure 3: overlapping samples – which should we trust more?

T statistic

Once we have (1) the difference between the two means and (2) the Standard Error we can calculate the t-statistic – we are finally ready to evaluate whether the difference we observed is legit!

The t statistic divides the difference between the two samples by the standard error (in Figure 4, the denominator is 'Standard Error'). This means that we are expressing the difference between the sample means in terms of the number of standard errors. This gives us a nice *comparable* way of talking about the difference between the sample means. A big t-statistic comes from one of two sources: either a very large difference between the two sample means or very small standard error. In either case, bigger t-statistics indicate that the difference we have observed is more likely to exist in the population. It is given as:

$$t - statistic = \frac{Mean\ of\ sample\ 1 - Mean\ of\ sample\ 2}{Standard\ Error\ of\ the\ sampling\ distribution\ of\ the\ two\ means}$$

Figure 4: t-statistic calculation for two-sample t-test

So, we have the t-statistic – are we all done? …Sadly not! On its own the t-statistic is not very meaningful. To make sense of it, we need to convert it into a p-value *or* calculate a critical value.

Evaluating significance

Two options are available to us for evaluating significance. We can either (1) take the test statistic and convert it into a p-value, which we then compare with alpha (as outlined at the start of this section) or (2) take the value of alpha, calculate a critical value and compare it with the test statistic. Either way, we are making a comparison between the test statistic and our alpha level.

Alpha

Alpha is the significance threshold. The lower the value of alpha (typically, we set it to 0.05, 0.01 or 0.001), the more *stringent* and *rigorous* our criteria is and the more that we can trust our result. If we set alpha to 0.05 we are saying that we only have grounds to reject the null hypothesis if the probability of observing a test statistic *this extreme or more given that the null hypothesis is true* is equal to or less than 1 in 20. Now, that sounds complicated but it just means that we only reject the null if it would be super unlikely for us to see this result if there was no actual difference between the two groups we are studying. This is an odd style of reasoning, so don't worry if it at first seems counter-intuitive. The basic logic is: Ask, "if the null hypothesis were true, how likely is it that I would see this result?" If the answer is, 'Very unlikely, the probability is less than 0.05!' then you have grounds to reject the null and accept the alternative (i.e. you accept that there is a difference between the two groups).

Think about the impact of lowering alpha – what does it mean for Type I and Type II errors if we reduce alpha from 0.05 to 0.01? Well, lowering alpha means we only accept results which are even less likely to have occurred if the null is true. All other things being equal, this

means that we are reducing the Type I error (because we are setting a higher bar for rejecting the null) but are also increasing the Type II error.

- For a one-tailed test, we just use the value of alpha as it is stated.
- For a two-tailed test, we cut the value of alpha in half, so if alpha equals 0.05, then we actually use 0.025 for each end of the test distribution.

P-value

The easiest way of measuring significance – and the one which you will do most often when conducting your own research and evaluating other people's – is to take the test statistic (here, t-statistic) and convert it into the probability of observing a result this extreme or more (the p-value). Statistically, we do this using the t-distribution (see below). In practice, it can be done using either R or lookup tables in the back of a statistics textbook.

Then, compare the p-value with alpha and decide whether to reject or accept the null. Low p-values mean that the result is very unlikely and give us grounds to reject the null hypothesis.

Definition: The p-value is the probability of observing a test statistic this extreme or more extreme given that the null hypothesis is true.

The t-distribution

The t-distribution is a special type of symmetric distribution, the shape of which is determined by its *degrees of freedom* – a parameter which is directly related to the sample size (see below). This means that as the size of the sample increases, the shape of the distribution changes. In fact, as the sample size gets bigger the distribution becomes a closer approximation of the normal distribution, as illustrated by the diagram below. The key point about the t-distribution is that it accounts statistically for a fairly basic intuition: namely, that smaller samples are less trustworthy. When the t-distribution is based off a small sample size (e.g. just 10 or 15 values) then the tails are heavier – which means that we are more likely to observe means which fall far from the centre (i.e. we expect more variation).

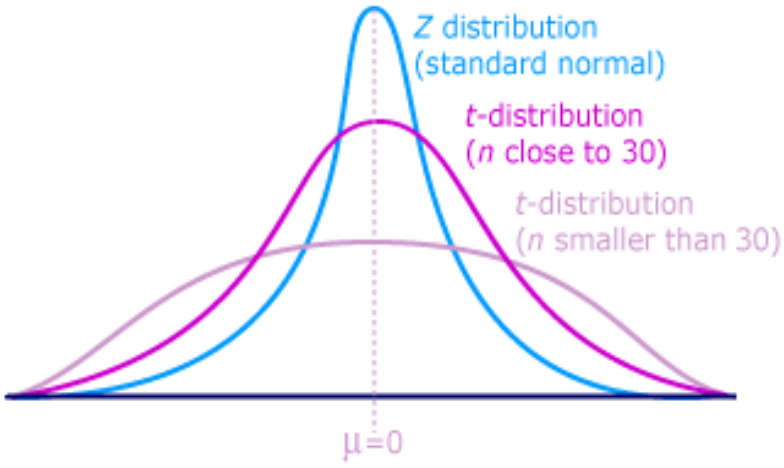

Figure 5: shape of the t-distribution depends upon the degrees of freedom

We use the t-distribution to turn a test statistic into a p-value. By plotting the t-distribution (based on the correct degrees of freedom) we can see the probability of observing a test statistic as extreme or more extreme than the one that we observed– that is, finally (!) we can evaluate statistical significance.

Critical value

The critical value is how we evaluate significance the other way around. The critical value takes our alpha (which is expressed as a probability) and turns it into a value in the t-distribution (i.e. something which can be meaningfully compared with our t-statistic). If the t-statistic is equal to or more extreme than the critical value, we can reject the null hypothesis. Again, the critical value can be found either using R or lookup tables.

Definition: The critical value is a point on the test distribution (here, the t-distribution) that is compared to the test statistic (here, t-statistic) to determine whether to reject the null hypothesis.

Confidence intervals

The confidence interval gives you a range of values which might contain the true value for the population. Typical confidence intervals include 90%, 95% and 99%. The confidence interval is calculated by taking the mean plus/minus the critical value multiplied by the standard error. For the normal distribution, a 95% confidence interval requires multiplying the standard error by 1.96.

A 95% confidence interval means that if you took an infinite number of samples then 95% of them would contain the true mean. This is **not** the same as saying that there is a 95% chance this particular confidence interval contains the mean! It simply lets you know that, given the parameters that you have set, 95% of all the confidence intervals calculated over an infinite number of samples would contain the true mean. In general, a smaller confidence interval (for a given level of confidence) is better.

NOTE: if your confidence interval includes ‘zero’ then, for this level of confidence, you do **not** have a statistically significant result.

3 | Some other stuff related to NHST

Assumptions of normality

A big research area in statistics is determining the shape of distributions and, in particular, assessing whether a variable is normally distributed. There are many tests available, such as the K-S test and Anderson-Darling. You can also 'eyeball' your data to see whether it looks normal (not very precise but a good starting point). In most cases, normality is assumed even if the actual distribution diverges quite considerably. Please note that testing for normality is **not** part of this term's work.

Degrees of freedom

Degrees of freedom states how many variables in a sample are 'free to vary'. This is an odd concept if you're new to statistics. But it's got a simple intuition behind it. Let's say we take a sample with just three values and am a mean of 12. The first value can be whatever we like – let's say it is 14. The second value can also be whatever we like – let's say 16. But the third value *cannot* be whatever we like. If we want to have a mean of 12 then the third variable must be 6 – because our first two values equal 30 (14 + 16 = 30); so to make an average of 12 – which is a total of 36 for the three values – the third value must be 6 (i.e., 36 – 30 = 6). So, we say that there is 2 degrees of freedom because two of the three values are *free to vary* – but once those two values are set then the third value is *determined*.

This holds as we increase the number of values in our sample; e.g. if we have 30 values then 29 of them are free to vary but the final value (number 30) is determined.

Type I error

When you reject a true null hypothesis – you think that you can reject the null (hurray, there really is a difference between the means!) but actually the difference does not truly exist. You think you have found something (i.e. accepted the alternative hypothesis) but really you haven't.

Type II error

When you fail to reject a false null hypothesis – or, in other words, you did not identify the true difference which actually exists between the means. In effect, you accepted the null hypothesis (and rejected the alternative) when you should not have done so! You *missed* the true difference...

Depending on the use case, type I or type II errors might be more concerning. For instance, if you are looking for cancer drug treatments then missing a treatment might be incredibly worrying. But, generally, in the social sciences we are super worried about type I errors because they bring the field into disrepute and lead to false knowledge (#FakeNews) being circulated.

4 | Some other terms

Leptokurtic

A distribution which has higher kurtosis than the normal distribution – 'kurtosis' is a term for describing a distribution. A leptokurtic distribution has very light tails and looks like a stretched up normal distribution.

Platykurtic

A distribution which has lower kurtosis than the normal distribution. The tails are very heavy, and it looks like a squashed down normal distribution. See the image below.

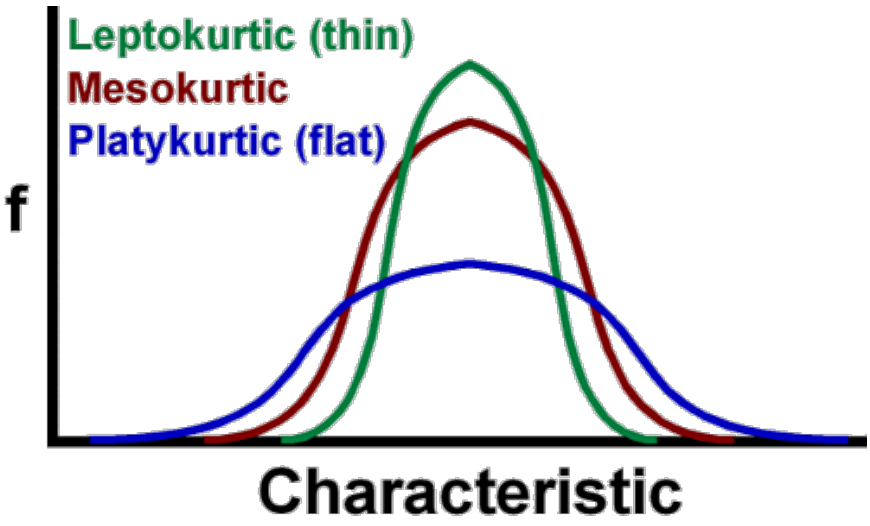

Figure 6: The kurtosis of different distributions

5 | The End

You made it! You got to the end… If you have any queries, comments or feedback then please email me at bertievidgen@gmail.com