Contents

Understanding significance testing

This white paper provides an overview of significance testing – one of the most commonly used but also most frequently misunderstood terms in statistics and data analysis. First, we'll provide a brief overview of what statistical testing is before delving in more depth into some of the challenges associated with using it correctly and interpreting the results of significance tests correctly.

What is significance testing?

The term 'tests of significance' was coined by the statistician R.A Fisher who has since been described as 'the founder of modern statistics' (Rao 1992). In his book Statistical Methods for Research Workers (1925), Fisher wrote that 'Critical tests of this kind may be called tests of significance, and when such tests are available, we may discover whether a second sample is or is not significantly different from the first.' It's hard to imagine that Fisher would have been comfortable with the casual way in which researchers, journalists and even politicians in subsequent decades have often argued that the veracity and importance of certain findings are almost indisputable due to them being 'statistically significant', something we'll examine more later in this white paper.

Statistical significance helps you to quantify whether the observed differences between two groups in your data are real or simply due to random chance.

If you are running an experiment, taking a poll, conducting market research or analysing data, you're typically going to be looking at a sample of the population that you're interested in, not the whole population. Let's imagine you're testing out a new marketing campaign. You send the new campaign to a sample of 10% of the people on your customer list. The remaining 90% receive the old campaign. Afterwards, people who saw the new campaign spent a mean amount of £21.99 compared to a mean spend of £19.50 amongst those who saw the old campaign. Does this mean that the new campaign is better that the old one? Maybe, but maybe not.

Perhaps the difference in the results is just due to random chance and doesn't really indicate any meaningful difference in the effectiveness of the two campaigns. Or perhaps the sample of people to whom you sent the new campaign was not truly random – a phenomenon known as sampling error. Sampling error typically occurs in one of two ways – either the size of the sample isn't large enough or there's some underlying variation in the population of interest that isn't being accounted for.

The issue of sample size is relatively intuitive. Imagine you flip a coin 10 times and you get 7 heads and 3 tails. Does that mean you can conclude that overall there's a 70% chance of getting a head on a coin flip? Clearly it doesn't. The sample is too small to provide a meaningful result. The more coins you flip, the closer you'll get to a 50/50 split between heads and tails. It's highly unlikely that if you flipped a hundred coins you'd get seventy heads, and vanishingly unlikely (although not impossible!) that if you flipped a thousand coins you'd get seven hundred heads. What this means is that the larger your sample, the more likely it is that any difference between the two groups in your results will reflect a true difference due to some factor of interest rather than simply random variation.

The second issue that can trip you up here is the level of underlying variation in your data. Imagine two different scenarios. In one, most people that respond to your campaign spend roughly the same amount – there's not a vast amount of variation in the data. Any customer that you pick at random is likely to have spent an amount that's pretty close to the average. In this situation any sample that you pick at random is unlikely to vary too much from the total population. However, another scenario might be that there's a lot of variation in your data. Some people spend very little whilst others spend a lot. Overall, the mean spend is still the same but if you pick a customer at random it's much less likely that their individual spend will be close to the mean. In this scenario, if you see a difference in spending between the two groups in your campaign test you can't be so confident that it's due to the campaign they received because the underlying level of variation in the data is higher.

What do we mean when we say a finding is statistically significant?

When we say that a finding is 'statistically significant' what we mean is that we can be confident that the finding is real and not due to an issue such as sampling error or some other underlying variance in the population. Let's break that down a bit more.

What is the null hypothesis?

When you run a marketing test (or any other type of experiment) you should be testing a null hypothesis against an alternative hypothesis. The null hypothesis is expressed in terms of an assumption that there will not be any difference between the two groups. In the example of the marketing campaign that we're using here the null hypothesis might be "There will be no difference in customer spend between the two campaign groups". The alternative hypothesis could be "Customers who receive the test campaign will have a higher average spend than those who receive the old campaign".

A null hypothesis can be thought of as a kind of default theory. It represents the current understanding of the researchers. Let's look at another example. Imagine you're a sports scientist researching reflex reactions and how they may or may not differ between the sexes. If you have no reason to believe that women have faster reflex reactions than men (or vice-versa), then that same lack of data leads to it being the null hypothesis.

To test this, you would collect data from a representative sample of male and female subjects and then examine it to see if it supports the notion that average reflex reaction times are probably the same for both groups. So, the null hypothesis is an implied stance that all statistical tests are measured against. It is characterised by the assumption that there is no relationship or no disparity between groups or factors in a study. As such it's analogous to the legal position of 'innocent until proven guilty'.

Other examples of null hypotheses might include assuming that blue eyed people in a given population exhibit the same variation in blood pressure as brown eyed people; that users of a music app in London are just as likely to cancel their subscription as those in Edinburgh; or that there is no relationship between height and memory recall ability.

What is the significance level?

The significance level – typically expressed as the p-value – is an expression of how rare your results would be if the null hypothesis were true. The lower the p-value, the lower the likelihood than an observed difference between your two groups is due to random chance. For example, a p-value of 0.05 or lower indicates that the chance of the null hypothesis being correct and the results being due to random chance is 5% or lower, so you reject the null hypothesis.

The purpose of calculating a probability value in a statistical test is to establish the likelihood that the data supports your null hypothesis. So, if the resultant probability value is quite large, there's not enough evidence to reject the null hypothesis. Note that this does not mean the null hypothesis is true, it just means that until we know better, we have no reason to reject it. However, if the probability is quite small, we can't really say that the data supports the null hypothesis, and so it is rejected.

The question of how small the p value should be before we reject the null hypothesis has proven to be one of the most vexatious in the field of statistics. Once again, a suggestion by R.A. Fisher has exerted enormous influence on the work of millions of statisticians, researchers and students in the decades since. Fisher suggested that probability values below 0.05 (equivalent to 5%, or 1 in 20) could serve as a default threshold when deciding whether an effect was 'significant'.

This arbitrary probability threshold of 5% is known in statistics as an alpha level. To be fair to Fisher, he didn't insist on it being set at 0.05 and he even suggested that that there may be times when one should use stricter alpha levels, 'If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty or one in a hundred'. Nevertheless, for generations of data analysts, the 5% alpha level has wielded an almost totemic power.

However, it is up to you as the researcher to set the threshold of significance with which you are comfortable. The level of 0.05 is commonly used but there is no rule that says this is the 'right' level for every situation. Often the level of significance with which you are comfortable will depend on the business decision that you're going to take and what the consequences might be of taking the 'wrong' decision.

If you're looking at differences between two marketing campaigns, as in our earlier example, you might be comfortable with higher p-value because the consequences of launching the campaign and then finding out that there's no real difference in spend are not very serious. However, if you were testing the effectiveness of new drug interventions however or trying to predict the outcome of a general election you might want to work with a p-value as a 5% chance of error might still be too high for comfort.

Just because something is statistically significant doesn't mean it's practically significant

The phrase 'statistically significant' occurs so frequently in data analysis reports, training courses and books about statistics, that most people rarely give it a second thought. But as a technical expression it is often at best unhelpful and at worst downright misleading.

It's not unreasonable to assume that saying a finding is 'statistically significant' implies some kind of revelation. After all, in common usage, the term 'significant' indicates something that is meaningful or notable. Unfortunately, in statistics 'significant' often means anything but that. Let's look at an example of how that could be the case.

Imagine that a researcher conducts an online survey of 20,000 adults visiting a gaming website. They compare the median amount bet by right-handed and left-handed gamblers over a two-week period and find that for right-handed people, this is £15.50, whereas for left-handed people the amount is £15.65, suggesting that perhaps right-handed people are more risk averse than those who are left-handed. The researcher choses a statistical procedure to test the null hypothesis that there is no difference in the levels of risk averseness between the two groups, however the procedure returned a probability value that shows this difference is in fact statistically significant – there is a difference in the median amount bet by left-handed people and their right-handed counterparts.

What's immediately obvious is that just because the difference in median bets might technically be deemed 'statistically significant' it's not 'practically significant'. The gaming company is hardly likely to redesign their entire website to be more attractive to left-handed people based on such a small difference between the two groups, even if the results are accurate (which might be questionable).

The arbitrary alpha level and the replication crisis

Another issue with significance testing stems from the arbitrary setting of an alpha level of 0.05, as we discussed briefly above. Many researchers nowadays argue that an alpha level of 0.05 has often proven to be insufficiently strict. So much so that it has led to a 'replication crisis' whereby many of the findings of influential studies, especially in the areas of social science and medicine, have been difficult or impossible to replicate or reproduce. The seminal paper Redefine statistical significance (Nature, 2017) recommended that in 'fields where the threshold for defining statistical significance for new discoveries is $p < 0.05$, we propose a change to $p < 0.005$. This simple step would immediately improve the reproducibility of scientific research in many fields.' The authors' concerns stem from the fact that they argue a level of 0.05 leads to too many false positives. Let's examine why that is.

Type I and Type II errors

The problem with a weak alpha level is that you are at a greater risk of incorrectly rejecting a null hypothesis, reporting that the effect observed in your findings is significant when in fact it is driven by random chance. This kind of false positive is known as a Type I error. At worst, Type I errors can mean drugs that have no effect are selected for treating patients.

Conversely, if you are too strict with your alpha level (by making it much smaller) you expose yourself to the danger of false negatives. A false negative occurs when you fail to reject a null hypothesis which actually is false. You fail to detect important effects when they really exist. This is known as a Type II error.

By changing the generally accepted default alpha level from 1 in 20 (P= 0.05) to 1 in 200 (p=0.005), as the authors of the Nature paper propose, we would see a substantial decline in the number of false positives being reported as significant findings (the authors' primary motivation for making the proposal). It might also mean that researchers would have to collect more data in order reject an existing null hypothesis. This is simply because a stricter alpha level may require more evidence to find an effect. The flip side of this is that if larger samples are not available then researchers might fail to demonstrate that a particular treatment is effective or that the life chances of two social groups vary when in reality these effects are present.

What does the 'P' value actually signify?

The probability value shown in a statistical test (often called the 'P' value) remains one of the most commonly misinterpreted figures in statistical analysis. The key to interpreting this kind of probability is to remember that it is based on some kind of analytical result in the form of a difference or relationship between groups or factors. This information effectively acts as the evidence used to potentially challenge the null hypothesis. It's therefore understandable that people often view this value either as the probability of the null hypothesis given the evidence, or the probability of the evidence given the null hypothesis. However, it is important to understand that in fact these two things are not the same.

In fact, it's easy to show that the probability of A given B is not the same as the probability of B given A. For example, what is the probability that someone has a driving license given they are aged over 30? Now ask yourself if this number is equal to the probability that someone is aged over 30, given that they have a driving licence? The two numbers are likely to be different as they may refer to different group sizes. It's important to keep this in mind as the probability values in statistical tests are often incorrectly described as 'the probability of the null hypothesis being true'.

In reality, these values show the probability of observing a result as extreme as the one obtained assuming the null hypothesis is true. This definition reminds us that it is always possible to observe relationships or differences that lead us to rejecting the null hypothesis simply by chance. It's important to bear in mind that the null hypothesis is never accepted. We can't prove that it is true, but we can collect evidence that shows it probably can't be maintained. Therefore, we either reject the null hypothesis or fail to reject it.

Examining a significance test

As a final example, let's look at a situation where researchers have conducted a recall memory test on two groups of 15 male and female subjects each (as indicated by label 1 in the image below). The initial group statistics show that the average score for the male subjects was 9.73, whereas the female subjects achieved a slightly higher mean score of 10.73 (as indicated by label 2). The researchers would like to know the probability that a difference at least as large as the one observed could exist between males and females in the wider population.

To check this, they have chosen to perform a T-Test, which tests the null hypothesis that the two means are in fact equal in the population. The resulting test has generated a probability value under the heading 'Sig. (2-tailed)' as indicated by label 3. This shows that the chance of observing a result as 'extreme' as the one observed (i.e. a difference of 1 in the mean recall scores from comparable samples) is likely to occur around 28.4% (p= 0.284) of the time, assuming that null hypothesis is true. On the basis of this analysis, they must conclude that they have insufficient evidence to reject the null hypothesis that sex does not influence memory recall.

Group Statistics

	Sex	1 N	2 Mean	Std. Deviation	Std. Error Mean
Memory Recall Score	Male	15	9.73	2.463	.636
	Female	15	10.73	2.549	.658

Independent Samples Test

	t	df	3 Sig. (2-tailed)	Mean Difference	Std. Error Difference
Memory Recall Score	-1.093	28	.284	-1.000	.915

The menagerie of statistical tests

A T-Test is just one of a plethora of standard statistical tests devoted to comparing group means and other summary statistics. Any student of statistics knows that these procedures have a number of data assumptions that must be met and that some approaches generate complex results that need to be interpreted with care. There are tests that compare variations in groups, check that data are normally distributed, measure linear relationships between factors and test differences in proportions. Each respective test has a default null hypothesis that assumes the variances are equal, the data are normally distributed, there is no linear relationship between the factors and that the proportions are the same between groups. Each test in turn generates a 'P' value that measures the probability of getting a result as extreme as the one observed assuming that the associated null hypothesis is true. In this way, the interpretation of a 'test of significance' is exactly the same from one test to another, only the context changes.

Conclusions

Significance testing remains one of the most important and commonly used statistical techniques but, as we hope this white paper shows, it's important to really understand what it is that you're saying when you say that something is statistically significant, and to be precise in the way that you use the language of significance otherwise you risk your findings being misrepresented or misinterpreted.

Next steps

- Take a look at our free on demand webinar – How to interpret significance tests
  - Calculating and interpreting confidence intervals correctly
  - How does a Chi Square test actually work?
  - How to interpret P values
  - When is significant not significant?
- Check out the Eat your greens series of articles on our website – each one examines a core statistical concept (such as significance testing) that is often misunderstood or misapplied.
- Get in touch with us for help with your project – our SPSS Boost service lets you book anything from a chat over the phone to few hours of consultancy and hand-holding to a fully personalised training plan, so if you have further questions about statistical significance (or any other statistical concept) let us know.

Contact us

info@sv-europe.com

020 7786 3568

www.sv-europe.com

Registered address

Level 17, Dashwood House,
69 Old Broad Street, London, EC2M 1QS