23

What are Tests of Significance?

What Problem are Significance Tests Designed to Address?

Typically the goal of research is to generalize results beyond the particular cases participating in the study. But how do we know when we can extrapolate from our sample? If 52 per cent of a sample indicate that they intend to vote for a particular political party, can we confidently say that 52 per cent of the population will vote this way? If the males in a sample earn $5000 more than the females, is this likely to be true of the population? How confidently can we apply our sample findings to the population?

Tests of statistical significance are a widely used method for addressing these questions. However, the method is not without its problems. Problems 24 and 25 examine some of the criticisms of tests of significance and ways of avoiding these misuses.

However, a clear understanding of the logic of significance tests is required before the criticisms and solutions make much sense.

What are Significance Tests?

- Tests of statistical significance are a subset of a wider category of statistics called *inferential statistics*. Inferential statistics are designed to assist in making inferences from samples to populations.
- Inferential statistics tell us *nothing* about the nature of patterns within a sample.
- Significance tests indicate the probability that results found in a sample are due to sampling error, or reflect patterns in the population from which the sample is drawn.

There is a wide range of tests of significance. Later sections will explain the difference between the different tests and when to use one rather than another (Problem 39).

What is a Null Hypothesis and how does it Relate to Significance Tests?

To make sense of tests of significance, you have to understand the concept of the null hypothesis. In some respects the null hypothesis is a bizarre concept that appears to be based on double negatives and designed to confuse people. In fact the logic of using null hypotheses reflects a conservative approach to scientific generalization.

In brief, a null hypothesis is:

- a statement of a pattern we *assume* exists in the population;
- a hypothesis that we seek to *reject*.
- normally the converse of the *substantive hypothesis* – the pattern we *expect*.

The logic of null hypothesis testing requires that we begin by *assuming* a particular pattern in the population. This pattern will be the opposite to that which we *expect* (on the basis of theory, etc.) to find.

Normally this assumption will be that variables are unrelated in the population – that groups do *not* differ from one another, that two variables are not related, etc. When we expect the absence of patterns, differences or relationships this is presented symbolically as $H_0 = 0$.[1]

With this starting point in mind, data are collected from a sample. These sample data are analysed with a view to checking whether the null hypothesis of no relationship fits the sample. Suppose we wanted to see if gender is related to marital satisfaction. Marital satisfaction is measured on a scale of 0 to 10, where a high score indicates a high level of marital satisfaction. Are men more satisfied with their marriage than are women? We can calculate the average level of marital satisfaction of men and of women.

The null hypothesis would be that there is no difference in the levels of marital satisfaction of men and women. That is, the difference in the male and female means will be 0. On analyzing the data from a sample we learn that men have an average of 7.5 on the scale, while women average 6.3. That is, the difference in means is 1.2.

We now have a 'problem'. We *assumed* a difference between men and women of 0, but have *observed* a difference of 1.2. How can we account for this discrepancy? There seem to be two options:

- *The sample data are wrong*. These data are misleading and do not provide any basis to revise our original assumption that the difference in satisfaction is 0.
- *The initial assumption is wrong*. The difference between male and female marital satisfaction means in the population is *not* zero.

If the first explanation is accepted, the reasoning is that the nature of the sample is such that sampling error (chance) could produce a misleading sample. By attributing the discrepancy between the null hypothesis of a zero difference and an observed sample difference of 1.2 to sampling error we *fail to reject the null hypothesis*. We continue to work on the *assumption* that the mean difference in the population is 0.

We must be very clear what it means to fail to reject the null hypothesis (forgive the confusing use of double and triple negatives, but this is the language of null hypothesis testing). Failing to reject the null hypothesis does *not* mean that:

- the null hypothesis is *proven;*
- the opposite to the null hypothesis (the substantive hypothesis that the sample difference of 1.2 reflects a real difference between men and women in the population) is wrong.

Failure to reject the null hypothesis simply means that we cannot rule out the chance that the differences between men and women in the sample could be due to sampling error.

If the second explanation is correct (i.e. the assumption of no difference is wrong) we should reject the null hypothesis. Since the null hypothesis is that $H_0 = 0$, rejecting it means that there is a good chance that in the population the male–female difference in marital satisfaction is something other than zero. Indeed, there is a good chance that a difference as great as we have observed in the sample exists in the population.

How to Decide to Accept or Reject the Null Hypothesis

Which of these alternative explanations should we accept? The choice is traditionally made on the basis of probability theory. When a probability sample is drawn from a population it is possible to estimate the probability of obtaining a discrepant sample – that is, a sample that does not mirror population patterns. Thus we can estimate the chance of obtaining a sample in which there *is* a difference between men and women even though they exhibit *no* difference in the population. For example, if the men and women in the population have the same level of marital satisfaction how many samples from that population would produce a male–female difference of at least 1.2?

For many years a magical cut-off point has been used to decide at what point the null hypothesis is rejected. Traditionally researchers have said that the null hypothesis should only be rejected if very few samples would produce the observed results because of sampling error. The precise cut-off point has varied a little but the 5 per cent (0.05) level is frequently applied. In larger samples the 1 per cent level often is applied.

What does it mean to apply the 5 per cent level? Using the example above of a difference of 1.2 points, if 100 samples were drawn from the same population in which there was *no* difference, how many samples would nevertheless produce a difference of at least 1.2 points? The 5 per cent level means that five out of 100 comparable samples could result in such a misleading finding of at least 1.2 points' difference. If our sample has found a difference of 1.2 it *could* be one of those five misleading samples.

When we say that a sample could produce a pattern that does not exist in its population we are describing *sampling error*. Clearly, the less likely it is that sampling error is responsible for the observed sample pattern the more confident we can be that the sample finding reflects something real in the population – the more confidently we can reject the null hypothesis.

As indicated earlier, researchers have traditionally used the 5 per cent or 1 per cent levels for rejecting the null hypothesis. If the probability (of sampling error) is *less* than this critical point, then researchers have taken their chances that their sample is probably not one of the faulty samples – there is at least a 95 per cent (or 99 per cent) chance that their sample reflects something real from the population.

- If we observed a difference of 1.2 points in a sample and the likelihood of this being due to sampling error was one in 100 we would report this mean difference of 1.2 points as being *statistically significant* at the 1 per cent level. The null hypothesis of no difference would be rejected.
- If there was a chance of four in 100, we would report the correlation as being statistically significant at the 4 per cent level. The null hypothesis would be rejected (assuming 5 per cent the critical cut-off).
- If there was a six in 100 chance of sampling error, the difference between male and female means would be reported as not statistically significant and the null hypothesis would *not* be rejected.

Using these probabilities is a cautious approach to research. All that failing to reject the null hypothesis means is that we cannot rule out the chance that the correlation we have observed *could* be due to sampling error. Unless we can very confidently rule out sampling error as an explanation of the sample findings we will assume that there is insufficient support for the substantive hypothesis.

The Use of Tests of Significance

The only use of a test of significance is to provide an estimate of the likelihood that a particular sample result differs from an assumed population level due to sampling error. The significance test provides an estimate of the probability of sampling error. Where the probability of sampling error is *less*

than a set level (e.g. 5 per cent or 1 per cent) the probability of sampling error is usually considered too low to worry about so the null hypothesis (of no relationship in the population) is rejected.

The rejection of a null hypothesis lends support to the substantive hypothesis – that a difference between groups, a correlation or some other pattern at least as substantial as found in the sample, is likely to exist in the population.

Should One-tailed or Two-tailed Significance Tests be Used?

Tests of significance come in two flavours: one-tailed and two-tailed. Since many statistics packages provide the option it is important to know what the difference is.

Choosing between one- and two-tailed tests depends on the nature of the substantive hypothesis being tested. We can distinguish between directional and non-directional hypotheses.

Non-Directional Hypotheses

Non-directional hypotheses predict that the sample value will simply be statistically different from a particular value. For example, we might anticipate that men and women will have different levels of marital satisfaction without predicting in which direction the difference will be – whether men or women will be more satisfied. We might predict that age and prejudice will be correlated without knowing whether older people will be less or more prejudiced than younger people. When your substantive hypothesis is non-directional, use a two-tailed test.

Directional Hypotheses

A directional hypothesis predicts an effect (correlation, difference between means of groups, etc.) *and* states the nature of the difference – that the correlation will be positive, that group A will have a higher mean than group B (or whatever the predicted direction is). One-tailed tests are used for testing directional hypotheses.

Notes

1 The null hypothesis does not have to be that $H_0 = 0$. As Cohen (1994) points out, we could begin with a proposition derived from a theory or previous research that specifies a correlation of a particular size and direction. In the Popperian tradition of trying to falsify a theory, this hypothesis would be the null hypothesis that we tried to reject.