Hypotheses and p values

When using frequentist statistics, we are always asking what would happen if we continually sampled from a population *where the effect we are interested in is not present*. This idea of a hypothetical 'no effect' situation is so important that it has a special name; it is called **the null hypothesis**. Every kind of statistical test (in this course at least) works by first specifying a particular null hypothesis. It is not possible to fully understand the results of a statistical test if we don't know the null hypothesis it relies on.

Hypotheses and null hypotheses

When discussing the scientific process, a hypothesis is a statement of a proposed process or mechanism which might be responsible for an observed pattern or effect. We have also seen that in statistics, we encounter 'hypothesis' used in a different, and quite specific way. In particular we frequently see the term: *null hypothesis* (often written in statistics books as H_0).

The null hypothesis is simply a statement of what we would expect to see if there is no effect of the factor we are looking at (e.g., plant morphology) on the variable that we measure (e.g., dry weight biomass). So in the plant morph example (encountered during t-tests) our null hypothesis is *There is no difference in mean biomass of purple and green plants*. All frequentist statistical tests work by specifying a null hypothesis and then evaluating the observed data to see if they deviate from the null hypothesis in a way that is inconsistent with sampling variation. This may seem like a rather odd approach, but this really is how frequentist tests work.

It is important to be aware of what a null hypothesis is, and what it is used for, so that we can interpret the results of statistical tests. However, in a general discussion of an analysis we normally refer to the effect we are actually interested in. This is called the *test hypothesis*, or the *alternative hypothesis* (often denoted H_1 in statistics books). The alternative hypothesis is essentially a statement of the effect we are expecting (or hoping!) to see, e.g., purple and green plants differ in their mean size. It is a statement of whatever is implied if the null hypothesis is not true.

Having got all the types of hypothesis sorted out, we can then use a particular frequentist technique to evaluate the observed result against that expected if the null hypothesis was true. The test gives us a probability (*p*-value) telling us how likely it is that we would have got the result we observe, or a more extreme result, if the null hypothesis was really true. If the value is sufficiently small we judge it unlikely that we would have seen this result if the null hypothesis was true. Consequently, we say we *reject the null hypothesis* (i.e. reject the notion that there is no difference). This is not the same as 'proving' the alternative hypothesis is true. We can't prove anything by collecting data or carrying out an experiment.

If the *p*-value is large, then it is quite likely that we could have got the observed result if the null hypothesis was true, i.e. it is due to sampling variation. In this case we cannot reject the null hypothesis. Note that in this situation we say that we "*do not reject the null hypothesis*". This is not the same as accepting that the null hypothesis is true, paradoxical though this may seem. One obvious reason for this is that if we only have a small sample then there may be an effect of the factor we are looking at, but we simply can't detect it because we don't have enough data.

Interpreting and reporting *p*-values

It is important to understand the meaning of the probabilities generated by frequentist tests. We have already said a *p*-value is the proportion of occasions on which we would expect to see a result at least as extreme as the one you actually observed if the null hypothesis (of no effect) was true. Conventionally, we accept a result as statistically significant if $p < 0.05$ (also expressed as 5%). This threshold is called the **significance level** of a test. We've said it before but it is worth repeating: there is nothing special about the $p < 0.05$ significance level! It is just a widely used convention.

Which significance level should you use?

We will always use the $p = 0.05$ threshold in this course. You need to remember this fact, because we aren't always going to remind you of it.

A probability of 0.05 is a chance of 1 in 20. This means that if there really was no effect of the factor we are investigating, we would expect to get a result significant at p=0.05 about 5 times in 100 samples. To envisage it more easily, it is slightly less than the chance of tossing a coin 4 times and getting 4 heads in a row (p=0.0625). It's not all that rare really. This puts a 'significant' result into context. Would we launch a new drug on the market or bring a prosecution for pollution on the evidence of the strength of four heads coming up in a row when a coin is tossed? Well of course such things are unlikely to hinge on a single test, but it is always worth bearing in mind what 'significance' actually means.

In general, the smaller the *p*-value the more confident one can be that the effect we see is 'real'. For a given analysis, a probability of p=0.001 provides stronger evidence for an effect being present than p=0.01. For this reason, in some critical applications such as drug testing the significance threshold may be lower than we use in biology. The costs of using a more stringent threshold is that this increases the possibility of false negatives—we are more likely to fail to detect an effect when it is present by adopting a lower significance threshold.

Careful with those *p*-values

This is a good time to issue an important warning about *p*-values. Frequentist *p*-values are counter-intuitive quantities that are easily (and often) misinterpreted. Whole books have been written about the problems associated with them. We don't have time to really cover the issues here, but here are a few key observations:

- Scientists tend to use p=0.05 to define 'significance', but p=0.055 is really no different from p=0.045. It would be irrational to reject an idea completely just on the basis of a result of p=0.055, while at the same time being prepared to invest large amounts of time and money implementing policies based on a result of p=0.045.

- The exact value of p is affected by the size of the true effect being studied, the amount of data being analysed, and how appropriate a statistical model is for those data. It's very easy to arrive at a tiny p-value, when an effect is weak or even absent, by using a statistical model that is inappropriate for the data in hand.
- The relationship between the 'strength of an effect' and its associated p-value is a complicated one. It is simply not correct to equate the size of a p-value with the weight of evidence for the effect being present, nor is correct to interpret a p-value a statement about how big the effect is.

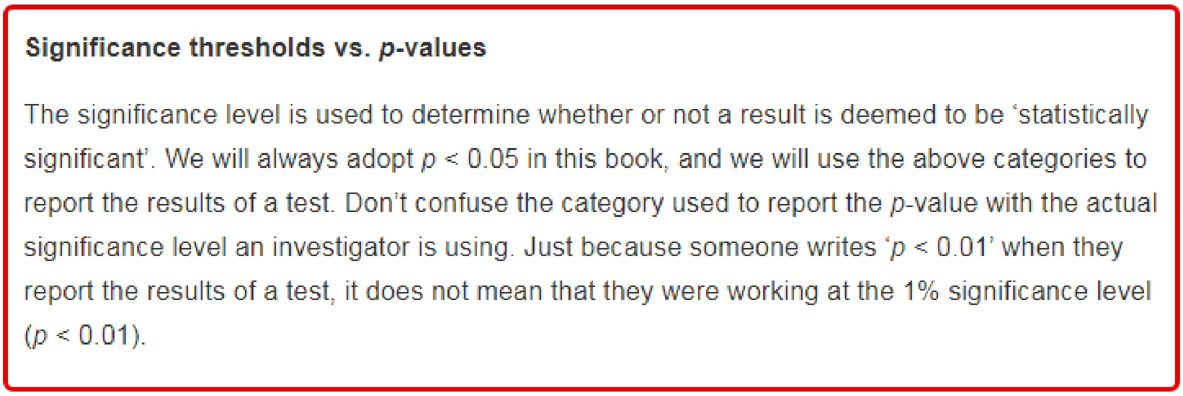
Take home message: p-values are hard to interpret, and should only be used as one line of evidence when answering scientific questions. They are not the 'final word' on truth.

Presenting *p*-values

R will typically display p-values from a statistical significance test to six decimal places (e.g. p = 0.003672). However, when we write about them, the results from tests are usually presented as one of the following four categories:

- $p > 0.05$, for results which are not statistically significant (sometimes also written as 'NS'),
- $p < 0.05$, for results where $0.01 < p < 0.05$,
- $p < 0.01$, for results where $0.001 < p < 0.01$,
- $p < 0.001$ for results where $p < 0.001$,

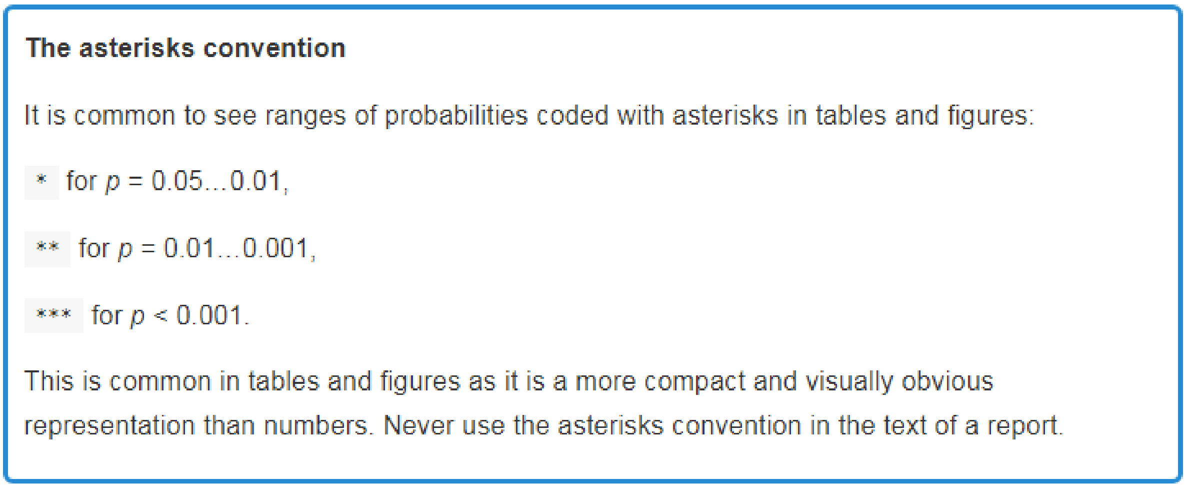
This style of presentation stems from the fact that statistical tests often had to be calculated by hand in the days before everyone had access to a computer. The significance of the result was difficult to calculate directly, so it would have been looked up in a special table. We still use this style because the value of p does not have a simple interpretation in terms of weight of evidence or effect sizes. Knowing which category a p-value falls into provides sufficient information to roughly judge 'how significant' the result is.



Significance thresholds vs. *p*-values

The significance level is used to determine whether or not a result is deemed to be ‘statistically significant’. We will always adopt $p < 0.05$ in this book, and we will use the above categories to report the results of a test. Don’t confuse the category used to report the p-value with the actual significance level an investigator is using. Just because someone writes ‘$p < 0.01$’ when they report the results of a test, it does not mean that they were working at the 1% significance level ($p < 0.01$).

It’s usually sufficient to use the four categories above when writing about the significance of a statistical test, though occasionally, giving the actual probability can be appropriate. For example, it can be informative to know that a test yielded $p = 0.06$ rather than simply quoting it just as $p > 0.05$ or NS. This is because p is so close to the significance threshold. While not wholly convincing, it is still suggestive of the possibility that an effect is present.



The asterisks convention

It is common to see ranges of probabilities coded with asterisks in tables and figures:

`*` for $p = 0.05 \ldots 0.01$,

`**` for $p = 0.01 \ldots 0.001$,

`***` for $p < 0.001$.

This is common in tables and figures as it is a more compact and visually obvious representation than numbers. Never use the asterisks convention in the text of a report.

Biological vs. statistical significance

A final, but vital, point: do not confuse statistical significance with biological significance. A result may be statistically highly significant (say $p < 0.001$) but biologically trivial. To give a real example, in a study of the factors determining the distribution of freshwater invertebrates in a river, the pH of water was measured in the open water and in the middle of the beds of submerged vegetation. There was a statistically significant difference in pH ($p < 0.01$) but the mean pH values were 7.1 in the open water and 6.9 in the weeds. This is a very small effect, and almost certainly of no importance at all to the invertebrates.

The significance of a result depends on a combination of three things (1) the size of the true effect in the population, (2) the variability of the data, and (3) the sample size. Even a tiny effect can be significant if the sample size is very large. Do not automatically equate a significant result with a large biological effect. Plot the data, inspect the estimates, and consider the biological implications of the difference. The statistical results provide some guidance in separating genuine differences from random variation, but they can’t tell us whether the difference is biologically interesting or important—that’s the scientist’s job!