NuMBerS
numerical methods for biosciences students

Biostatistics: HYPOTHESES AND HYPOTHESIS TESTING

1. Introduction

This helpsheet is concerned with the use of hypotheses in science and other areas of knowledge, and particularly how their use is supported by statistical tests such as those set out in the NuMBerS toolkits. In everyday use, the words 'hypothesis' and 'hypothetical' often carry the connotation of something imaginary and vague. However, the Austrian philosopher Karl Popper defined a scientific hypothesis as one that is 'falsifiable' - that is it can be subject to a rigorous test that is capable of demonstrating that the hypothesis is incorrect. So, defining a hypothesis in science is a process that sets out a description or explanation in a way that can be tested. Although Popper's criterion would look to a test that could unequivocally reject a hypothesis, in practice we often end up with a likelihood that a hypothesis is false rather than a black-and-white answer.

2. Overview of hypotheses

How does the process of hypothesis generation work? A 'research hypothesis' is a plausible explanation for a phenomenon or observation. It should logically provide one or more predictions, which can then be tested. Testing of predictions could involve making further observation and measurement, or could assess whether the predictions are compatible with other established facts related to the current observation.

A research hypothesis is a deductive approach to a problem. It generates predictions about the problem, which can be tested. In the formal testing of predictions using statistical methods, a 'statistical hypothesis' is defined that provides a rigorous test of a prediction. In Section 4, you will see that the statistical hypothesis is typically a 'null hypothesis', that is the prediction is false.

3. Analysis of quantitative data - descriptive- and inferential- statistics

Most science is ultimately quantitative, and the hypothesis-led approach demands a quantitative approach both to developing the hypothesis and its predictions, and testing the predictions. Descriptive statistics provide the tools to inspect data, from the simple calculation of averages and measures of variability through to the multivariate techniques used to make sense of very large and complex datasets.

Inferential statistics, on the other hand, are designed to test for patterns in data. Estimation techniques are used to quantify variability, for instance through the construction of confidence intervals. Statistical-hypothesis testing procedures are, as their name suggests, used to test particular types of hypotheses in relation to predictions.

4. The statistical null hypothesis

The use of inferential statistics to test the predictions arising originally from a research hypothesis typically refers to a so-called 'null hypothesis' (written H_0). As its name suggests, a null hypothesis sets out to establish that nothing unusual is present, and by inference the prediction is untrue. The alternative hypothesis (H_1) posits that something unusual is present, and that the prediction is likely to be true.

We use the null hypothesis, because it is mathematically and logically easier to construct the test. Essentially, all of the inferential statistical tests in the NuMBerS toolkits are concerned with establishing the likelihood that particular pattern suggested by the prediction arose simply by chance.

The procedure comprises four steps, that are used in each of the toolkits:

- Construct a null hypothesis that is appropriate to your prediction, eg if the prediction is that there is a difference between the means of samples from two populations, the null hypothesis is that the means are the same.
- Decide on a critical significance level (denoted by α, the Greek letter alpha). As we indicated in the introduction, we are rarely in a position to reject a hypothesis categorically, and in these tests we set a level of uncertainty that we are prepared to accept in testing the null hypothesis.
- Calculate the statistic that is appropriate to your null hypothesis and is consistent with your data. For instance, when testing for difference between two unrelated samples where data fulfil parametric criteria, calculate the t-statistic.
- Reject or accept your null hypothesis, either by comparing the value of the calculated statistic with published values for the statistic for given critical significance level and degrees of freedom, or by directly using a probability (or significance level). The probability, P, is the likelihood of obtaining data equal to or more extreme than the observations were the null hypothesis to be true. So a very low probability ($P \leq \alpha$) indicates that you should reject the null hypothesis that the pattern in the data arose purely by chance, and that the prediction is probably correct.

5. Probability, error and power

The critical significance level introduced in Section 4 implies that there is a finite possibility of error in testing the null hypothesis. If we select $\alpha = 0.05$, then we are saying that we will accept our null hypothesis even if the probability that the data we are testing conform with the hypothesis is as low as 5%. This still leaves a small possibility that we will reject a null hypothesis when we really should have accepted it. In the case of $\alpha = 0.05$, we are likely to make the wrong choice once in every twenty tests on average, for $\alpha = 0.01$ it is once in every hundred and for $\alpha = 0.001$ once in every thousand.

This error is termed a type 1 error, and is pre-defined by the critical significance level. A type 2 error occurs if you accept a false null hypothesis, and has a probability denoted by β (Greek letter beta). This is essentially an indication of some form of mismatch between the null hypothesis and the observations. The quantity $(1 - \beta)$ is termed the

'power' of the test, and is an estimate of the effectiveness of the test in avoiding a type 2 error.

Power can be increased by:

- Increased sample size
- A high 'signal to noise' ratio in the data, that is a strong effect of the source of variation under study and a low effect from other sources of variation

6. Experimental design

From the discussion of hypothesis testing, error and power, it is clear that the effectiveness of testing depends ultimately on the data. Where data have been collected before the hypothesis is defined, there is often little that can be done to improve matters. However, where an experiment is designed specifically to test a hypothesis, it is often possible to optimise the measurements so that the test is effective, that is to increase the power of the test (see Section 5).

Good experimental design is often quite complex, especially when it comes to isolating the source of variation under study and being able to exclude other sources of variation. Agricultural field trials are a good example, where differences in test plots have to be attributable to crop variety or fertilizer treatment, not to soil conditions or exposure to wind. But even in the more controlled conditions within a laboratory, it is important that an experiment is thought through and matches the prediction and the hypothesis used to test it. There are various techniques to optimise the effectiveness of experimental design, including power analyses.