A Brief Overview of Statistical Testing

Sylvia L. Halkin, Central Connecticut State University

You have learned in your introductory science courses about hypotheses that are explanations proposed by scientists to explain phenomena they observe, or to predict the outcome of an experiment testing the effect of a particular variable. In statistical testing, another kind of hypotheses, called null hypotheses, are used. A null hypothesis predicts that the variable being tested has no effect.

For some statistical tests, you actually need to calculate what data would be expected if the variable being tested has no effect. Often this expected value is simply the average of the values from your different test groups. For example, if the null hypothesis is that a cold medicine has no effect, and colds of people taking the medicine last an average of 4 days, while colds of people not taking the medicine last an average of 10 days, then the expected value you'd use for both groups in statistical testing would be $(4 + 10) / 2 = 7$ days.

Each null hypothesis has corresponding alternative hypotheses. The most general kind of alternative hypothesis simply states that the variable being tested does have an effect. A more specific alternative hypothesis would specify what effect the variable has (for example, that the cold medicine being tested shortens the duration of colds, or that flowers with their nectar at the end of a long tube will have fewer species of pollinators than flowers with their nectar near the surface). The kind of explanatory hypotheses with which you are familiar are examples of this specific type of alternative hypothesis.

In statistical testing, we test whether our data are significantly different from those predicted by the null hypothesis. More precisely, we calculate the probability that results as different as ours are from those predicted by the null hypothesis, or more different, would occur by chance alone, if the null hypothesis were true. Statistical tests generally involve calculations that produce a number that you look up in a table to find this probability, also known as p, or the p-value. If our data closely resemble those predicted by the null hypothesis, there will be a high p-value and thus a high probability that chance alone can account for any differences that exist, and we conclude that the null hypothesis is probably true. On the other hand, if our data are very different from those predicted by the null hypothesis, there will be a low p-value and thus a low probability that these differences are due to random chance. In this case we conclude that the null hypothesis is probably not true, that the variable(s) we were testing probably did have an effect, and that an alternative hypothesis is probably true.

How low should the p-value be, in order to allow you to reject the null hypothesis? It is standard in scientific work that if the p-value falls below 0.05, the difference between your data and those predicted by the null is considered to be "statistically significant", and the null hypothesis is rejected. However, this decision really depends upon the risks of accepting the null if it is actually false, compared to the risks of rejecting the null if it is actually true. For example, if the null hypothesis states that a new mushroom is the same in toxicity as known poisonous mushrooms, even with a probability of only 1% that the null is true and the mushroom is poisonous ($p = 0.01$, well below the “normal” cutoff of $p = 0.05$), it would be wise not to eat it: the risks of rejecting the null if it is actually true are too high, while accepting the null, even if it is false, just means that you’ll have one less mushroom to eat. On the other hand, if the null hypothesis states that a potential new cancer drug has no more effect than a placebo, even with a probability as high as 30% that the null is true and the drug is useless, most people would opt to continue testing the drug. Wrongly accepting the null hypothesis in this case could mean ignoring a drug that could eliminate a lot of human suffering, while wrongly rejecting it just means that some more time and money will be spent in testing.

In statistical tests whose results you report, you should give the name of the test you used (e.g., Chi-square test of independence, or Mann-Whitney U test), the value of the test statistic (e.g., $\chi^2 = 24.8$, or Mann-Whitney $U = 15$), the degrees of freedom relevant to the test you are using, and the corresponding p-value. If your sample sizes are not clear from your data tables, you should also give these.

Note that failure to reject a null hypothesis does not necessarily mean that it is true. Small sample sizes limit your ability to generate low p-values. For example, if you toss a coin 100 times and every time it comes up heads, there is a very low probability that it is a fair coin. However, if you tossed the same coin just 4 times and got 4 heads in a row, that result is not statistically significantly different from what would be expected with a fair coin ($p = 0.0625$).

Importantly, even if an experiment produces very low p-values that <u>support</u> an alternative hypothesis, scientists never say that they have 100% proved that a particular alternative hypothesis is correct. There are two reasons for this: first, statistics tell us that if the null hypothesis is true, there is a small but finite probability that even results extremely different from those predicted by the null hypothesis will occasionally occur by chance alone. The second reason is that it is always possible that someone else will think of an alternative hypothesis that better explains our results than the alternative hypothesis that we proposed. For example, they may take into account a variable we had not thought of, that was actually important. This is particularly likely to be the case if the variable we thought we were testing is actually only correlated with the important variable, and we did not think of the correlation. An alternative hypothesis is therefore tentatively accepted on the basis of experimental results that support it, and as long as it is making more accurate predictions than other alternatives, scientists are perfectly comfortable making decisions that assume it is correct. However, scientists also acknowledge that the particular hypothesis (or theory) in favor at any given time may in the future be replaced or supplemented by a hypothesis that allows even more accurate predictions to be made.

Thanks to David Spector for helpful discussions on the topics covered on this and other handouts.