Lecture 9

Guide to comparing two samples

This section is a review and contains guidance about how to choose the appropriate testing and confidence interval method for your data. Not being good at writing flow charts, I'll write the guide as an outline:

Guide to comparing two samples

1. Determine if your data is paired. If it is paired, then

   (a) Create a column for the pairwise differences and work with that column.

   (b) For the purposes of assessing normality, all that matters are whether the differences are normal - the individual distributions are irrelevant.

   (c) For large sample sizes, a one sample t-test or confidence interval on the data will almost always be appropriate. (As always, given a sample size, I can create data where this statement is false. But it is not common with real data.)

   (d) For small sample sizes and roughly normally distributed data, a one sample t-test or confidence interval on the data is appropriate.

   (e) The sign test is always appropriate. It is not very powerful because it only considers the signs of the data. But, when it is significant, it says it loudly and clearly with no distributional assumptions made on the data at all. (The observations should still be random and independent of each other.)

   (f) For small sample sizes with moderate to severe skewness, for instance, the Wilcoxon Ranked Sign Test is appropriate with the interpretation that a significant result means that the differences are not symmetric about zero.

   (g) Note: You will never use a log transform on the paired differences. Why? If you could, it would mean that all the differences were positive and thus a sign test would conclude that there is a difference for all but the smallest of sample sizes.

2. If your data is not paired, determine whether it needs a log transform by looking at the normal probability plot and/or histogram for both samples. If a log transform for both data sets is appropriate, then

   (a) Use 2-sample t-test procedures on the logged data.

   (b) But interpret the results on the scale of the original data by exponentiating the confidence interval limits, for instance. The interpretation on the original scale will be an interpretation about the median rather than the mean and will be multiplicative rather than additive in nature.

3. If your data is not paired, highly non-normal (or non-normal with small sample sizes) and a log transform is not appropriate or does not work (in the sense of improving normality), then

   (a) If your data is continuous or does not have too many ties, you can use the Mann-Whitney Rank test. However, recall, that the null hypothesis for this test is that the two distributions are equal. To conclude that it is their medians that are not equal, other aspects about the two distributions, such as their variances, have to be the same.

   (b) In a worst case scenario, with small sample sizes, many ties, or other problems, you can conduct a permutation test, a simulation approximation to a permutation test, or a bootstrap test. Simulation methods require programming and are thus harder to implement than other methods. They are also not as powerful or as accurate as normality-based procedures when normality-based procedures apply. But they do provide a good "method of last resort."

4. If you have large datasets without too many repeats or small data sets that are roughly normal then

   (a) Standard 2-sample t-test procedures are appropriate. I recommend using Welch's t-test unless there is a compelling or philosophical reason to use a pooled-variance t-test. The pooled-variance t-tests is sensitive to the assumption of equal variances when the two sample sizes are not equal (unbalanced design.) It is not overly sensitive to the assumption of equal variances when the design is balanced (equal sample sizes in the two groups). This is one reason why it is best to aim for a balanced design when possible when you are designing your own experiment.

These tests are only as good as your data - meaning, if your data are not representative of their underlying population distributions, by flawed design or flawed sampling, then the results of these methods will also be off. If your data is not random and if the values are not independent of each other, your results may be off. This can always be an issue when you have missing data. For instance, if some of your individuals die during your experiment and those are not random, and if you conduct your analysis ignoring the dead individuals, your reported results may be biased.

Exercises for Lecture 9

1. –

2. –