Lecture 8 – A postscript

Do not confuse statistical significance with biological consequence

The use of the term "significant" to refer to statistical deviations from a null hypothesis is in some ways unfortunate, since of course in every day language "significant" means very important.

It must therefore be noted that with a large enough sample, even tiny deviations from the null hypothesis can be statistically significant, but of little or no consequence biologically at all.

Always look at the *magnitude* of the difference, as well as the p-value. Does this much difference mean anything biologically?

"Not rejecting the null hypothesis" is not the same as "accepting the null hypothesis" - Beware Type II errors

In truth, almost any null hypothesis is not strictly true. There is most probably at least a very small difference between two groups, even if that difference is not measurable or statistically significant. Therefore, we can never truly say that a null hypothesis is

"true" but we can say something about how much deviation from that hypothesis there is likely to be.

In evaluating whether to use the null hypothesis as our current state of knowledge, when we fail to reject it, we need to ask ourselves:

Is the sample large enough?
Is there a small, but important effect?

If p<0.05 happens once in 20 tries: Remain skeptical!

The standard significance level of α=0.05 is actually fairly lax: 1 in 20 significant results is a Type I error. When reading the literature, it is important to keep in mind potential publication bias, that significant results are more likely to be written up, submitted and accepted for publication than non-significant results. As a result, it is almost certainly true that the Type I error rate of published results is higher than the statedα!

The journal *Nature*, the primary journal of biology, rejects more than 95% of the papers submitted to it, and that surprising results are much more likely to be published there. Many Type I errors, with apparently exciting, novel results, are published in these journals as a result.)

The data can not tell you whether to use a one- or two-tailed test

The determination of what the predicted or interesting result will be must be made before examining the data, otherwise the data can bias the choice of alternate hypotheses. The data will always deviate form the null hypothesis in a particular direction -- if we were allowed to use the data to generate our hypotheses then our hypotheses would never differ from our data, and statistics would be useless!

Do not confuse correlation with cause

Cause and effect cannot be proven statistically. At best, statistics can answer questions pertaining to correlation. Only a good experiment can prove cause.

No snooping!

Similarly, all hypotheses must be stated before starting to examine a data set (i.e. before taking a data set!) . If one takes a particular data set and asks every question possible from it, eventually a significant result will be found even if by Type I error. But if that results were then reported in isolation of the other unsuccessful tests, the p value associated with it would be misleading. If you do enough tests, one will

eventually be significant at the 5% level, but there is a much higher than 5% chance of a Type I error for that analysis as a whole.

If $\alpha = 0.05$, the probability of *not* getting a Type I error is 0.95 if one does only one test. But if one did *t* tests, then the probability of not getting an error is 0.95^t, and the probability of at least one Type I error becomes $1- 0.95^t$. For 10 tests, say, the probability of at least one Type I error is about 40%!