HYPOTHESIS TESTING

The null hypothesis

We come now to a peculiarity of inferential statistics which often seems illogical. The basic ideas, however, are as simple as usual. In presenting the example of testing for the significance of differences between the mean statistics scores of anxious and non-anxious students, I referred earlier to the working hypothesis. On the basis of research and theory, there might be grounds for hypothesizing that non-anxious students will do better as a group on statistics than anxious ones. A researcher who collected the data might very well undertake the research with a strong expectation that significant differences will occur. However, in order to test whether a difference between two means is statistically significant, it is necessary to formulate a null hypothesis. In the present example, the null hypothesis which the researcher would have to test, in order to determine if the two sample means differed significantly, would have the following form: <u>There is NO difference between anxious and non-anxious students.</u>

This does not mean that a researcher seriously believes that there is no difference between the group means (or whatever sample statistics are being compared); formulation of the null hypothesis is a step in the faunal procedure for testing for significant differences. It is required by the logic of inferential statistics, not by the logic with which researchers develop their working hypotheses.

This can be explained by returning to the original definition of the role of inferential statistics. Suppose we have two sample means which are numerically unequal. If the numerical difference is large, it seems reasonable to speculate that it is significant; if the difference is small, common sense suggests that it might not be significant. To take a simple example from everyday life: The Adelaide Crows defeated Hawthorn by one point on June 8, 1992. This one-point result could have gone either way had circumstances been slightly different, and can hardly be regarded as a sign that Adelaide were significantly better than Hawthorn on the day. A small difference is inconclusive. However, when Adelaide defeated Hawthorn by 13 goals in March, 1991, you might well have decided that there was a significant difference between the two sides that day. Such a large win can hardly be attributed to chance. A big difference encourages you to regard the result as conclusive: The difference in scores indicates a genuine difference between the two clubs.

The following question now arises: How big must the difference in scores be before we can say with confidence that it was not just chance, the luck of the game, a small slipup which would never be repeated? What we are doing here is starting from the position that there is no systematic difference between the two teams until we have evidence to the contrary. A small difference in scores is not enough to make us reject the opinion that there is no difference, apart from chance factors. The difference must be large before we reject the idea of **no difference.** When it is huge, for instance 13 goals, it is pretty clear that the difference was genuine, Adelaide **really were** better in that particular match and we reject the hypothesis of **no difference.**

Inferential statistics work in a similar way. We start from the idea that there is **no difference:** When the difference becomes huge, we reject the idea of no difference. The basic supposition of no difference is, in fact, a **null hypothesis.** Take the case of anxious and non-anxious students: A small difference between the sample means is not big enough to convince us that non-anxious students really do get better statistics marks than anxious ones. We do **not reject the null hypothesis.** A very big difference convinces us that the non-anxious students, really did do better than the anxious ones, and we **reject** the null hypothesis. **The contribution of inferential statistics is to provide rules which tell us how big a difference must be in order to cause us to reject the null hypothesis.** The logic of inferential statistics goes like this:

1. Formulate a null hypothesis.

2. Calculate an appropriate inferential statistic.

3. Check whether the inferential statistic is so large that you must reject the null hypothesis.

The alternative hypothesis

It is usual to formulate not only a null hypothesis, but also a so called **alternative hypothesis.** If the null hypothesis is rejected, then the alternative hypothesis must be correct. The null hypothesis is usually abbreviated as H_o and the alternative hypothesis as H_1. To return to the anxious students and statistics marks, the procedure would have the following form:

1. Formulate H_o = There is **no** difference in statistics marks between anxious and non-anxious students.

2. Formulate H_1 = There **is** a difference between anxious and non-anxious students.

3. Calculate an appropriate inferential statistic.

4. Check if this statistic is so large that H_o must be rejected,

5. If it is not, do not reject H_o. If it is, reject H_o and assume that H_1 is true

Types of Error

At this stage you may ask:

1. What are "appropriate inferential statistics"?

2. How do you know if the value of this statistic is so large that the null hypothesis must be rejected?

You already have one answer to both questions: You could calculate the confidence intervals of the two samples, and if they did not overlap you could reject the null hypothesis. However, the requirement that the confidence intervals do not overlap is too demanding; when you reject the null hypothesis according to this criterion you do so correctly, but when you **fail** to reject it you may make a mistake. It is as though you would only be willing to conclude that one football team had played better than the other if the losing team scored 0 - 0, and the winner at least 20 goals. By applying this harsh decision rule you would be quite correct in concluding that the difference between 20 - 10 and 0 - 0 was significant, but you would be forced to conclude that there was no significant difference between the teams in the case of a score of let us say 18 - 20 to 0 - 1[1]. You have set too demanding a criterion for rejecting the null hypothesis. In the case you have made an error because you demanded too big a difference before rejecting the null hypothesis. This kind of error is referred to as p - error or Type II error. Type II error occurs when we **do not reject** the null hypothesis, although we should have. This occurs when there really is a difference, but our criterion for rejecting the null hypothesis fails to detect the difference.

In psychological research more emphasis is placed on t:4.- error (Type I error). This error occurs when we incorrectly reject the null hypothesis, i.e., there is really no difference, but our criterion for rejecting the null hypothesis is too "soft" and leads us to reject it incorrectly.

To summarize:

1. **Type I** error occurs when the null hypothesis is **incorrectly rejected.**
   We incorrectly conclude that there is a significant difference when there really isn't one.
2. **Type II** error occurs when we **incorrectly fail to reject the** null hypothesis.
   We incorrectly conclude that there is no difference when there really is one.

The application of statistics has been oriented towards rejecting the null hypothesis only if there are very strong grounds for doing so -- better to be too cautious than too reckless. To take the statistics-anxiety example, we are very cautious about rejecting the null hypothesis and concluding that the mean statistics scores of anxious and non-anxious students are not significantly different. The justification for this is that it is worse to take an unjustified action than to fail to take a justified one; when in doubt, do nothing. The sense of this position is obvious if we consider an example from medicine. It is more dangerous to give a new drug if we are not 100% sure about it than to withhold it. Better to accept the null hypothesis, the drug is **not** safe, than to reject it too hastily.