Homogeneity of Effects: Interaction

Introduction

Many medical studies are carried out to investigate the size of some effect, for example the treatment effect in a controlled clinical trial or the risk of exposure to some substance in an observational study. To this end it is usual to compare groups of subjects, such as those receiving active treatment with those getting placebo or those exposed with those not exposed. The comparison is often made using a statistical technique, such as a *t*-test, a χ^2 test, ANCOVA. If these techniques (or indeed many other statistical techniques) are applied to the groups as a whole then only one estimate of the size of an effect will be obtained, i.e. there is an implicit assumption that the effect size is constant across different types of patients or subjects. In practice this may not be sensible, for example women may respond to a hormonal treatment differently from men or the effect of exposure to an environmental toxin may be different for smokers and non-smokers. Such *heterogeneity of effect* is usually investigated by repeating analyses on selected subgroups. Virtually any trial report will include a series of *subgroup analyses*, and manoeuvres with similar motivation occur in many epidemiological studies. Although the biology or medicine behind this approach may be sensible, there are several statistical traps for the unwary and some of these are discussed in this note.

There are two aspects to the statistical complications of the investigation of heterogeneity. The first concerns the logical and technical issues about how to assess heterogeneity of an effect between specified subgroups. The second concerns more subtle problems about how the subgroups are selected in the first place. These will be dealt with in turn.

Nomenclature

The statistical term used for the notion of heterogeneity is interaction. An *interaction* between two variables, such as sex and treatment, means that the effect of treatment is different in the sexes. Because an interaction is when the effect of one variable depends on the level of another variable, the term *effect modifier* is sometimes used (especially in social science applications).

An Example: Vitamin D supplementation and of breast- and bottle-fed infants

An example which will be used to illustrate the assessment of heterogeneity between specified groups is a controlled trial of vitamin D supplementation for the prevention of neonatal hypocalcaemia [1], in which expectant mothers were randomised to receive either supplements of vitamin D or placebo: several endpoints were measured but we will consider only the serum Ca of the baby at one week of age. The effect of vitamin D supplementation was assessed separately in those infants who were breast- and bottle-fed : table 1 is based on data from the study.

Comparison of the treatment groups in breast-fed infants gives P=0.44, whereas the same comparison amongst bottle-fed infants gives P=0.0002. There is a temptation to claim that the difference in P-values establishes a difference because

"there is an effect in the bottle-fed group but not in the breast-fed group". This is false: the key to realising this is to recall that a statement such as P=0.44 does not mean there is no difference, merely that we have found no evidence for an effect. A P-

	Breast-fed		Bottle-fed	
	Supplement	Placebo	Supplement	Placebo
Treatment Mean	2.445	2.408	2.300	2.195
n	64	102	169	285
SE	0.0365	0.0311	0.0211	0.0189
Treatment Effect	0.037		0.105	
SE	0.0480		0.0283	
P-value	0.44		0.0002	

Table 1: summary statistics for serum Ca (mmol/l) for breast and bottle fed subgroups from [1]

value is a composite which depends not only on the size of an effect but also on how precisely the effect has been estimated (its standard error). Thus we cannot tell whether the difference in P-values arose because there was no effect in the breast-fed group or there was an effect, perhaps even one of similar size to that in the bottle-fed group, but which was less precisely estimated.

Before indicating the correct analysis it is useful to rehearse the arguments that led to the P-value in one of the groups, take the breast fed group as an example. In table 1 the mean serum calciums are 2.445 and 2.408 mmol/l in the vitamin D and placebo groups respectively: the difference in means, which we will refer to as the *treatment effect*, is 0.037 mmol/l (= 2.445 - 2.408). The standard errors of the two means are 0.0365 and 0.0311 respectively. In order to assess whether the true treatment effect is different from zero, the standard error of the difference in means, that is the standard error (SE) of the value 0.037 is required. Finding the SE of a difference between two quantities which themselves have SEs se_1 and se_2 is a standard statistical manoeuvre (see [2], page 160-1) giving:

$$se_{diff} = \sqrt{se_1^2 + se_2^2} = \sqrt{0.0365^2 + 0.0311^2} = 0.0480.$$

The P-value corresponding to the observed difference of 0.037 is found from the ratio 0.037/0.048=0.771 using standard methods (see [2] p. 165-7) which give P=0.44; how this calculation can be done in Minitab is shown in the Appendix.

As pointed out above, it is wrong to conclude that the effects of vitamin supplementation differ between breast- and bottle-fed infants because of the differences in P-values, essentially because P-values measure more than the effect size. The correct way is to compare the *effect* sizes directly: the effect in the breast fed group is 0.037 mmol/l and in the bottle-fed group 0.105 mmol/l, so the difference in effects is

0.105 - 0.037 = 0.068 mmol/l. The question of whether or not the treatment effect differs between the feeding groups is down to an assessment of whether this difference could have arisen by chance. However, this is exactly the same question that arose in the assessment of a treatment effect within one of the groups, so the method of answering it is the same. That is, we must apply to the numbers in the boxes in table 1 with the heavier shading the same procedure that was applied to the numbers in the lightly shaded boxes to get the P-value 0.44. Thus by analogy with the above, the SE of the difference in treatment effects is:

$$se_{diff\ of\ effects} = \sqrt{se_1^2 + se_2^2} = \sqrt{0.0480^2 + 0.0283^2} = 0.0557$$

so the ratio of the difference in effect sizes to its SE is 0.068/0.0557 = 1.22, and the P-value is 0.22. Therefore, there is no evidence that vitamin D supplementation has different effects on the serum Ca of breast- and bottle-fed babies. A 95% confidence interval for the difference in effects is $0.068 \pm 1.96 \times 0.0557$ = (-0.0412, 0.1772) mmol/l.

Thus, the correct method of comparison gives a different conclusion to the simple comparison of P-values.

	Steroid group		Placebo group		P-value
Sex of Baby					
Boys	14.9%	24/161	14.1%	24/170	0.96
Girls	4.8%	7/146	18.8%	24/128	<0.001
Pre-eclampsia groups					
with pre-eclampsia	21.2%	7/33	27.3%	9/33	0.57
without pre-eclampsia	7.9%	21/267	14.1%	37/262	0.021

Table 2: subgroup results from antenatal steroid trial: figures are percentages & fractions with RDS.

Another Example

Further illustration of the problems of comparing P-values is in another controlled trial of antenatal prophylaxis, this time of maternal steroids for the prevention of neonatal respiratory distress syndrome (RDS) in the baby. Unlike the previous example, the outcome is binary, i.e. whether or not the baby had RDS. In the trial report [3] numerous subgroups were analysed separately, including the effect on boys and girls and the effect on children born to mothers with and without pre-eclampsia.

A formal analysis of this table will not be presented: once the standard errors of the proportions have been calculated (see [2], p.161-2) the methods the same as those just described for a continuous outcome. However, it is instructive to consider the differences between the two types of subgroups.

Effect of sex of baby

As in the first example, the P-values for treatment comparisons in the boys and girls are quite different. The treatment effects, as measured by the difference in percentages, are 14% in girls and -0.8% in boys. A correct test of interaction gives P=0.007, showing there is strong evidence that antenatal steroid therapy has different effects on boys and girls. Indeed, the treatment effect appears to be confined to girls.

Effect of pre-eclampsia

Again, there is a marked difference between the P-values for the two groups, with a "significant" effect in the pre-eclampsia group but a "non-significant" difference in the other group. However, in this case the treatment effects in the two subgroups are very similar, being 6.06% in those born to mothers with pre-eclampsia and 6.26% for those whose mothers did not have pre-eclampsia. With such similar treatment effects it is not surprising that a test for interaction provides no evidence at all of any difference, P=0.99. In this instance it seems reasonable to postulate that the steroids have the same effect on children regardless of whether or not their mother suffered from pre-eclampsia but because only 66 mothers had pre-eclampsia as compared with 529 who did not, the precision with which the treatment effect is measured is much higher in the latter group, and it is this difference, rather than any difference in treatment effects, that is responsible for the disparity in the P-values.

In the above, only pairs of variables each with two levels, boys and girls, steroids and placebo, etc. have been considered. It is possible to investigate interactions between categorical variables with more than two levels or even with continuous variables. However, these are not considered here because they introduce no essentially new ideas but are technically considerably more involved.

How Should the Subgroups be Selected?

So far three examples of interactions have been discussed, and in two cases the *prima facie* evidence of an interaction based on P-values alone was shown to be misleading. In the third example the correct analysis revealed evidence that antenatal steroids did affect male and female children differently. However, in the study numerous subgroups were selected and it is not surprising in these circumstances that results in some of them are "significant". This is not to say that the effect is not there, merely that interpreting the results of such hypothesis tests is difficult. In a study measuring ten variables, interactions between all 45 pairs of variables could be assessed, so it is clear there is much scope for hunting through the data looking for "significant" findings.

It is clearly more convincing if the difference makes sense in terms of the biology of the study and to this end it is sensible *at the start of a study* to specify which subgroups are to be analysed. Of course, an unanticipated finding may be biologically plausible and clinically important, but *post hoc* rationalisation can only carry a limited amount of conviction. In this example, the evidence for the difference in effect between boys and girls was not sustained in subsequent trials [4]. It is not uncommon for unanticipated, but quite definite effects in a study not to be supported in other studies. It is therefore wise to interpret such effects cautiously and, for those of sufficient importance, perform a subsequent study specifically designed to investigate them.

Summary

i) When assessing interactions, make sure the correct test is applied, just comparing P-values will lead you astray

ii) At the planning stage, specify the subgroup analyses you intend to perform

iii) Treat unanticipated subgroup findings with caution

References

1. Cockburn, F, Belton, NR, Purvis, RJ, Giles, MM, Brown, JK, Turner, TL, Wilkinson, EM, Forfar, JO, Barrie, WJM, McKay, GS, Pocock, SJ (1980) Maternal vitamin D intake and mineral metabolism in mothers and their newborn infants. *British Medical Journal*, 281, 11-14.
2. Altman, DG, (1991) *Practical Statistics for Medical Research: London*, Chapman and Hall
3. Collaborative Group on Antenatal Steroid Therapy (1981) Effect of antenatal dexamethasone administration on the prevention of respiratory distress syndrome. *American Journal of Obstetrics and Gynecology*, 141, 276-287.
4. Crowley, P, Chalmers, I, Keirse, MJNC. (1990) The effects of corticosteroid administration before preterm delivery: an overview of the evidence from controlled trials. *British Journal of Obstetrics and Gynaecology*, 97, 11-15

Appendix: calculating a P-value in Minitab

Since the samples in the neonatal hypocalcaemia study are quite large, the P value from the ratio of the mean difference to its standard error can be found from the Normal distribution rather than the *t*-distribution. In Minitab the calculation can be done through the **Calc** menu item, selecting successively **Probability Distributions** and

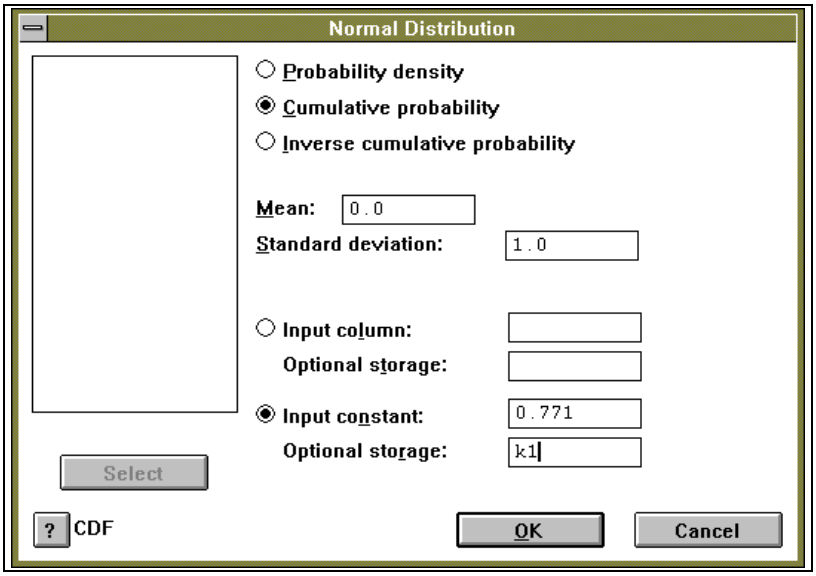

Figure A1

Normal. This leads to the screen in figure A1, into which the value of 0.771 should be entered as shown (for the comparison in the breast-fed group). When the OK button is clicked the proportion of a standard Normal population that is below 0.771 is entered into k1. The P-value is the proportion of this population that is *above* 0.771 or below -0.771: these last two are the same, so the P-value is twice the proportion above 0.771, which is one minus the contents of k1, which can be found as shown in figure A2.

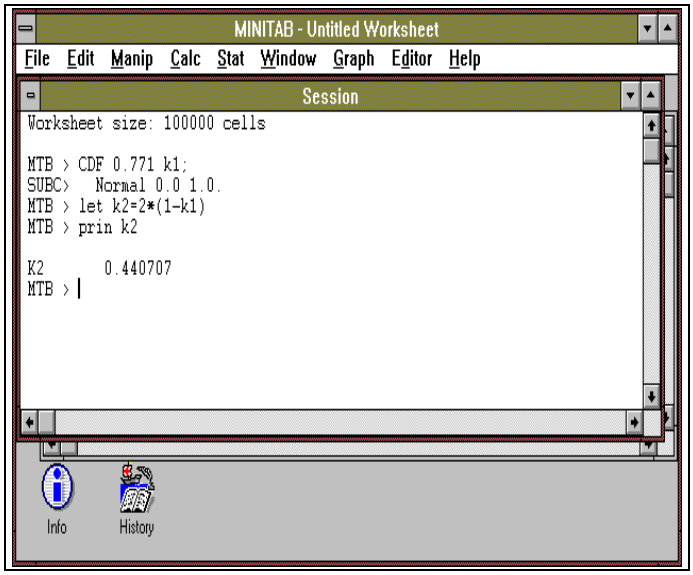

Figure A2