Clinical Biostatistics

Sugested answers to exercise: Significance Tests

(a) *What is meant by 'P<0.001'?* This is the result of a significance test. P is the probability of getting a difference as far from expectation as that observed when the null hypothesis is in fact true. Here the null hypothesis is that there is no difference between the two treatments as measured by a visual assessment. P is very small showing that the probability of the observed difference occurring if there is really no difference between the treatments was small. We say that the difference is statistically significant and conclude that there is good evidence for a difference between the treatments in the whole population.

(b) *What is meant by type I error and type II error?* A type I error is when we get a significant result when the null hypothesis is true. The probability of a type I error is equal to the P value. The maximum type I error probability is set in advance, usually 0.05. A type II error occurs when we get a non-significant result when the null hypothesis is false. In other words, we fail to detect a real difference. The probability of a type II error depends on the size of the difference in the population, as well as on the sample size and the significance level chosen.

(c) *What is wrong with this approach to the analysis?*

(d) *Suggest a better method.* The authors are looking at each group separately, testing the null hypothesis that there is no change in adrenaline. If we compare in this way, we are wrongly interpreting a non-significant result as meaning that there is no effect. It would be much better to compare the two groups directly, testing the null hypothesis that the change is the same in the two groups. This could be done by the large sample Normal comparison of two means.

(e) *What problems are there in this approach to testing?* In a one sided test, the alternative hypothesis is that there is a difference in a specified direction. The null hypothesis is then that there is no difference or a difference in the opposite direction. This is reasonable if a difference in the opposite direction would have the same meaning or result in the same action as would no difference. For the comparison with the control group, this argument does not hold. There is no a priori reason to suppose that the exposed group would be any different in cancer risk than the controls. In fact, the controls were chosen so that the risk would be the same, apart from any risk due to the exposure. Thus an excess of cancer in the control group would be very surprising and lead us to conclude either that radiation exposure protected against cancer or that the groups were not comparable. If we found no difference, on the other hand, we would conclude that there was no evidence that the radiation influenced cancer risk. The conclusions would be different and a one sided test in the direction of more cancers in the exposed group cannot be justified. It is even harder to justify a one sided test in the direction of fewer cancers in the exposed group, opposite to the research hypothesis. For the comparison to the general population, it could be argued that the exposed group, predominantly servicemen, are selected and would have a reduced cancer risk, a phenomenon known as the 'healthy worker effect'. A one sided test in the direction of more cancers in the exposed group would be arguable, because if radiation had no effect the number of cancers in the exposed group would be the same as or less than that in the general population.

(f) *What two sided test would be equivalent to a one sided test at the 5% level in the direction of the difference?* To test in the direction of the observed difference is in fact

to carry out tests in both directions simultaneously. As one of these tests assumes that fewer cancers in the controls is equivalent to no difference, and the other assumes that more cancers in the controls is equivalent to no difference, the tests are contradictory. The procedure is the same as a two sided test at the 10% level, and is not truly one sided at all.

(g) *What problems are there in the design of this study?* The control cycles were the three cycles preceding the treatment cycles, thus there was no placebo and so the assessment could not be blind. Hence there could be assessment bias. It would have been better to have included placebo cycles in a randomized design.

(h) *What problems are there in the analysis?* In this table, the effect of each drug has been analysed for each treatment separately, calculating the difference between the control and the treatment cycles. A comparison between treatments would have been more informative than the within treatment tests presented. This would have indicated if there were any differences between the drugs.

(i) *The authors reported that they found no significant differences between the groups at baseline. Why were these tests of significance unnecessary?* Because the subjects were allocated at random to one of two groups, any differences in the characteristics of the groups would have occurred by chance, by definition of randomization. Therefore the null hypothesis is true and so tests of significance would be meaningless.

(j) *The authors stated that 'the difference between the groups in the number of women falling during the whole two year period was not significant (P=0.58), but between 12 months and 18 months into the study the difference was significant (P=0.011)'. Does the 'P=0.011' add anything of value to the results of this study?* This is multiple testing. If we keep testing the data as we collect them, the chance of a spurious significant difference increases. In other words, if the null hypothesis were true, the probability of a significant result would not be 0.05 but something bigger. We therefore test only at the end of the study. The intermediate test does not mean anything.