University of York
Department of Health Sciences
Applied Biostatistics
Suggested answers to exercise: Comparing the means of small samples

Question 1

a) *What method could be used to calculate P in this study?* In this study we have means from two independent samples. Hence we can use an unpaired or two sample t test to test the null hypothesis that the means are the same in the populations from which samples were drawn.

b) *What conditions, if any, do the data have to fulfil for the method to be valid?* This method would assume that the data are from Normal distributions, with the same variance.

c) *Are they likely to be fulfilled here?* These variables are measurements of skeletal size which often follow a Normal distribution. The standard deviations are similar. They are also small in comparison with the means, providing no evidence of skewness.

d) *From these P values, can we conclude that the length of femoral neck in elderly women has increased over time?* There is good evidence that the length of the femoral neck is different in the populations which these samples represent because the difference was statistically significant. To conclude that there is a change over time we must assume that the samples are truly representative of elderly white women in these two decades.

e) *Can we conclude that the width of femoral neck in elderly women has not increased over time?* There is no evidence that the width of the femoral neck is different in the populations which these samples represent because the difference was not significant. However, this does not mean that it has not changed. The sample may be too small to detect a change which has occurred. A confidence interval would be more informative. The difference is 0.5 with 95% confidence -1.0 to 2.0. The clinical implication is that elderly women in the 1990s had longer, but not thicker, femoral necks than similar women in the 1950s. Longer bones may be more likely to break and so this may explain any increase in fractured neck of femur.

Question 2

a) *What is meant by no significant difference?* No significant difference is the result of a significance test comparing PEFR in children at two points in time after exposure. The null hypothesis is that there is no change in PEFR in the population from which the sample of children was drawn. The data was used to give a test statistic which has a known distribution if the null hypothesis is true. We then find the probability of data as or more extreme as that observed if the null hypothesis were true. If the probability is small then we have good evidence against the null hypothesis. Conventionally, we use P=0.05 as the cut-off. Here the probability must be >0.05, therefore we conclude that the study failed to detect a difference. This does not necessarily mean that no difference exists since we cannot prove that the null hypothesis is true on the basis of a significance test.

b) *What is meant by a paired t test?* The paired t test is used to test the null hypothesis as above. It is used here because we have two measurements on each subject at two points in time and we are interested in the differences within individuals.

c) *What assumptions are involved and are they likely to be justified here?* The t test can be used for small samples but requires that the differences follow a Normal distribution. This is likely to be true because PEFR follows a Normal distribution and the difference of two variables from Normal distributions will also follow a Normal distribution.

d) *What can we conclude about the effect of spillage on the respiratory function of children?* There is no evidence for a reduction in lung function in the children between 3 and 12 days after the spillage. However, we should not conclude that the spillage had no effect on lung function. There may be an effect which is too small to be statistically significant in a sample of this size. Alternatively, lung function may have already been reduced before the initial reading was taken and remained reduced.

Question 3

a) *What is meant by a 'two-tailed, unpaired Student's t test'?* The two-tailed, unpaired Student's t test is used to compare means from two independent samples. It tests the null hypothesis that the means are the same in the populations from which the samples are drawn against the alternative hypothesis of a difference in either direction. If the null hypothesis is true then the difference between means divided by the standard error of the difference follows the t distribution with $38 (= 20-1 + 20-1)$ degrees of freedom.

b) *What conditions must the data satisfy for these t tests to be valid?* The assumptions are that the cholesterol data are from Normal distributions with the same variance.

c) *Are these likely to satisfied here?* In general serum concentrations are often positively skew with the variance increasing with the mean. Here, the standard errors are consistently bigger for the larger of the two means for each type of cholesterol. Since the sample size is equal for the smokers and non-smokers, the standard deviations must also increase with the means. Hence, the assumptions of the tests may not be met. However, with equal numbers in 2 groups the test is very robust, i.e. gives true P-values when the null hypothesis is true, though some power may be lost.

d) *What extra information could be given in the table?* It would be useful to show the difference between the means and a 95% confidence interval for that difference. In addition, the actual P-value is more informative than `NS'.

e) *What aspect of the data has been ignored in the analysis?* The analysis has ignored the matching for age and sex and has treated the groups as independent. A matched analysis, such as using a paired t test would take the structure of the data into account. If cholesterol is actually related to the matching variables age and sex, a paired test would remove some of the variation which is included in the standard error in the unpaired test. The paired test would be more powerful. If cholesterol were unrelated to the matching variables then a paired test would not be necessary. When the sample is very small, the loss of degrees of freedom may even make a paired test less powerful and so be counter-productive in these circumstances.