Methods of Clinical Evaluation and Analysis II

Suggested Answers: Comparing means

a) *What method would be used to carry out the tests of significance shown in the figure, and why?* Paired t test, because we have the same subject measured on two occasions.

b) *What can we conclude from these tests?* We can conclude that there is good evidence that in group A creatinine fell between the start of treatment and the sixth month of follow-up, and between six months before treatment and the sixth month of follow-up. There is no evidence for any fall in creatinine in group B. We cannot conclude that creatinine did ***not*** fall in group B, or that groups A and B differ.

c) *What test of significance would be better in this study?* We should compare the two groups directly, using two sample t test for the creatinine at six months follow-up. Even better would be to use the three month data as well, perhaps by averaging the two. We should also control for the pre-treatment levels. We could compare the mean difference between pre- and post-treatment creatinine between the two groups. An even better way to do this, which we will look at later in the course, would be to use analysis of covariance of the post-treatment creatinine using the pre-treatment creatinine as a covariate.

d) *What method would be used to calculate the confidence interval, and why? What condition should the data meet for this method?* This is a comparison of two samples, so we would use an unpaired or two sample t method. The data should be from Normal distributions with the same variance for the standard two sample t method. However, in this study the standard deviations in the two samples are very different, 48·7 µmol/L in study patients compared with 14·3 µmol/L in controls. We should therefore use the two sample t method for unequal variances, the Satterthwaite correction to the degrees of freedom.

e) *The standard deviations are bigger than the means. Why should we NOT conclude that change in serum creatinine has a skew distribution?* The changes could be negative, so it is quite possible to have observations two standard deviations below the mean.

(f) *Is there anything to suggest that the t test may not be valid?* It is not likely that semen count follows a Normal distribution. The standard deviation is more than half the mean, so mean minus 2 SD would be negative. We would expect 2.5\% of observations to be below this if the distribution were Normal. This would imply negative values, which would be impossible.

(g) *What are the implications of this for the t test? What could be done about it?* The distribution must be positively skew, which means that the t test is not strictly speaking valid. However, the samples are almost the same size and the test is robust. This means that if the null hypothesis were true, we would get 5% of tests 'significant'. If the null hypothesis were false and there was a difference, the power to detect it would be reduced by the skewness and we would have less chance of detecting the difference than if assumptions were met. The distribution could be made more symmetric by a log transformation, which we will discuss next week. Alternatively, we could use a test which does not rely on such assumptions, such as the Mann Whitney U test, outside the scope of this module.

(h) *Are the t tests important to the conclusions of the study?* The t tests are not important to the conclusions of the study. What we want to know is how good the semen index is at discriminating between fertile and infertile donors, not whether the mean level is different. The degree of overlap between the fertile and infertile populations is much more important. As the means and standard deviations suggest that there is much overlap, the conventional semen indices do not discriminate well. The sensitivity and specificity of the semen tests would be more useful.