Clinical Biostatistics

Suggested answers to exercise: Regression and correlation

1. *"The mean body temperature was 37.9^{o}C in climbers with cerebral oedema, compared with 36.9^{o}C in climbers with a score ≤3 (mean difference 1.0^{o}C (95% confidence interval 0.5 to 1.5))".* What does this mean and what method could be used to calculate the 95% confidence interval? The difference in the average temperature for the sample was 1.0, but we are using this as an estimate for the population of climbers as a whole. We estimate that in this population the mean body temperature in climbers with cerebral oedema will be somewhere between 0.5 and 1.5 degrees greater than mean temperature in climbers with mild or no mountain sickness. The two sample t method could be used, provided the data followed an approximately Normal distribution.

2. *"The correlation coefficient between the body temperature and arterial oxygen pressure ... [was] −0.52 (P<0.001)". What does this mean?* The correlation coefficient measures the strength of the linear relationship. -0.52 indicates a relationship of moderate strength. The negative sign tells us that high body temperature is associated with low arterial oxygen pressure. P<0.001 shows that if they were no relationship in the population we would get a correlation as strong as this for fewer than 1 in 1000 samples. Hence there is good evidence that this relationship is present in the whole population of climbers.

3. *Figure 1 shows axillary temperature plotted against mountain sickness score, with correlation coefficients and associated P values. What condition must the data meet for the P value to be valid?* At least one of the variables must follow a Normal distribution. Do you think the condition is met for Day 1? Sickness score looks positively skew. Temperature looks roughly symmetrical, although it may be slightly positively skew. However, there are not 60 points here, which means that some points are coincident. This makes it rather difficult to judge. The distribution looks sufficiently close to the Normal for the P value to be OK.

4. *Do you think this study was ethical?* This is a very unusual study, in that they deliberately set out to make people ill. Presumably, the climbers had experience of altitude and knew what they were letting themselves in for, though no mention is made of how consent was obtained. 15/60 (25%) of the subjects had to be evacuated by helicopter. I have great doubts about this. It is, of course, a matter of opinion.

5. *What will be the effects of using patients attending their general practice?* We would like a sample of the general population. Instead, we have people selected by attending their GP. This means that they will be sicker and hence older then the general population. Thus the sample is biased. It seems implausible that the relationship between ear size and age will be different for this sample, so in practice it seems unlikely to have any effect.. If we were trying to estimate mean ear size, the age bias would have an effect.

6. *Will the dropping of patients due to the seriousness of their presenting problem or the late running of the surgery have any effect?* The sample is biased, but again this seems unlikely to matter here. .

7. *Is this study blind? Does this have any implications for the interpretation of the results?* The study is not blind. GPs clearly knew the patients' ages before they

measured their ears. The measurement of ear size with a ruler has a subjective element. GPs may have been biased in their measurements, elevating the measurements of the older subjects and lowering those of the younger subjects. It is hard to see how this measurement could be made without the observer being aware of the subject's age.

8. *Is the distribution of ear size skew or symmetrical, and why?* It appears to be symmetrical. The mean (67.5) is almost exactly in the middle of the range (52.0-84.0).

$$\frac{52.0+84.0}{2}=68.0$$

The figure also suggests this.

9. *What is a regression equation? What does the one in the paper tell us?* A regression equation predicts one continuous variable from another. The line shows the estimated mean ear size for given age. The line has slope 0.22, i.e. mean ear size increases by 0.22 mm for each year of age. When age is zero, the line would cross the vertical axis at 55.9.

10. *Can we conclude that the mean ear size at birth is 55.9 mm?* When age is zero, the line would cross the vertical axis at 55.9. This does not mean that babies have ears of this size, because we would be extrapolating beyond the data. We cannot do this because we cannot assume that the straight line relationship will also be valid for children.

11. *What assumptions about the data are required for regression analysis and do you think they satisfied here?* We assume that the deviations from the regression line follow a Normal distribution and have uniform variation along the line. The second assumption looks very reasonable from the figure. The spread of the data about the same is very similar all the way along. It is difficult to tell about the Normal distribution.

12. *What are the conclusions and are they justified by the data?* `It seems therefore that as we get older our ears get bigger (on average by 0.22 mm a year)' --- last paragraph of the `Methods and results'. This is a cross-sectional not a longitudinal study, and the uncertainty in the estimate should be included. Strictly speaking, the conclusions should read `It seems therefore that older people have bigger ears (on average by between 0.17 and 0.27 mm per year of age)'. It could be that the ears of different birth cohorts differ. After all, different birth cohorts have different mean heights at the same age.

13. *What further investigations could be done?* For these data, it would be interesting to look at men and women separately. Is it old men who have big ears, or old people? Ideally, we would like to follow people over time, if not from cradle to grave at least for several years, to see whether this is a phenomenon of individual growth or of differences between generations.