UNIVERSITY OF NEWCASTLE UPON TYNE

Premedical course

MOCK Examination I

Outline solutions

1.
a) Mention histograms, stem and leaf plots, cumulative frequency curves.
b) Hypothesis testing is perhaps better suited to clinical decision making but beware, "non-significant" P-values, these do not mean the null hypothesis is true, merely that you do not have evidence that it is false. Confidence intervals provide an interval estimate of the parameter under consideration - linked to hypothesis testing, as testing that a parameter has a particular value, e.g. $\mu=0$ will be significant at the 5% level if and only if the value 0 is outside a 95% confidence interval for μ. However, 95% confidence interval gives more information, as it indicates all the values that would not be rejected by a hypothesis test at 5% level.
c) The apparent reversal of a relationship by a lurking variable. Should give an example.

2.
Proportion of children with ulcer who are Rh+ is 36/42=0.86. For control children this value is 28/40=0.7.
Odds ratio is (36x12)/(6x28) =2.57. The natural log of this is 0.9445 and the s.e. of this is $\sqrt{(1/36+1/6+1/28+1/12)}=0.5599$. Therefore the 95% confidence interval for the *log* of the odds ratio is $0.9445\pm1.96\text{x}0.5599$= (-0.1529, 2.0419), and taking antilogs gives a 95% confidence interval for the odds ratio of (0.86, 7.71).
The odds ratio indicates that children with peptic ulcers are 2.57 times more likely to be Rh+ than are children without peptic ulcers. However, this factor could be between 0.86 (i.e. children with peptic ulcers are *less* likely to be Rh+ than the control children) or as large as 7.71 times *more* likely to be Rh+. That the confidence interval stretches between cases being less likely to be Rh+ up to more likely to be Rh+ indicates that there is no conclusive evidence one way or the other from this study.

3.
The analyst has done the following
a) Produced stem and leaf plots describing both the age at starting playing and the peak number of games per season separately by those who have been injured and those who have not. The distribution of age of starting is slightly skewed to older ages, whereas the number of games is when most active is highly right-skewed, with considerable evidence of digit preference (respondents tend to give answers ending in 0 or 5)
b) Produced the mean, median, SD and SE for each variable, again separately for those who have and have not been injured.

Together these analyses provide a description of the variables underlying the analysis. It is plausible that injuries may be more prevalent amongst those who play more games, or perhaps amongst those who started later in life, and these analyses may confirm or refute such notions.

c) Performed two-sample (i.e. unpaired) t-tests with pooled errors, comparing both age at starting and number of games per season between the injured and uninjured groups. There is strong evidence that the uninjured group started playing at an older age (mean difference in age 7 years, 95 % confidence interval for difference (2.4, 11.5). There is no evidence of a difference between the groups in the number of games per season but this result is suspect because of the marked skewness in the data. Note that the means are always much less than the SDs. These analyses go some way to answering whether these variables are associated with injury to the wrist etc.
d) The analyst then tabulates the data on injury with respect to the sex of the player and finds that there is no evidence of a difference between the proportion of males and females who sustain injury. However, the values of these proportion have not been calculated and should have been (although this is trivial to do from the output presented).
e) Finally the analyst investigates whether there is a relationship between the age at starting playing and the peak number of games per season, both through a scatterplot and regression analysis. It is may be that younger players play more intensively and this accounts for the increased risk affecting younger players (with the direct analysis of the variable `'ngames'` not picking this up because of its skewed nature). As the analysis shows, this turns out not to be a possible explanation.
f) Some information that would be useful is not available from these variables. The time after starting playing at which the (first) severe injury was sustained would be useful, as would the time of injury in relation to the time when playing was at its peak.