University of York

Department of Health Sciences

Applied Biostatistics

Suggested answers to exercise: The analysis of cross-tabulations

Question 1

a) *What is meant by 'chi-squared = 23.98, P<0.001?'* This is the result of the chi-squared test which tests the null hypothesis that there is no association between *P. alcalifaciens* and foreign travel. The value 23.98 is the test statistic which will follow a chi-squared distribution with 1 degree of freedom if the null hypothesis is true. P<0.001 tells us that the probability of these data or more extreme data occurring if the null hypothesis were true is smaller than 0.001 and so we have good evidence that the null hypothesis is not true and conclude that an association exists.

b) *What conditions do the data have to meet for the test to be valid?* The chi-squared test is a large sample test and the usual rule is that the large sample approximation holds if all expected frequencies are greater than 5 for a 2 by 2 table. Although one observed frequency is 5, no expected values will be as small. This is because if the null hypothesis were true then the overall probability of being positive for *P. alcalifaciens* would be 28/627 = 0.04 and this proportion would apply to those who have and those who have not travelled abroad. Thus the expected numbers positive for *P. alcalifaciens* would be 254 × 28/627 = 11.3 for those who have travelled abroad and 373 × 28/627 = 16.7 among those who have not travelled abroad. The other expected values can be calculated in a similar way but will be large because the expected values must add to the marginal totals for each row and column.

c) *What conclusions can be drawn from these data?* The study shows that there is a statistically significant association between travelling abroad and being positive for *P. alcalifaciens* among people with gastroenteritis. We cannot conclude from this that *P. alcalifaciens* was the cause of the gastroenteritis. We can only conclude that an association between *Providencia* and foreign travel exists.

d) *What other information would be useful in deciding whether P. alcalifaciens was a likely cause of gastroenteritis in travellers?* We need a control group. We could look at the number of positive screens for *P. alcalifaciens* among subjects without diarrhoea cross-classified according to whether or not they had recently travelled abroad. This would tell us if the observed association between travel and *P. alcalifaciens* was a general one or one specific to those with diarrhoea.

Question 2

a) *What is wrong with this statement and what analysis should they have done?* The authors appear to have tested each line of the three by two contingency table separately. This would involve doing three significance tests using the same data. This increases the chance of a type I error, a significant difference where there is none in the population. The authors could have done a chi-squared test for a three by two contingency table.

Question 3

a) *What method could we use to test the null hypothesis that the two classifications are related, and why?* The expected values here must all be small, as all four of them must sum to 8. Hence we cannot use a chi-squared tests, but must use Fisher's exact test. This is not significant, P = 0.067. The invalid chi-squared test would give chi-squared = 6.00, df = 1, P = 0.014.

Question 4

a) *What is a trend test and how would you interpret the one presented here?* The trend test is the Armitage chi-squared test for trend or the Mantel-Haenszel test for trend. It works by assigning numerical values to each category and then estimating the best prediction of one variable by the other as a simple y = constant$_1$ + constant$_2$×x. The chi-squared test for trend tests the null hypothesis that there is no such prediction and constant$_2$ = 0. The calculated chi-squared value is 20.6 with 1 degree of freedom with an associated P-value smaller than 0.001. This provides strong evidence for a trend in the proportions and so we would conclude that the proportion of males who are seropositive increases with age.

b) *What would be the advantages and disadvantages compared to a chi-squared test for association in a contingency table?* The test for trend takes into account the ordering of the age groups, which the ordinary contingency table chi-squared does not. Hence the test for trend has much greater power to detect a steady increase (or decrease) in seropositivity with age. However the test has less power to detect non-linear relationships, such as seropositivity being higher among young men and older men than among those in the middle of the age range. Such a relationship would produce a non-significant trend test.

c) *Suggest an alternative way of testing the difference in age in the two seropositivity groups, assuming that the raw data were available.* Assuming that age was recorded in years and given the large samples (176 seropositive, 305 seronegative), we could compare the difference in mean age between the two groups using the large sample Normal comparison method (z test).