Chapter 8 – Categorical analyses

Using the Salaries.csv data set, examine the different ranks of academics to see if the proportions are to be expected. To give you some idea of what is going on, you may find it useful to start by studying the descriptive statistics (which is usual good practice) for the 'rank' variable.

What is the most appropriate test?

The Chi Squared Goodness of Fit test analyzes the frequencies of more than two categories within a single variable.

What are the null and alternative hypotheses?

H0: There is no significant difference from the expected proportion.

H1: There is a significant difference from the expected proportion.

Interpret the results. For simplicity's sake, use the default equal proportions as expected proportions.

file = read.csv("Salaries.csv")

rank = file$rank

summary(rank)



The categories are decidedly unequal.

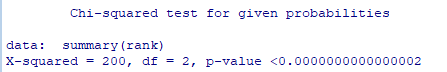
rank[1: 10]



We examine the top 10 cases (material on data handling is to be found in Chapter 21). We see that file is not in categories, but in individual cases. So we want the test to handle proportions:

freqs = summary(rank)

chisq.test(freqs)



If you wish to see proportions, produce this, which is the same as in the text, but replacing data$religion with summary(rank).

total = sum(freqs)

observed = chisq.test(freqs)$observed

proportions = observed/total

counter = 0 # Initialisation

for (n in summary(rank)) {

cat(n, "\n")

counter = counter + 1

cat(proportions[counter], "\n")

}

rm(counter) # Clean up afterwards



This provides the proportions, out of 1, beneath each category frequency.

chisq.test(summary(rank))$expected



The results show the observed and expected counts for each professorial rank. Since we are testing for equal proportions, the expected count is 132 for each level. However, it can be seen that the Prof level has a higher observed than expected count while both AsstProf and AssocProf level have low observed scores compared to the expected counts.

We have a large Chi squared statistic and a low *p* value which is less than 0.001. Thus, we reject the null hypothesis that there is no significant difference between the expected proportions.

Let's do a visualisation:

barplot(summary(rank), names.arg = colnames(rank),

sub="The numbers come first in the formula,

then the grouping")

