Handout 11. Tabular data

1. Single proportions

Tests of single proportions are generally based on the binomial distribution with size parameter N and probability parameter p. For large sample sizes, this can be well approximated by a normal distribution with mean Np and variance Np(1 − p). As a rule of thumb, the approximation is satisfactory when the expected numbers of "successes" and "failures" are both larger than 5.

Denoting the observed number of "successes" by x, the test for the hypothesis that $p = p_0$ can be based on $u = \frac{x-Np_0}{\sqrt{Np_0(1-p_0)}}$ which has an approximate normal distribution with mean zero and standard deviation 1. Or on u^2, which has an approximate chi-square distribution with 1 degree of freedom.

We consider an example (Altman, 1991, p. 230) where 39 of 215 randomly chosen patients are observed to have asthma and one wants to test the hypothesis that the probability of a "random patient" having asthma is 0.15. This can be done using *prop.test*:

```
> prop.test(39, 215, 0.15)

        1-sample proportions test with continuity correction

data:  39 out of 215, null probability 0.15
X-squared = 1.425, df = 1, p-value = 0.2326
alternative hypothesis: true p is not equal to 0.15
95 percent confidence interval:
 0.1335937 0.2408799
sample estimates:
        p
0.1813953

 > binom.test(39, 215, 0.15)

        Exact binomial test

data:  39 and 215
number of successes = 39, number of trials = 215, p-value =
0.2135
alternative hypothesis: true probability of success is not equal
to 0.15
95 percent confidence interval:
 0.1322842 0.2395223
sample estimates:
probability of success
             0.1813953
```

2. r X c tables

For the analysis of tables with more than two classes on both sides, you can use chisq.test or fisher.test, although you should note that the latter can be very computationally demanding if the cell counts are large and there are more than two rows or columns. An r × c table looks like this:

$$\begin{array}{cccc|c}
n_{11} & n_{12} & \cdots & n_{1c} & n_{1\cdot} \\
n_{21} & n_{22} & \cdots & n_{2c} & n_{2\cdot} \\
\vdots & \vdots & & \vdots & \vdots \\
n_{r1} & n_{r2} & \cdots & n_{rc} & n_{r\cdot} \\
\hline
n_{\cdot 1} & n_{\cdot 2} & \cdots & n_{\cdot c} & n_{\cdot\cdot}
\end{array}$$

Such a table can arise from several different sampling plans, and the notion of “no relation between rows and columns” is correspondingly different. The total in each row might be fixed in advance, and you would be interested in testing whether the distribution over columns is the same for each row, or vice versa if the column totals were fixed. It might also be the case that only the total number is chosen and the individuals are grouped randomly according to the row and column criteria. In the latter case, you would be interested in testing the hypothesis of statistical independence, that the probability of an individual falling into the ijth cell is the product of the marginal probabilities. However, the analysis of the table turns out to be the same in all cases.

If there is no relation between rows and columns, then you would expect to have the following cell values:

$$E_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n_{\cdot\cdot}}$$

This can be interpreted as distributing each row total according to the proportions in each column (or vice versa) or as distributing the grand total according to the products of the row and column proportions. The test statistic

$$X^2 = \sum \frac{(O - E)^2}{E}$$

has an approximate chi-squared distribution with (r − 1) × (c − 1) degrees of freedom. Here the sum is over the entire table and the ij indices have been omitted. O denotes the observed values and E the expected values as described above.

```
caff.marital <-matrix(c(652,1537,598,242,36,46,38,21,218,327,
106,67), nrow=3,byrow=T)
 colnames(caff.marital) <- c("0","1-150","151-300",">300")
 rownames(caff.marital) <- c("Married","Prev.married","Single")
 caff.marital
```

```
               0 1-150 151-300 >300
Married      652  1537     598  242
Prev.married  36    46      38   21
Single       218   327     106   67
```

```
chisq.test(caff.marital)

        Pearson's Chi-squared

test  data:    caff.marital
X-squared = 51.6556, df = 6, p-value = 2.187e-09
```

The test is highly significant, so we can safely conclude that the data contradict the hypothesis of independence. However, you would generally also like to know the nature of the deviations. To that end, you can look at some extra components of the return value of chisq.test.

```
chisq.test(caff.marital)$expected


                     0      1-150   151-300      >300
Married      705.83179 1488.01183 578.06533 257.09105
Prev.married  32.85648   69.26698  26.90895  11.96759
Single       167.31173  352.72119 137.02572  60.94136
> chisq.test(caff.marital)$observed
               0 1-150 151-300 >300
Married      652  1537     598  242
Prev.married  36    46      38   21
Single       218   327     106   67
 E <- chisq.test(caff.marital)$expected
 O <- chisq.test(caff.marital)$observed
 (O-E)^2/E
```

```
data(juul)
attach(juul)
chisq.test(tanner,sex)
    Pearson's Chi-squared test

data:  tanner and sex
X-squared = 28.867, df = 4, p-value = 8.318e-06
table(tanner,sex)
```

It may not really be relevant to test for independence between these particular variables. The definition of Tanner stages is gender-dependent by nature.