8

Tabular data

This chapter describes a series of functions designed to analyze tabular data. Specifically, we look at the functions `prop.test`, `binom.test`, `chisq.test`, and `fisher.test`.

8.1 Single proportions

Tests of single proportions are generally based on the binomial distribution (see Section 3.3) 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 χ^2 distribution with 1 degree of freedom.

P. Dalgaard, *Introductory Statistics with R*,
DOI: 10.1007/978-0-387-79054-1_8, © Springer Science+Business Media, LLC 2008

The normal approximation can be somewhat improved by the *Yates correction*, which shrinks the observed value by half a unit towards the expected value when calculating u.

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,.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
```

The three arguments to `prop.test` are the number of positive outcomes, the total number, and the (theoretical) probability parameter that you want to test for. The latter is 0.5 by default, which makes sense for symmetrical problems, but this is not the case here. The amount 15% is a bit synthetic since it is rarely the case that one has a specific a priori value to test for. It is usually more interesting to compute a confidence interval for the probability parameter, such as is given in the last part of the output. Notice that we have a slightly unfortunate double usage of the symbol p as the probability parameter of the binomial distribution and as the test probability or p-value.

You can also use `binom.test` to obtain a test in the binomial distribution. In that way, you get an exact test probability, so it is generally preferable to using `prop.test`, but `prop.test` can do more than testing single proportions. The procedure to obtain the p-value is to calculate the point probabilities for all the possible values of x and sum those that are less than or equal to the point probability of the observed x.

```
> binom.test(39,215,.15)

        Exact binomial test

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

```
probability of success
            0.1813953
```

The "exact" confidence intervals at the 0.05 level are actually constructed from the two one-sided tests at the 0.025 level. Finding an exact confidence interval using two-sided tests is not a well-defined problem (see Exercise 8.5).

8.2 Two independent proportions

The function `prop.test` can also be used to compare two or more proportions. For that purpose, the arguments should be given as two vectors, where the first contains the number of positive outcomes and the second the total number for each group.

The theory is similar to that for a single proportion. Consider the difference in the two proportions $d = x_1/N_1 - x_2/N_2$, which will be approximately normally distributed with mean zero and variance $V_p(d) = (1/N_1 + 1/N_2) \times p(1-p)$ if the counts are binomially distributed with the same p parameter. So to test the hypothesis that $p_1 = p_2$, plug the common estimate $\hat{p} = (x_1 + x_2)/(n_1 + n_2)$ into the variance formula and look at $u = d/\sqrt{V_{\hat{p}}(d)}$, which approximately follows a standard normal distribution, or look at u^2, which is approximately $\chi^2(1)$-distributed. A Yates-type correction is possible, but we skip the details.

For illustration, we use an example originally due to Lewitt and Machin (Altman, 1991, p. 232):

```
> lewitt.machin.success <- c(9,4)
> lewitt.machin.total <- c(12,13)
> prop.test(lewitt.machin.success,lewitt.machin.total)

        2-sample test for equality of proportions with continuity
        correction

data:  lewitt.machin.success out of lewitt.machin.total
X-squared = 3.2793, df = 1, p-value = 0.07016
alternative hypothesis: two.sided
95 percent confidence interval:
 0.01151032 0.87310506
sample estimates:
   prop 1    prop 2
0.7500000 0.3076923
```

The confidence interval given is for the *difference* in proportions. The theory behind its calculation is similar to that of the test, but there are some technical complications, and a different approximation is used.

You can also perform the test without the Yates continuity correction. This is done by adding the argument `correct=F`. The continuity correction makes the confidence interval somewhat wider than it would otherwise be, but notice that it nevertheless does not contain zero. Thus, the confidence interval is contradicting the test, which says that there is *no* significant difference between the two groups with a two-sided test. The explanation lies in the different approximations, which becomes important for tables as sparse as the present one.

If you want to be sure that at least the p-value is correct, you can use Fisher's exact test. We illustrate this using the same data as in the preceding section. The test works by making the calculations in the conditional distribution of the 2×2 table given both the row and column marginals. This can be difficult to envision, but think of it like this: Take 13 white balls and 12 black balls (success and failure, respectively), and sample the balls without replacement into two groups of sizes 12 and 13. The number of white balls in the first group obviously defines the whole table, and the point is that its distribution can be found as a purely combinatorial problem. The distribution is known as the *hypergeometric distribution*.

The relevant function is `fisher.test`, which requires that data be given in matrix form. This is obtained as follows:

```
> matrix(c(9,4,3,9),2)
     [,1] [,2]
[1,]    9    3
[2,]    4    9

> lewitt.machin <- matrix(c(9,4,3,9),2)
> fisher.test(lewitt.machin)

         Fisher's Exact Test for Count Data

data:  lewitt.machin
p-value = 0.04718
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.9006803 57.2549701
sample estimates:
odds ratio
  6.180528
```

Notice that the second column of the table needs to be the number of negative outcomes, not the total number of observations.

Notice also that the confidence interval is for the *odds ratio*; that is, for the estimate of $(p_1/(1-p_1))/(p_2/(1-p_2))$. One can show that if the ps are not identical, then the conditional distribution of the table depends only on the odds ratio, so it is the natural measure of association in connection with the Fisher test. The exact distribution of the test can be worked out also when the odds ratio differs from 1, but there is the same complication as with `binom.test` that a two-sided 95% confidence interval must be pasted together from two one-sided 97.5% intervals. This leads to the opposite inconsistency as with `prop.test`: The test is (barely) significant, but the confidence interval for the odds ratio includes 1.

The standard χ^2 test (see also Section 8.4) in `chisq.test` works with data in matrix form, like `fisher.test` does. For a 2×2 table, the test is exactly equivalent to `prop.test`.

```
> chisq.test(lewitt.machin)

      Pearson's Chi-squared test with Yates' continuity
      correction

data:  lewitt.machin
X-squared = 3.2793, df = 1, p-value = 0.07016
```

8.3 k proportions, test for trend

Sometimes you want to compare more than two proportions. In that case, the categories are often ordered so that you would expect to find a decreasing or increasing trend in the proportions with the group number.

The example used in this section concerns data from a group of women giving birth where it was recorded whether the child was delivered by caesarean section and what shoe size the mother used (Altman, 1991, p. 229).

The table looks like this:

```
> caesar.shoe
    <4  4 4.5  5 5.5  6+
Yes  5  7   6  7   8  10
No  17 28  36 41  46 140
```

To compare $k > 2$ proportions, another test based on the normal approximation is available. It consists of the calculation of a weighted sum of squared deviations between the observed proportions in each group and the overall proportion for all groups. The test statistic has an approximate χ^2 distribution with $k-1$ degrees of freedom.

To use `prop.test` on a table like `caesar.shoe`, we need to convert it to a vector of "successes" (which in this case is close to being the opposite) and a vector of "trials". The two vectors can be computed like this:

```
> caesar.shoe.yes <- caesar.shoe["Yes",]
> caesar.shoe.total <- margin.table(caesar.shoe,2)
> caesar.shoe.yes
 <4   4 4.5   5 5.5  6+
  5   7   6   7   8  10
> caesar.shoe.total
 <4   4 4.5   5 5.5  6+
 22  35  42  48  54 150
```

Thereafter it is easy to perform the test:

```
> prop.test(caesar.shoe.yes,caesar.shoe.total)
         6-sample test for equality of proportions without
         continuity correction

data:  caesar.shoe.yes out of caesar.shoe.total
X-squared = 9.2874, df = 5, p-value = 0.09814
alternative hypothesis: two.sided
sample estimates:
    prop 1     prop 2     prop 3     prop 4     prop 5     prop 6
0.22727273 0.20000000 0.14285714 0.14583333 0.14814815 0.06666667

Warning message:
In prop.test(caesar.shoe.yes, caesar.shoe.total) :
  Chi-squared approximation may be incorrect
```

It is seen that the test comes out nonsignificant, but the subdivision is really unreasonably fine in view of the small number of caesarean sections. Notice, by the way, the warning about the χ^2 approximation being dubious, which is prompted by some cells having an expected count less than 5.

You can test for a trend in the proportions using `prop.trend.test`. It takes three arguments: `x`, `n`, and `score`. The first two of these are exactly as in `prop.test`, whereas the last one is the score given to the groups, by default simply $1, 2, \ldots, k$. The basis of the test is essentially a weighted linear regression of the proportions on the group scores, where we test for a zero slope, which becomes a χ^2 test on 1 degree of freedom.

```
> prop.trend.test(caesar.shoe.yes,caesar.shoe.total)

      Chi-squared Test for Trend in Proportions

data:  caesar.shoe.yes out of caesar.shoe.total ,
 using scores: 1 2 3 4 5 6
X-squared = 8.0237, df = 1, p-value = 0.004617
```

So if we assume that the effect of shoe size is linear in the group score, *then* we can see a significant difference. This kind of assumption should not be thought of as something that must hold for the test to be valid. Rather, it indicates the rough type of alternative to which the test should be sensitive.

The effect of using a trend test can be viewed as an approximate subdivision of the test for equal proportions ($\chi^2 = 9.29$) into a contribution from the linear effect ($\chi^2 = 8.02$) on 1 degree of freedom and a contribution from deviations from the linear trend ($\chi^2 = 1.27$) on 4 degrees of freedom. So you could say that the test for equal proportions is being diluted or wastes degrees of freedom on testing for deviations in a direction we are not really interested in.

8.4 $r \times c$ 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. We have already seen `chisq.test` in a simple example, but with larger tables, some additional features are of interest.

An $r \times c$ 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 $p_{i\cdot}p_{\cdot j}$ 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 χ^2 distribution with $(r-1) \times (c-1)$ 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.

We consider the table with caffeine consumption and marital status from Section 4.5 and compute the χ^2 test:

```
> 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`.

Notice that `chisq.test` (just like `lm`) actually returns more information than what is commonly printed:

```
> 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
```

These two tables may then be scrutinized to see where the differences lie. It is often useful to look at a table of the contributions from each cell to the total χ^2. Such a table cannot be directly extracted, but it is easy to calculate:

```
> E <- chisq.test(caff.marital)$expected
> O <- chisq.test(caff.marital)$observed
> (O-E)^2/E
                     0    1-150   151-300      >300
Married       4.1055981 1.612783 0.6874502 0.8858331
Prev.married  0.3007537 7.815444 4.5713926 6.8171090
Single       15.3563704 1.875645 7.0249243 0.6023355
```

There are some large contributions, particularly from too many "abstaining" singles, and the distribution among previously married is shifted in the direction of a larger intake — insofar as they consume caffeine at all. Still, it is not easy to find a simple description of the deviation from independence in these data.

You can also use `chisq.test` directly on raw (untabulated) data, here using the `juul` data set from Section 4.5:

```
> attach(juul)
> chisq.test(tanner,sex)

      Pearson's Chi-squared test

data:  tanner and sex
X-squared = 28.8672, df = 4, p-value = 8.318e-06
```

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

8.5 Exercises

8.1 Reconsider the situation of Exercise 3.3, where 10 consecutive patients had operations without complications and the expected rate was

20%. Calculate the relevant one-sided test in the binomial distribution. How large a sample (still with zero complications) would be necessary to obtain statistical significance?

8.2 In 747 cases of "Rocky Mountain spotted fever" from the western United States, 210 patients died. Out of 661 cases from the eastern United States, 122 died. Is the difference statistically significant? (See also Exercise 13.4.)

8.3 Two drugs for the treatment of peptic ulcer were compared (Campbell and Machin, 1993, p. 72). The results were as follows:

	Healed	Not Healed	Total
Pirenzepine	23	7	30
Trithiozine	18	13	31
Total	41	20	61

Compute the χ^2 test and Fisher's exact test and discuss the difference. Find an approximate 95% confidence interval for the difference in healing probability.

8.4 (From "Mathematics 5" exam, University of Copenhagen, Summer 1969.) From September 20, 1968, to February 1, 1969, an instructor consumed 254 eggs. Every day, he recorded how many eggs broke during boiling so that the white ran out and how many cracked so that the white did not run out. Additionally, he recorded whether the eggs were size A or size B. From February 4, 1969, until April 10, 1969, he consumed 130 eggs, but this time he used a "piercer" to create a small hole in the egg to prevent breaking and cracking. The results were as follows:

Period	Size	Total	Broken	Cracked
Sept. 20–Feb. 1	A	54	4	8
Sept. 20–Feb. 1	B	200	15	28
Feb. 4–Apr. 10	A	60	4	9
Feb. 4–Apr. 10	B	70	1	7

Investigate whether or not the piercer seems to have had an effect.

8.5 Make a plot of the two-sided p-value for testing that the probability parameter is x when the observations are 3 successes in 15 trials for x varying from 0 to 1 in steps of 0.001. Explain what makes the definition of a two-sided confidence interval difficult.