Dr Isabella Gollini

Probability and Statistics

14 Contingency Tables

14.1 Example — data on midges in streams

Midges of the genus *Dixa* are found in running water and are indicators of surface pollutants. A biologist investigating habitat preferences of four species of *Dixa* (*D. nebulosa*, *D. submaculata*, *D. dilatata*, *D. nubilipennis*) obtains a random selection of pupae from three streams with different levels of pollution by chemical nutrients (oligotrophic, mesotrophic and eutrophic). He records the numbers of pupae from each of the species that he finds in each stream. The observed frequencies are shown in Table 1. This is what is known as a *contingency table*, in this case a 3×4 contingency table.

	D. nebulosa	*D. submaculata*	*D. dilatata*	*D. nubilipennis*	Totals
Site 1: oligotrophic	12	7	5	17	41
Site 2: mesotrophic	14	6	22	9	51
Site 3: eutrophic	35	12	7	11	65
Totals	61	25	34	37	157

Table 1: Observed frequencies of *Dixa* species

The biologist wishes to determine from these data if there is evidence of *association* between species and habitat: do some of the species tend to be associated with particular levels of pollution?

14.2 Two-way classification

To put the example of Section 14.1 in a more general setting, suppose that we have a sample of total size n, where each observation is classified according to two different criteria. In other words, we have *bivariate* data, where each observation has associated with it two categorical variables. Suppose that the first variable has r categories and the second has s categories.

The data may be summarized in a $r \times s$ contingency table of observed frequencies

$$O_{ij} \qquad (i = 1, 2, \dots r, j = 1, 2, \dots s),$$

as shown in Table 2, where O_{ij} is the observed count of sample members who are classified into category i according to the first variable and into category j according to the second variable.

O_{11}	O_{12}	$\dots$	O_{1s}	R_1
O_{21}	O_{22}	$\dots$	O_{2s}	R_2
$\vdots$		$\ddots$	$\vdots$	$\vdots$
O_{r1}	O_{r2}	$\dots$	O_{rs}	R_r
C_1	C_2	$\dots$	C_s	n

Table 2: Table of observed frequencies

Let R_i be the ith row total, the total number of individuals in the sample who are classified into category i according to the first variable ($i = 1, 2, \dots r$) and let C_j be the jth column total, the total number of individuals in the sample who are classified into category j according to the second variable ($j = 1, 2, \dots s$). So we have that

$$\begin{aligned}
\sum_{j=1}^{s} O_{ij} &= R_i \ , \\
\sum_{i=1}^{r} O_{ij} &= C_j \ , \\
\sum_{i=1}^{r} R_i = \sum_{j=1}^{s} C_j &= n.
\end{aligned}$$

In many cases, such data are obtained when random samples are taken from a number of different populations. The first (row) variable, say, then indicates from which population an individual is drawn, in which case R_i is the sample size from the ith population. The second (column) variable will then represent a characteristic of each sample member that is being recorded. It may then be that the sample size R_i from each population is fixed in advance by the design of the sampling experiment, but the C_j are random variables, depending on the outcome of the experiment. On other occasions neither the row totals R_i nor the column totals C_j are fixed in advance, but both depend on the outcome of the experiment. In either case, the form of the analysis will be the same, but, in developing the concepts that underlie the method of analysis, it will be easier to think in terms of the second scenario, where both variables are random.

The basic null hypothesis H_0 that we shall be testing is that there is no association between the two variables, i.e., that they are independent. The alternative hypothesis H_1 is that there is an association between the two variables, i.e., they are not independent.

For a randomly chosen individual, let the random variable X denote the value of the first (row) variable and let the random variable Y denote the value of the second (column) variable. Write

$$\Pr(X = i) = p_i \qquad (i = 1, 2, \dots, r),$$

where $\sum_{i=1}^{r} p_i = 1$, and

$$\Pr(Y = j) = \pi_j \qquad (j = 1, 2, \dots, s),$$

where $\sum_{j=1}^{s} \pi_j = 1$. Under the null hypothesis of independence, it follows that for a randomly chosen individual

$$\Pr(X = i, Y = j) = p_i \pi_j \qquad (i = 1, 2, \ldots, r; j = 1, 2, \ldots, s), \tag{1}$$

which is the (i, j)th cell probability as illustrated in Table 3.

$p_1\pi_1$	$p_1\pi_2$	$\ldots$	$p_1\pi_s$	p_1
$p_2\pi_1$	$p_2\pi_2$	$\ldots$	$p_2\pi_s$	p_2
$\vdots$		$\ddots$	$\vdots$	$\vdots$
$p_r\pi_1$	$p_r\pi_2$	$\ldots$	$p_r\pi_s$	p_r
π_1	π_2	$\ldots$	π_s	1

Table 3: Probability structure under the null hypothesis of independence

14.3 The chi-square test of independence

The test statistic that we use for testing the null hypothesis of independence is a chi-square statistic, similar in form to those used in Sections 12 and 13. Here

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} , \tag{2}$$

where E_{ij} is the expected frequency for the (i, j)th cell under the null hypothesis. If the p_i and the π_j were known, using Equation (1), we would have $E_{ij} = np_i\pi_j$. As the p_i and the π_j are unknown, we use the natural estimates, the corresponding sample proportions,

$$\begin{aligned} \hat{p}_i &= \frac{R_i}{n} \qquad (i = 1, 2, \ldots, r), \\ \hat{\pi}_j &= \frac{C_j}{n} \qquad (j = 1, 2, \ldots, s). \end{aligned}$$

It follows that

$$E_{ij} = n\hat{p}_i\hat{\pi}_j = n\frac{R_i}{n}\frac{C_j}{n} ,$$

i.e.,

$$E_{ij} = \frac{R_i C_j}{n} \qquad (i = 1, 2, \ldots, r; j = 1, 2, \ldots, s). \tag{3}$$

- Note that the expected frequencies E_{ij} have the same row and column totals as the observed frequencies O_{ij}:

$$\begin{aligned} \sum_{j=1}^{s} E_{ij} &= R_i , \\ \sum_{i=1}^{r} E_{ij} &= C_j . \end{aligned}$$

Recall from Section 13 that the degrees of freedom ν of the chi-square statistic are given by

$$\nu = k - 1 - d,$$

where k is the number of cells and d is the number of fitted parameters. In the present case, there are rs cells, so $k = rs$. We have estimated the p_i $(i = 1, 2, \ldots, r)$ and the π_j $(j = 1, 2, \ldots, s)$, $r + s$ parameters in all. However, once we have estimated the first $r-1$ of the p_i, the final one is determined by the condition that $\sum_{i=1}^{r} p_i = 1$. So, in effect, we have only estimated $r-1$ parameters. Similarly, in effect, we have only estimated $s-1$ of the parameters π_j. So the effective number of estimated parameters is $(r-1)+(s-1)$ and the degrees of freedom are given by

$$\nu = rs - 1 - (r-1) - (s-1),$$

i.e.,

$$\nu = (r-1)(s-1). \tag{4}$$

The formulae of equations (2), (3) and (4) provide us with all the information needed to calculate the chi-square statistic and evaluate its significance.

- Another way of thinking of the degrees of freedom is that, given the row and column totals of the contingency table, once we have filled in an $(r-1) \times (s-1)$ block of observed frequencies or expected frequencies, the remaining frequencies are automatically determined.

Example — data on midges in streams (continued)

We could calculate the chi-square statistic manually, using a calculator, but here we move straight into `R`. The simplest way of carrying out the chi-square test involves first entering the tabulated data of observed frequencies O_{ij} into a matrix `midges`. This has been done, naming the columns with the names of the four species, `nebulosa`, `submaculata`, `dilatata`, `nubilipennis`. And naming the rows with the types of the three sites, `Oligotrophic`, `Mesotrophic`, and `Eutrophic`.

```
midges <- matrix(c(12, 7, 5, 17,
                   14, 6, 22, 9,
                   35, 12, 7, 11),
                 byrow = TRUE,
                 nrow = 3, ncol = 4)

colnames(midges) <- c("nebulosa",  "submaculata",
                      "dilatata", "nubilipennis")
rownames(midges) <- c("Oligotrophic", "Mesotrophic", "Eutrophic")
midges

##              nebulosa submaculata dilatata nubilipennis
## Oligotrophic       12           7        5           17
## Mesotrophic        14           6       22            9
## Eutrophic          35          12        7           11
```

To carry out the chi-square test of independence, you can use the function `chisq.test`.

```
test <- chisq.test(midges)
test

##
##  Pearson's Chi-squared test
##
## data:  midges
## X-squared = 30.954, df = 6, p-value = 2.586e-05
```

The value of the chi-square statistic in the present case is 30.954. There are 6 degrees of freedom, as may also be seen from equation (4) with $r = 3$ and $s = 4$. The p-value, which is zero to 3 decimal places, is highly significant. It follows that we reject the null hypothesis of independence. There is very strong evidence of association between species and type of habitat.

From the output of `chisq.test` we can obtain the values of the expected frequencies E_{ij}.

- The E_{ij} have been calculated using the formula of Equation (3). For example,

$$E_{11} = \frac{41 \times 61}{157} = 15.93\ .$$

```
expected <- test$expected
expected

##                nebulosa submaculata  dilatata nubilipennis
## Oligotrophic 15.92994    6.528662  8.878981      9.66242
## Mesotrophic  19.81529    8.121019 11.044586     12.01911
## Eutrophic    25.25478   10.350318 14.076433     15.31847
```

We can also calculate the corresponding contribute in the calculation of the chi-square statistic,

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}\ ,$$

which may be regarded as a measure of the discrepancy between O_{ij} and E_{ij}.

```
contributions <- (midges - expected)^2 / expected
contributions

##                nebulosa submaculata  dilatata nubilipennis
## Oligotrophic 0.9695205  0.03402827  1.694619     5.572111
## Mesotrophic  1.7066399  0.55396028 10.866962     0.758377
## Eutrophic    3.7604517  0.26293386  3.557429     1.217432
```

We can check that the sum of these contributions gives the value of the chi-square statistic, which in the present case is 30.954.

```
sum(contributions)

## [1] 30.95446
```

Having arrived at the conclusion that there is very strong evidence of association, we should then investigate the nature of the association. A systematic way to investigate the nature of the association is to look in particular at those cells which provide the largest contributions to the value of the chi-square statistic.

The biggest contribution, 10.867, comes from the cell (2,3), where the observed frequency is about twice the expected frequency. The species *D. dilatata* is particularly strongly represented at the mesotrophic site 2, i.e., *D. dilatata* is positively associated with the mesotrophic site. The next biggest contribution, 5.572, is for the cell (1,4): the species *D. nubilipennis* is positively associated with the oligotrophic site 1. There are also large contributions to the chi-square statistic in cells (3,1) and (3,3). The species *D. nebulosa* is positively associated with the eutrophic site 3. The observed frequency in cell (3,3) is about half the expected frequency: *D. dilatata* is under-represented at the eutrophic site, i.e., *D. dilatata* is negatively associated with the eutrophic site.

14.4 2×2 contingency tables

The simplest type of contingency table is a 2×2 table. One way in which such tables arise is as an alternative formulation of the problem of comparing two proportions dealt with in Section 11.4.

Example — stated voting intentions

In the example dealt with in Section 11.4, a sample of 200 voters had been taken in Constituency 1, of whom 73 said that they were planning to vote for the Conservative party at the next general election. A sample of size 100 had been taken in Constituency 2, of whom 43 said that they were planning to vote Conservative. Is there any significant evidence that the proportions of those planning to vote Conservative differ between the two constituencies?

The data may be laid out in a 2×2 contingency table, as shown in Table 4.

	Planning to vote Conservative		
	Yes	No	Totals
Constituency 1	73	127	200
Constituency 2	43	57	100
Totals	116	184	300

Table 4: Observed counts of stated voting intentions

The null hypothesis may be formulated as: "there is no association between constituency and the stated intention to vote Conservative."

Using the formula of equation (3), we may evaluate the expected frequencies under the null hypothesis. For example,

$$E_{11} = \frac{200 \times 116}{300} = 77.33 \ .$$

Instead of repeatedly using equation (3), the remaining expected frequencies can be found more simply by using the fact that the row and column totals in the table of expected frequencies must be the same as the row and column totals in the table of observed frequencies. We obtain Table 5.

	Planning to vote Conservative		
	Yes	No	Totals
Constituency 1	77.33	122.67	200
Constituency 2	38.67	61.33	100
Totals	116	184	300

Table 5: Expected counts of stated voting intentions

The chi-square statistic may then be calculated using equation (2). For each cell, $|O_{ij} - E_{ij}| = 4.33$ and hence

$$X^2 = 4.33^2 \left(\frac{1}{77.33} + \frac{1}{122.67} + \frac{1}{38.67} + \frac{1}{61.33} \right) = 1.186 \ .$$

For a 2×2 table, from the formula of equation (4), there is 1 degree of freedom associated with the chi-square statistic.

From Table 8 of *Lindley and Scott*, $\chi^2_1(5) = 3.841$. Hence, at the 5% significance level, we do not reject the null hypothesis of no association between constituency and the stated intention to vote Conservative. There is no strong evidence of any association between constituency and the stated intention to vote Conservative. (Alternatively, we could use Table 7 of *Lindley and Scott* to find the p-value of the chi-square statistic.)

The following output shows the same analysis carried out using `R`.

```
voting <- matrix(c(73, 127,
                   43, 57),
                 nrow = 2,
                 ncol = 2,
                 byrow = TRUE)
rownames(voting) <- c("Constituency 1", "Constituency 2")
colnames(voting) <- c("Cons",  "No Cons")

voting

##                  Cons No Cons
## Constituency 1   73     127
## Constituency 2   43      57
```

```
test <- chisq.test(voting,
                   correct = FALSE)
test

##
##  Pearson's Chi-squared test
##
## data:  voting
## X-squared = 1.1877, df = 1, p-value = 0.2758

test$expected

##                      Cons   No Cons
## Constituency 1 77.33333 122.66667
## Constituency 2 38.66667  61.33333
```

In fact, the above chi-square test, when used for comparing two proportions, is equivalent to the test for comparing two proportions in Section 11.4. It can be shown algebraically that the square of the z-statistic used in Section 11.4 is equal to the chi-square statistic used here. If a two-tail test is used with the z-statistic then the p-values for both statistics are identical. This can be verified numerically from the two `R` outputs for our voting intentions example. In Section 11.4, $z = -1.09$. In the both cases in the `R` output we can notice that, $X^2 = 1.188$ and $p = 0.276$.

- It follows from our discussion of the chi-square distributions in Section 8 that if $Z \sim N(0, 1)$ then $Z^2 \sim \chi^2_1$.

Extra examples

Survey

A researcher is interested in whether people are more likely to return a survey if one incentive is offered.

He sends out 100 questionnaires, he promises to 20 respondents that they will get the survey results. He says to 30 respondents that will be entered in a prize draw. 50 respondents have no incentives.

The results were as follow:

	Response NO	Response YES
Survey Results	11	9
Prize Draw	14	16
Nothing	38	12

New Drug effectiveness

In order to test the effectiveness of a new drug in treating a particular disease, 70 patients were randomly divided into two groups. The first group was treated with the new drug and the second group was treated in the standard way. The results were as follow:

	Recover	Die
Drug	20	15
No drug	13	22

Test the hypothesis that the drug has no effect.