6. Contingency tables – association of two (or more) categorical variables

Contingency tables – introduction

Contingency tables are tables that summarize frequencies (counts) of two (or more) categorical variables. Their analysis allows to test (in)dependence between the two variables. Table 6.1 is a contingency table summarizing frequencies of people of different eye and hair colors.

Table 6.1. Contingency table of two variables: eye and hair color with basic frequency statistics (marginal sums and grand total).

		Hair color			
		black	brown	blonde	marginal sums
Eye color	blue	12	45	14	71
	brown	51	256	84	391
	marginal sums	63	301	98	grand total: 462

Basic analysis by goodness-of-fit test

Association between the variables (i.e. the **null hypothesis which states that the variables are independent**) can be tested by a goodness-of-fit test. This is a universal approach suitable for tables of any size and dimensions but its explanatory power is limited.

For goodness-of-fit test, we need expected frequencies under null hypothesis which are calculated on the basis of probability theory: P(event 1 and event 2) = P (event 1) x P (event 2), if the two events are independent. In contingency tables, this can be used to calculate expected frequencies as the product of ratios of corresponding marginal totals and the grand total.

For instance, expected probability of observing a blue-eyed and black-haired person in Table 6.1 can be calculated as P(blueE and blackH) = 63/462 x 71/462 = 0.02096. Multiplication of the probability then gives the expected frequency Freq(e) = 0.02096 x 462 = 9.68.

The same approach can be used to calculate expected frequencies in all cells but is done automatically by software nowadays. Goodness-of-fit test can consequently be computed (in the same way as described in chapter 5). Note, however, that the number of degrees of freedom is determined as DF = (number of rows – 1) x (number of columns – 1)

In our example: **We did not find a significant association between eye and hair color (χ^2 = 0.785, DF = 2, p = 0.6755).**

The goodness-of-fit test does not provide much more information on the result, though in case of significant result, it may make sense to report also the difference between observed-

expected frequencies (i.e. the residuals), or their standardized values (residuals divided by square root of corresponding expected frequencies) as supplementary information. In particular, standardized residuals are useful as they indicate excess or deficiency of which combinations cause association between the variables.

2x2 tables and their analysis

These tables represent a special and the simplest cases of contingency tables (Table 6.2).

Table 6.2. Structure of a 2x2 table.

		Var2		
		level 1	level 2	
Var 1	level 1	f11	f12	R1
	level 2	f21	f22	R2
		C1	C2	***n***

Their simplicity allows additional statistics to be computed to express how tight the association between the two variables is. Most important of these is the **phi-coefficient**:

$$\varphi = \frac{f11f22 - f12f21}{\sqrt{R1R2C1C2}} = \pm\sqrt{\frac{\chi^2}{n}}$$

where f, R C symbols correspond to cells in Table 6.2 and χ^2 is the χ^2 statistics of the table and *n* is the grand total.

The phi-coefficient can thus be viewed as an average contribution of each observation to the association between the variables. This implies its important advantage which lies in comparability of the phi coefficients between datasets with unequal numbers of observations.

The 2x2 tables may seem trivial and not of much use. However, they and especially the phi-coefficient is frequently used in vegetation ecology to measure association between occurrences of two species or as a fidelity measure of a species with a vegetation unit. In that case Var1 describes frequency of given species and Var2 frequency of the vegetation unit in the dataset.

Advanced analysis of contingency tables – odds and odds ratios

Odds and odds ratios are additional important statistics that can be used to analyze contingency tables. They are defined for 2x2 tables only but can also be used in larger (in particular n x 2) tables, which can be subdivided into a series of 2x2 tables. For table 6.1, we can calculate the odds for the level 1 of Var1 as:

$odds_1 = p/(1-p) = (f11/R1)/(f12/R1)$

where p is the probability of one outcome of the second variable and 1-p is probability of the second outcome of the second variable. We can do the same for the second level of Var1 to get $odds_2$. Odds ratio then equals:

$OR = odds_1/odds_2$

Odds ratio directly indicates how probability of observing level 1 of Var1 changes with respect to the levels of Var2.

OR values range between 0 and infinity, with OR < 1 indicating negative association, OR = 1 independence and OR > 1 positive association.

OR is a population parameter and the computation summarized above is actually its maximum-likelihood estimation procedure. As a result, OR estimate has associated standard error and confidence intervals (i.e. intervals within which the population OR lies with 95% probability). A confidence interval directly indicates significance – if a confidence interval of OR contains 1, the OR is not significantly different from 1 and thus independence between the two variables cannot be rejected.

A worked example

Malaria is a dangerous disease widespread in tropical areas. It is caused by protozoans of the genus *Plasmodium* and transmitted by mosquitos. To prevent infection, it is possible to take prophylaxis, i.e. treatment which blocks the infection after mosquito bite. This is only possible for short time journeys to areas with malaria since the prophylaxis drugs are not safe for long-term use. Here we asked whether the prophylaxis is efficient and whether there is significant difference between two types of prophylaxis. The data are summarized in Table 6.3.

Table 6.3. Table summarizing frequencies of travelers to the tropics infected by malaria (or not) and anti-malaria prophylaxis they used.

Prophylaxis	Infected by malaria	Frequency
none (control)	0	40
none (control)	1	94
doxycycline	0	130
doxycycline	1	80
lariam	0	180
lariam	1	15

Note here, that contingency table can also have a form of a table with individual factor combinations and corresponding frequencies. This is actually a bit better for computation than the cross-tabulated form.

Goodness of fit test demonstrates, that there is a significant association between the two variables:

Chisq = 137.45, df = 2, p-value = 1.42e-30

Odds ratios summary then follows. Two odds ratios are produced comparing the second and third level of to the first one (here control). The “lower” and “upper” values indicate limits of confidence intervals. We can see that both types of prophylaxis are associated with significantly decreased infection rate.

```
         infected
prophylax   0        p0  1         p1  oddsratio      lower      upper      p.value
  control  40 0.1142857 94 0.49735450 1.00000000         NA         NA           NA
  doxy    130 0.3714286 80 0.42328042 0.26186579 0.16479825 0.41610692 6.790312e-09
  lariam  180 0.5142857 15 0.07936508 0.03546099 0.01862937 0.06749997 8.847446e-34
```

To compare just the two prophylaxis types, we can select just the corresponding part of the data for analysis (specifying this by square brackets in R). The result shows that taking Lariam is associated with significantly lower infection rate than taking doxycycline.

```
         infected
prophylax   0        p0  1        p1 oddsratio      lower     upper      p.value
   doxy   130 0.4193548 80 0.8421053 1.0000000         NA        NA           NA
   lariam 180 0.5806452 15 0.1578947 0.1354167 0.07462922 0.2457171 1.531487e-13
```

In a paper/thesis, the result can by summarized as Table 6.4

Table 6.4. Summary of a contingency table analysis testing the association between malaria prophylaxis and infection. Overall test of independence χ^2 = 137.45, df = 2, $p < 10^{-6}$.

	Odds ratio	lower 95% conf. limit	upper 95% conf. limit	*p*
Lariam vs. none	0.035	0.019	0.067	$< 10^{-6}$
doxycycline vs. none	0.262	0.165	0.416	$< 10^{-6}$
Lariam vs. doxycycline	0.135	0.075	0.246	$< 10^{-6}$

Coincidence and causality

Note here, that significant results of a contingency table analysis indicate significant association. This can be caused either by coincidence or causality. Causality means that if we manipulate one variable, the other also changes, i.e. one variable has a direct effect on the other. By contrast coincidence may happen due to another variable affecting the two ones analyzed. In such case, manipulation of one variable has no effect on the other in case of coincidence.

Considering the malaria example, the travelers using prophylaxis are simultaneously more likely to use mosquito repellents, which in reality can strongly decrease infection risk. Therefore, if somebody from the no-prophylaxis travelers decided to take prophylaxis, it may have much lower (or even no) effect than our analysis suggests.

People in general like causal explanations (and expect them). As a result, association is frequently interpreted as causal relationship, which is however inappropriate. Association may only suggest causality at best, which can be consequently demonstrated by a **manipulative experiment**. In our case, this would mean to select a group of people, assign them randomly into three groups according to prophylaxis, send them to the tropics and see what happens. In this particular case however, such research would not be approved by an ethics committee.

How to do in R

1. Chisq analysis of contingency tables

Option 1: apply **chisq.test** on matrix containing frequencies

Option 2: If the data are formatted in data frame as in Table 6.3, they can be converted to contingency table by function **xtabs**

```
data.table<-xtabs(freq~var1+var2, data=data.frame)
```

chisq.test can then be applied on the contingency table. If its result is saved in an object:

```
test.res<-chisq.test(data.table)
```

running **test.res$std.resid** can then be used to display standardized residuals.

2. Phi – coefficient

function **phi** (package psych) applied on a 2x2 matrix

3. Odds ratios

function **epitab** (package epitools) applied on contingency table produced by xtabs. Square brackets can be used to select the levels to compare.