Chapter 9: Statistical inference for discrete data

Contents

1 Testing independence

1.1 Raw data: use `table` + `chisq.test`

Example *chol*

In order to test whether there is an association between the smoke behavior and the mortality, we use the `chol` data. Because we work here with raw data, we first have to construct a contingency table by the `table` function.

Importing the data: (data is available on Toledo)

```
Chol <- read.table(file=file.choose(), header=TRUE)
```

```
head(Chol, n=4)
```

```
## # A tibble: 4 x 7
##     AGE HEIGHT WEIGHT  CHOL SMOKE  BLOOD MORT
##   <dbl>  <dbl>  <dbl> <dbl> <chr>  <chr> <chr>
## 1    20    176     77   195 nonsmo b     alive
## 2    53    167     56   250 sigare o     dead
## 3    44    170     80   304 sigare a     dead
## 4    37    173     89   178 nonsmo o     alive
```

```
names(Chol)
```

```
## [1] "AGE"    "HEIGHT" "WEIGHT" "CHOL"   "SMOKE"  "BLOOD"  "MORT"
```

Null hypothesis:

H_0 : **mortality and smoke behavior is independent**

Alternative hypothesis:

H_1 : **There is an association between mortality and smoke behavior**

First constructing a contingency table:

```
tab.chol <- table(Chol$SMOKE, Chol$MORT)
tab.chol
```

```
##
##          alive dead
##   nonsmo    45    4
##   pipe      38    4
##   sigare    93   16
```

```
chisq.test(tab.chol)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab.chol
## X-squared = 1.6677, df = 2, p-value = 0.4344
```

1.2 Summary data: use `xtabs` and `chisq.test`

In case you have summarized data, you work with a weight variable (often `n`).

By using the `xtabs()` function, you can use weight variables

> **`xtabs(Freq ~ var1 + var2, DF)`**

Example *seatbelt*
In order to test whether wearing seat belts prevents for having fatal accidents we have following data *seatbelt.txt*.

Import seatbelts.txt

```
head(seatbelts)
```

```
##   seatbelts fatal   n
## 1       yes   yes   7
## 2       yes    no  89
## 3        no   yes  24
## 4        no    no 122
```

```
table <- xtabs(seatbelts$n ~ seatbelts$seatbelts + seatbelts$fatal)
table
```

```
##                     seatbelts$fatal
## seatbelts$seatbelts  no yes
##                  no 122  24
##                 yes  89   7
```

> H_0 : **wearing seatbelts and having a fatal accident is independent**
> versus
> H_1 : **there is an association between wearing seatbelts and having a fatal accident**

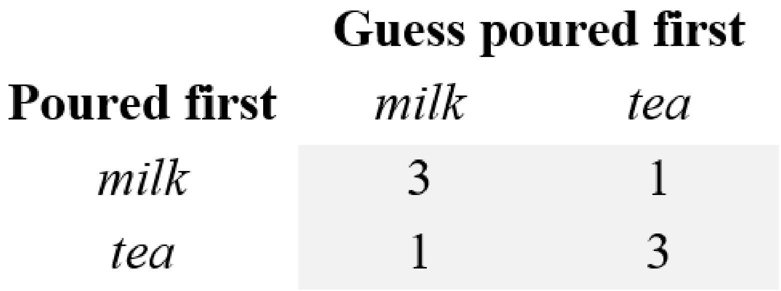
```
chisq.test(table)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table
## X-squared = 3.558, df = 1, p-value = 0.05926
```

1.3 In case of very few observations

Example *tea*
tea-drinking lady: *tea.txt*

Someone claimed that, when drinking tea, she could distinguish whether milk or tea was added to the cup first. To test her claim, she tasted 8 cups of tea. 4 cups had milk added first, and the other had tea added first. The cups were presented in random order.



	Guess poured first	
Poured first	*milk*	*tea*
milk	3	1
tea	1	3

```
tea # after importing tea.txt
```

```
##   poured guess n
## 1   milk  milk 3
## 2   milk   tea 1
## 3    tea  milk 1
## 4    tea   tea 3
```

```
table <- xtabs(tea$n ~ tea$poured + tea$guess)
table
```

```
##           tea$guess
## tea$poured milk tea
##       milk    3   1
##       tea     1   3
```

```
fisher.test(table)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table
## p-value = 0.4857
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    0.2117329 621.9337505
## sample estimates:
## odds ratio
##   6.408309
```

Remark:
In the package `rstatix`, you can as well use the functions `chisq_test()` and `fisher_test()`.

2 Some other functions for count data

Example

Two medications:

A: 121 deaths out of 1584 patients
B: 145 deaths out of 1998 patients

2.1 Test for two proportions

Test for two proportions: Do the two mortality rates differ significantly?

$H_0 : p_A = p_B$
versus
$H_1 : p_A \neq p_B$

```
trial.mort <- c(121, 145)
trial.siz <- c(1584, 1998)
prop.test(trial.mort, trial.siz)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  trial.mort out of trial.siz
## X-squared = 0.13579, df = 1, p-value = 0.7125
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.01408481  0.02171744
## sample estimates:
##     prop 1     prop 2
## 0.07638889 0.07257257
```

2.2 Test for one proportion: `binom.test`

`binom.test` does not provide confidence intervals.
This function test hypotheses about the parameter p in a $Binomial(n, p)$ model given x, the number of successes out of n trials.

Syntax:

`Binom.test(x, n, p)`

Hypotheses:

$H_0 : p = 0.08$ versus $H_1 : p \neq 0.08$

```
binom.test(121, 1584, p = 0.08)
```

```
##
##  Exact binomial test
##
## data:  121 and 1584
## number of successes = 121, number of trials = 1584, p-value = 0.6432
## alternative hypothesis: true probability of success is not equal to 0.08
## 95 percent confidence interval:
##  0.06378725 0.09058599
## sample estimates:
## probability of success
##             0.07638889
```

Remark:
In the package `rstatix`, you can use the functions `prop_test()` and `binom_test()`.

3 Exercises

1. The data for this exercise can be found in *operations.txt*.

There are 4 groups of operations (A = Not serious, B, C, D = Very serious) and 3 groups of side effects (0 = No side effects, 1 = one side effect, 2 = two side effects).

Make a table of this cross-classified data in R.

```
##                     operations$side_effect
## operations$operation  0  1  2
##                    A 61 28  7
##                    B 68 23 13
##                    C 58 40 12
##                    D 53 38 16
```

Check if the row and column variables are independent of each other.

2. The data set used in this exercise is `tips` from the `reshape` package.

Make a table of day of the week and gender of the bill payer. Check whether these two variables are independent of each other.