**Chi-squared Test of Independence**

Two random variables *x*and *y*are called **independent**if the probability distribution of one variable is not affected by the presence of another.

Assume *fij* is the observed frequency count of events belonging to both *i*-th category of *x*and *j*-th category of *y*. Also assume *eij* to be the corresponding expected count if *x*and *y*are independent. The null hypothesis of the independence assumption is to be rejected if the p-value of the following [Chi-squared](http://www.r-tutor.com/node/61) test statistics is less than a given significance level *α*.

                2
χ2 = ∑  (fij --eij)-
     i, j   eij


**Example**

In the built-in data set [survey](http://www.r-tutor.com/node/61), the **Smoke**column records the students smoking habit, while the **Exer**column records their exercise level. The allowed values in Smoke are "Heavy", "Regul" (regularly), "Occas" (occasionally) and "Never". As for Exer, they are "Freq" (frequently), "Some" and "None".

We can tally the students smoking habit against the exercise level with the table function in R. The result is called the **contingency table**of the two variables.

> library(MASS)       # load the MASS package   
> tbl = table(survey$Smoke, survey$Exer)   
> tbl                 # the contingency table   
   
        Freq None Some   
  Heavy    7    1    3   
  Never   87   18   84   
  Occas   12    3    4   
  Regul    9    1    7

**Problem**

Test the hypothesis whether the students smoking habit is independent of their exercise level at .05 significance level.

**Solution**

We apply the chisq.test function to the contingency table tbl, and found the p-value to be 0.4828.

> chisq.test(tbl)   
   
        Pearson’s Chi-squared test   
   
data:  table(survey$Smoke, survey$Exer)   
X-squared = 5.4885, df = 6, p-value = 0.4828   
   
Warning message:   
In chisq.test(table(survey$Smoke, survey$Exer)) :   
  Chi-squared approximation may be incorrect

**Answer**

As the p-value 0.4828 is greater than the .05 significance level, we do not reject the null hypothesis that the smoking habit is independent of the exercise level of the students.

**Enhanced Solution**

The warning message found in the solution above is due to the small cell values in the contingency table. To avoid such warning, we combine the second and third columns of tbl, and save it in a new table named ctbl. Then we apply the chisq.test function against ctbl instead.

> ctbl = cbind(tbl[,"Freq"], tbl[,"None"] + tbl[,"Some"])   
> ctbl   
      [,1] [,2]   
Heavy    7    4   
Never   87  102   
Occas   12    7   
Regul    9    8   
   
> chisq.test(ctbl)   
   
        Pearson’s Chi-squared test   
   
data:  ctbl   
X-squared = 3.2328, df = 3, p-value = 0.3571

<http://www.sthda.com/sthda/RDoc/data/housetasks.txt>.

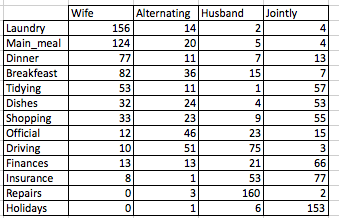
# Import the data

file\_path <- "http://www.sthda.com/sthda/RDoc/data/housetasks.txt"

housetasks <- read.delim(file\_path, row.names = 1)

# head(housetasks)

An image of the data is displayed below:



The data is a contingency table containing 13 housetasks and their distribution in the couple:

* rows are the different tasks
* values are the frequencies of the tasks done :
* by the *wife* only
* alternatively
* by the husband only
* or jointly

## Graphical display of contengency tables

Contingency table can be visualized using the function **balloonplot()** [in gplots package]. This function draws a graphical matrix where each cell contains a dot whose size reflects the relative magnitude of the corresponding component.

**library**("gplots")

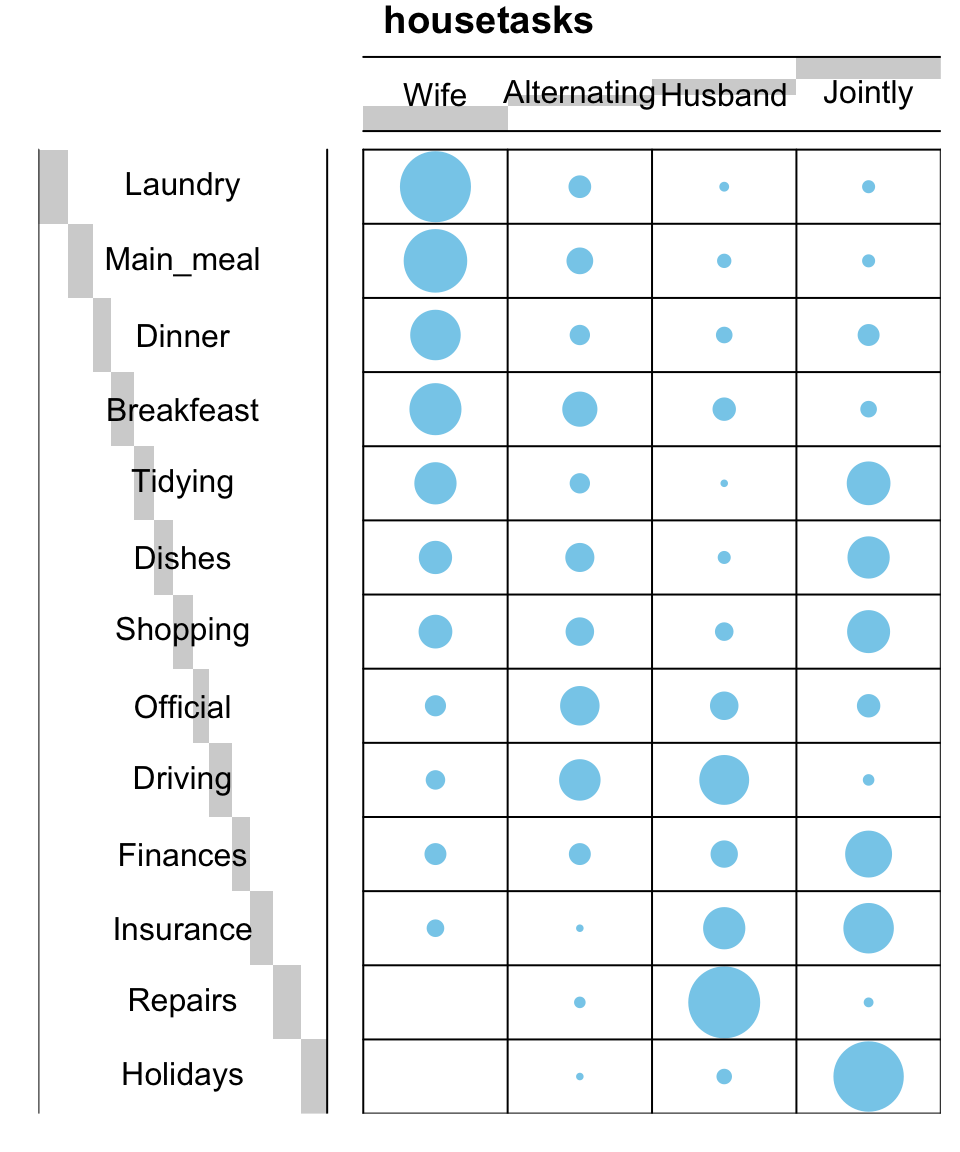
# 1. convert the data as a table

dt <- as.table(as.matrix(housetasks))

# 2. Graph

balloonplot(t(dt), main ="housetasks", xlab ="", ylab="",

label = FALSE, show.margins = FALSE)

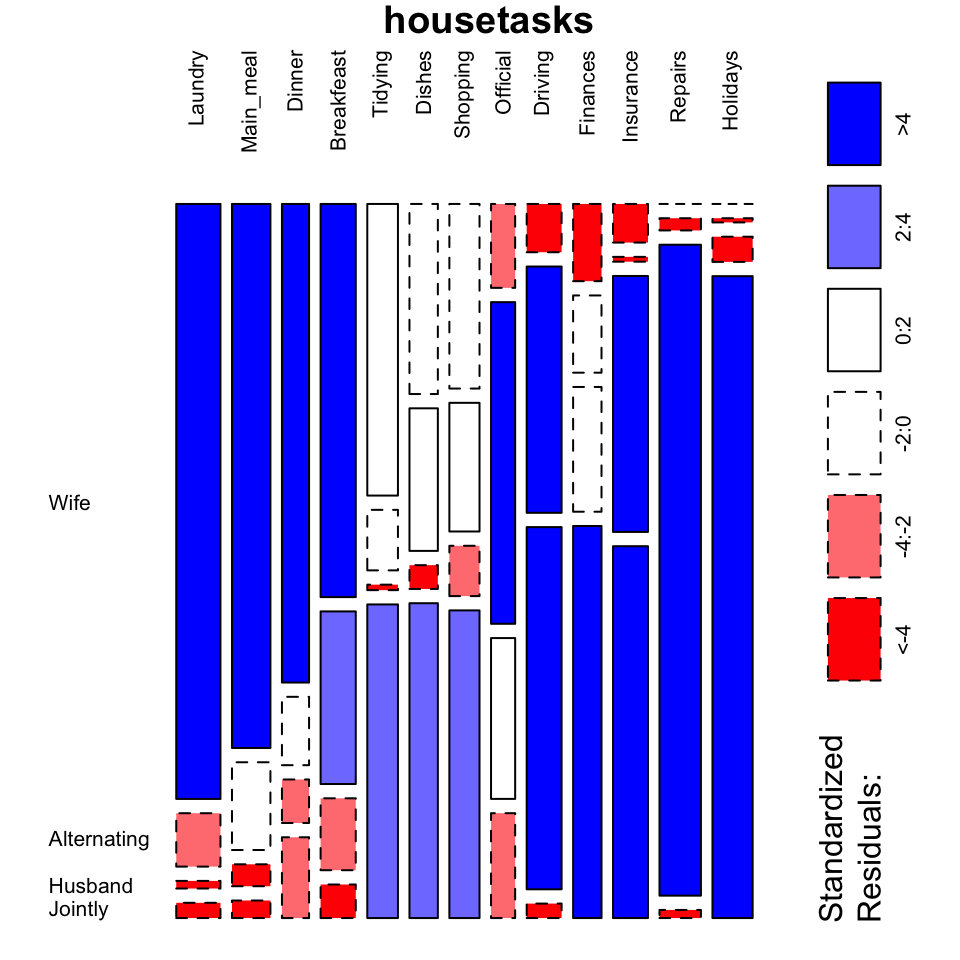


It’s also possible to visualize a contingency table as a *mosaic plot*. This is done using the function *mosaicplot*() from the built-in R package *garphics*:

**library**("graphics")

mosaicplot(dt, shade = TRUE, las=2,

main = "housetasks")



* The argument **shade** is used to color the graph
* The argument **las = 2** produces vertical labels

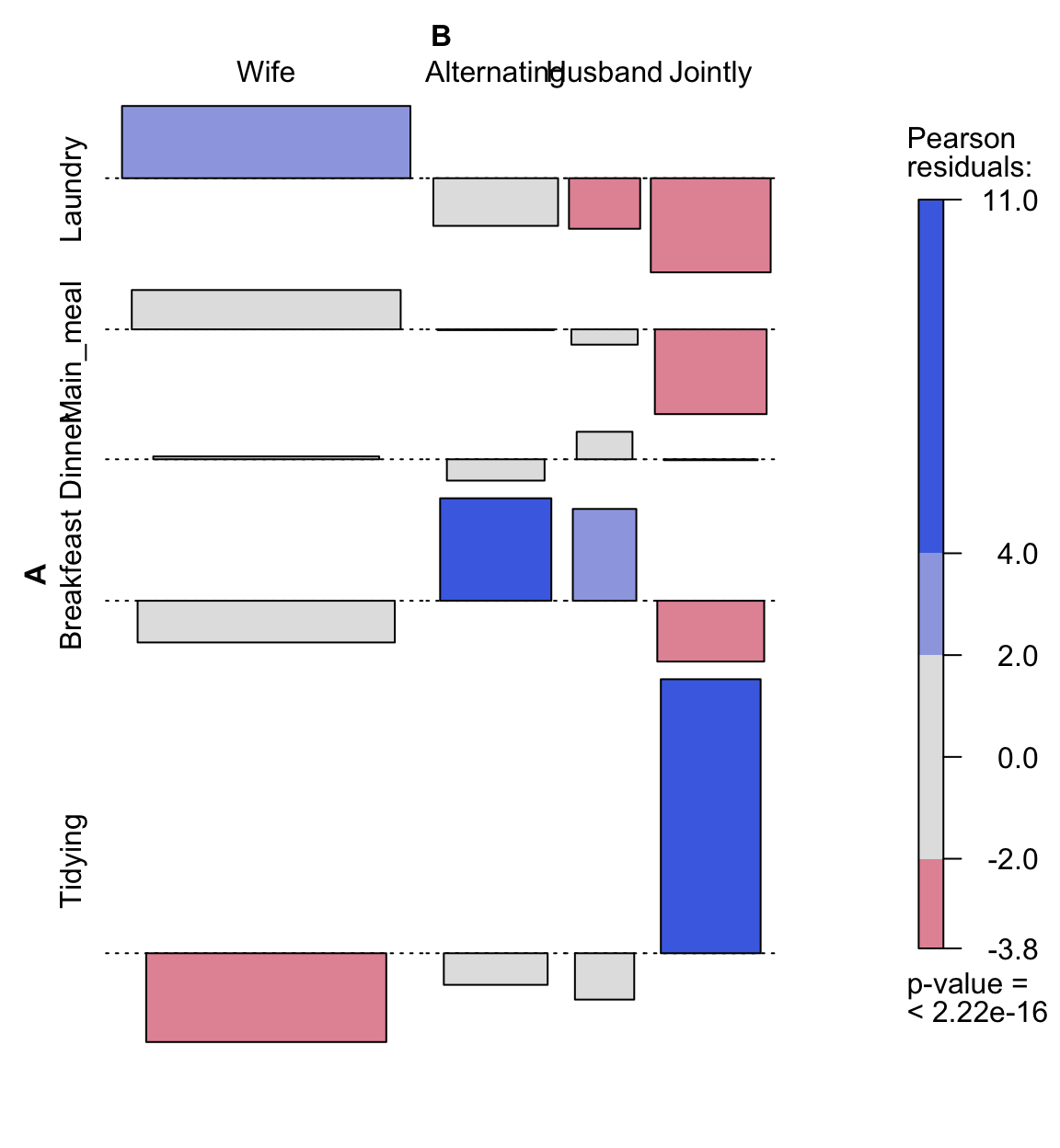
There is another package named vcd, which can be used to make a mosaic plot (function mosaic()) or an association plot (function assoc()).

# install.packages("vcd")

**library**("vcd")

# plot just a subset of the table

assoc(**head**(dt, 5), shade = TRUE, las=3)



## Chi-square test basics

**Chi-square test** examines whether rows and columns of a contingency table are statistically significantly associated.

* **Null hypothesis (H0)**: the row and the column variables of the contingency table are independent.
* **Alternative hypothesis (H1)**: row and column variables are dependent

For each cell of the table, we have to calculate the expected value under null hypothesis.

For a given cell, the expected value is calculated as follow:

e=row.sum∗col.sumgrand.totale=row.sum∗col.sumgrand.total

The Chi-square statistic is calculated as follow:

χ2=∑(o−e)2eχ2=∑(o−e)2e

* o is the observed value
* e is the expected value

This calculated Chi-square statistic is compared to the critical value (obtained from statistical tables) with df=(r−1)(c−1)df=(r−1)(c−1) degrees of freedom and p = 0.05.

* r is the number of rows in the contingency table
* c is the number of column in the contingency table

If the calculated Chi-square statistic is greater than the critical value, then we must conclude that the row and the column variables are not independent of each other. This implies that they are significantly associated

## Compute chi-square test in R

Chi-square statistic can be easily computed using the function **chisq.test()** as follow:

chisq <- chisq.test(housetasks)

chisq

Pearson's Chi-squared test

data: housetasks

X-squared = 1944.5, df = 36, p-value < 2.2e-16

The observed and the expected counts can be extracted from the result of the test as follow:

# Observed counts

chisq$observed

Wife Alternating Husband Jointly

Laundry 156 14 2 4

Main\_meal 124 20 5 4

Dinner 77 11 7 13

Breakfeast 82 36 15 7

Tidying 53 11 1 57

Dishes 32 24 4 53

Shopping 33 23 9 55

Official 12 46 23 15

Driving 10 51 75 3

Finances 13 13 21 66

Insurance 8 1 53 77

Repairs 0 3 160 2

Holidays 0 1 6 153

# Expected counts

round(chisq$expected,2)

Wife Alternating Husband Jointly

Laundry 60.55 25.63 38.45 51.37

Main\_meal 52.64 22.28 33.42 44.65

Dinner 37.16 15.73 23.59 31.52

Breakfeast 48.17 20.39 30.58 40.86

Tidying 41.97 17.77 26.65 35.61

Dishes 38.88 16.46 24.69 32.98

Shopping 41.28 17.48 26.22 35.02

Official 33.03 13.98 20.97 28.02

Driving 47.82 20.24 30.37 40.57

Finances 38.88 16.46 24.69 32.98

Insurance 47.82 20.24 30.37 40.57

Repairs 56.77 24.03 36.05 48.16

Holidays 55.05 23.30 34.95 46.70

## Nature of the dependence between the row and the column variables

As mentioned above the total Chi-square statistic is 1944.456196.