# Computer practical 1-cont: Topics in Statistics III/IV, Term 1

Georgios Karagiannis

Aim

- To check some descriptive statistics

---

## Contigency table: data manipulation

Below we load a table where refers to a 1992 survey by the Wright State University School of Medicine and the United Health Services in Dayton, Ohio. The survey asked 2276 students in their final year of high school in a nonurban area near Dayton, Ohio whether they had ever used alcohol, cigarettes, or marijuana. Denote the variables in this $2 \times 2 \times 2$ table by A for alcohol use, C for cigarette use, and M for marijuana use.

Load the observed counts in a data frame `obs.frame` and print the result. Use commands :

- `data.frame()` : as in SC2
- `factor()` : to encode a vector as a factor (aka category)
- `expand.grid()` : to produce all combinations of the supplied vectors or factors.

```
# I will do this for you

## load the data

obs.frame<-data.frame(count=c(911,538,44,456,3,43,2,279),
                      expand.grid(
  marijuana=factor(c("Yes","No"),levels=c("No","Yes")),
  cigarette=factor(c("Yes","No"),levels=c("No","Yes")),
  alcohol=factor(c("Yes","No"),levels=c("No","Yes")))
  )

## orint the obs.frame

obs.frame
```

```
##   count marijuana cigarette alcohol
## 1   911       Yes       Yes     Yes
## 2   538        No       Yes     Yes
## 3    44       Yes        No     Yes
## 4   456        No        No     Yes
## 5     3       Yes       Yes      No
## 6    43        No       Yes      No
## 7     2       Yes        No      No
## 8   279        No        No      No
```

Create 3 dimentional contingency table from `obs.frame`. Use command:

- `xtabs()`, to create a contingency table from cross-classifying factors in a dara.frame

```
# this is me again
obs.xtabs <- xtabs(count ~ marijuana+cigarette+alcohol, data=obs.frame)
## print
obs.xtabs
```

```
## , , alcohol = No
##
##          cigarette
## marijuana  No Yes
##       No  279  43
##       Yes   2   3
##
## , , alcohol = Yes
##
##          cigarette
## marijuana  No Yes
##       No  456 538
##       Yes  44 911
```

Compute the marginal contigency table of marijuana and cigarette.

- Use command `margin.table( , margin = )` and `obs.xtabs`.
- Save it in `obs.mc.xtabs`.

```
obs.mc.xtabs <- margin.table(obs.xtabs, margin=c(1,2))
obs.mc.xtabs
```

```
##          cigarette
## marijuana  No Yes
##       No  735 581
##       Yes  46 914
```

---

# Odds ratio calculations

Code an R function, named 'odds.ratio' with:

- Inputs:
  - x : a 2 by 2 matrix whose elements are the observed counts of a 2 by 2 contigency table
  - conf.level : with default input value 0.95 representing the confidence level
  - theta0 : with default value 1 representing a null hypothesis value of the odds ratio test
- Outputs:
  - estimator : representing mle of odds satio
  - log.estimator : representing the mle of log odds ratio
  - asympt.SE : representing the standard error / standard deviation of the mle of odds ratio
  - conf.interval: representing confidence interval of mle of odds ratio at sig level conf.level (from the inputs)
  - conf.level =representing confidence level
  - Ztest : representing the test statistic for the odds ratio test (2 tails)
  - p.value : representing the p value of the odds ratio test (2 tails)

– log.conf.interval : representing in log scale the confidence interval of mle of odds ratio at sig level conf.level (from the inputs)

```
odds.ratio <- function(x,conf.level=0.95,theta0=1)
{
  if (any(x==0)) x <- x+0.5
  theta <- x[1,1] *
    x[2,2]/(x[1,2]  *
              x[2,1])
  SE <- sqrt(sum(1/x))
  Za2 <- qnorm(0.5 *
                 (1+conf.level))
  Low <- exp(log(theta)-Za2 * SE)
  Up <- exp(log(theta)+Za2  *
              SE)
  CI <- c(Low,Up)
  Z=(log(theta)-log(theta0))/SE
  pv=2   *
    pnorm(-abs(Z))

  logCI <- log(CI)

  list (estimator=theta,
        log.estimator=log(theta),
        asympt.SE=SE,
        conf.interval=CI,
        conf.level=conf.level,
        Ztest=Z,
        p.value=pv,
        log.conf.interval=logCI
        )
}
```

For the marginal contigency table of marijuana and cigarette,

- compute the mle of the marginal odds ratio of marijuana and cigarette
- computet the 95% Confidence Interval of the marginal odds ratio of marijuana and cigarette
- perform a statiastical hypothesis test that marijuana and cigarette are independent at sig level 0.05

```
obs.mc.xtabs <- margin.table(obs.xtabs, margin=c(1,2))

odds.ratio.marijuana.cigarette <- odds.ratio(obs.mc.xtabs, conf.level=0.95, theta0=1 )
odds.ratio.marijuana.cigarette
```

```
## $estimator
## [1] 25.1362
##
## $log.estimator
## [1] 3.224309
##
## $asympt.SE
## [1] 0.1609812
##
```

```
## $conf.interval
## [1] 18.33463 34.46093
##
## $conf.level
## [1] 0.95
##
## $Ztest
## [1] 20.02911
##
## $p.value
## [1] 3.071215e-89
##
## $log.conf.interval
## [1] 2.908792 3.539826
```

The MLE of the marginal odds ratio of marijuana and cigarette is 25.136197 .

The 95% confidence interval of the marginal odds ratio of marijuana and cigarette is [18.33463, 34.4609298] .

The hypothesis test with $H_0 : \theta = 1$ versus $H_1 : \theta \neq 1$ at sig. level 0.05 has a p-value $3.0712155 \times 10^{-89}$ Hence I reject $H_0 : \theta = 1$ against $H_1 : \theta \neq 1$ at sig. level 0.05.

---

## Fourfold Plots

You can draw Fourfold Plots

### Tables 2 x 2

It is a graphical expression visualizing the odds ratio

$$\theta = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

in 2 x 2 contingency tables.

It shows the departure from independence as measured by the sample odds ratio,

Each cell $n_{ij}$ is represented as a quarter-circle with radius proportional to $\sqrt{n_{ij}}$ and area proportional to $n_{ij}$.

- If there is no association $\theta = 1$ between classification variables, the quarter-circles should form a circle.
- If there is positive association $\theta > 1$ between classification variables, the diagonal areas are greater than the off-diagonal areas
- If there is negative association $\theta < 1$ between classification variables, the diagonal areas are smaller than the off-diagonal areas

R provides a function to draw this kind of plots by using the function `fourfoldplot' from the package`vcd'

  - Install 'vcd' package and load it

```
# install.packages('vcd') # IF NOT ALREADY INSTALLED ON YOUR PC, THEN UNCOMMENT AND RUN THIS COMMAND
library(vcd)
```

```
## Loading required package: grid
```

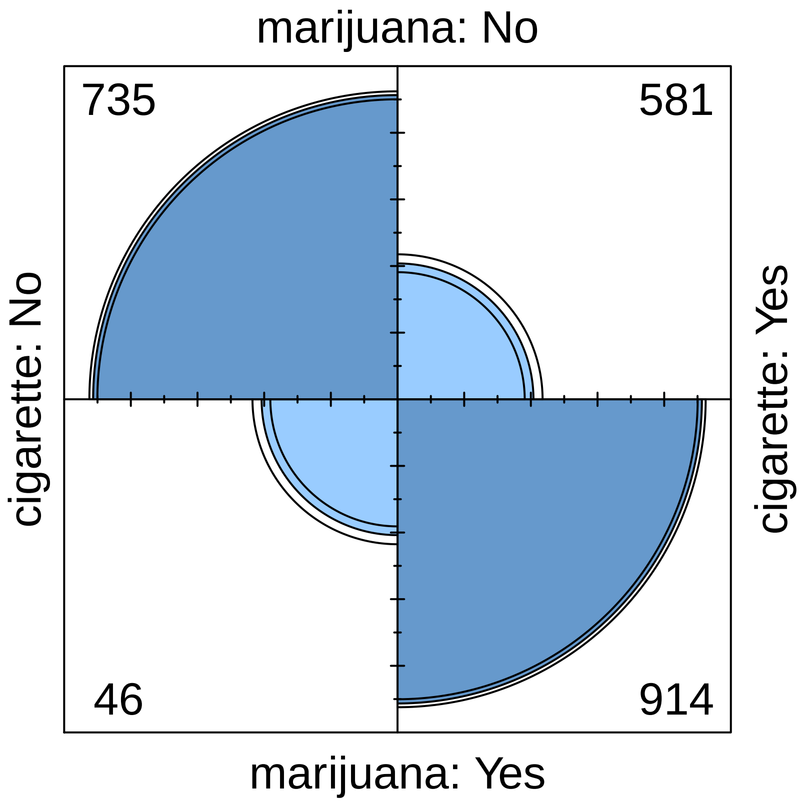
Check in help the function fourfoldplot by usign the command '?fourfoldplot'

  - Draw a Fourfold Plot for the marginal contigency table of marijuana and cigarette.

- Discuss what you can see

```
obs.mc.xtabs <- margin.table(obs.xtabs, margin=c(1,2))

fourfoldplot(obs.mc.xtabs)
```




Note that:
* The area of each shaded quadrant shows the observed counts.
* Circular arcs show the limits of confidence interval for the odds ratio.

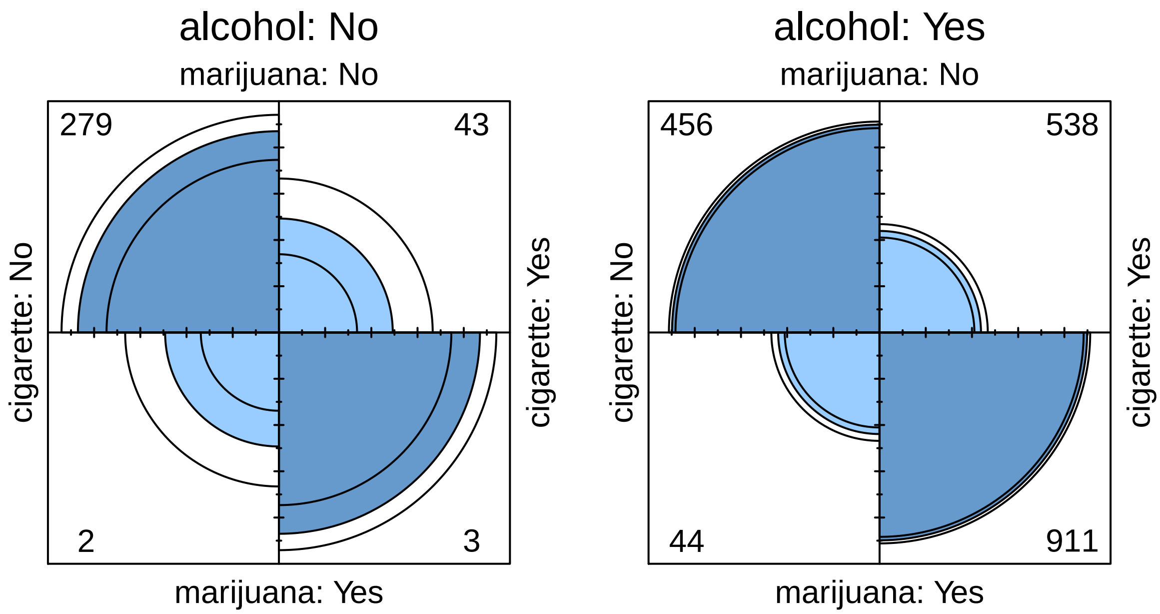
**Tables 2 x 2 x K**

Fourfold Plots can be also used for 2 x 2 x K contigancy tables

- Draw a Fourfold Plot for the contigency table of marijuana cigarette, and alcohol by controlling on the alcohol levels.

- inspect the plots

```
fourfoldplot(obs.xtabs,
             mfrow = c(1,2)
             )
```




### Mosaic plot

Mosaic plot display graphically the cells of a contingency table as rectangular areas of size proportional to the corresponding observed frequencies.

When the classification variables are independent the areas tend to be perfectly aligned in rows and columns.

The greater the deviation is, the worse the aforesaid alignment is.

Furthermore, specific locations of the table that deviate from independence the most can be identified and thus the pattern of underlying association can be explained.

The strength of individual cells contribution to divergence from independence as well as the direction of the divergence are reflected in the magnitude and sign of the corresponding independence model's residuals that can be incorporated in a mosaic plot.

R provides a function to draw this kind of plots by using the function `mosaic' from the package`vcd'

- Install 'vcd' package and load it

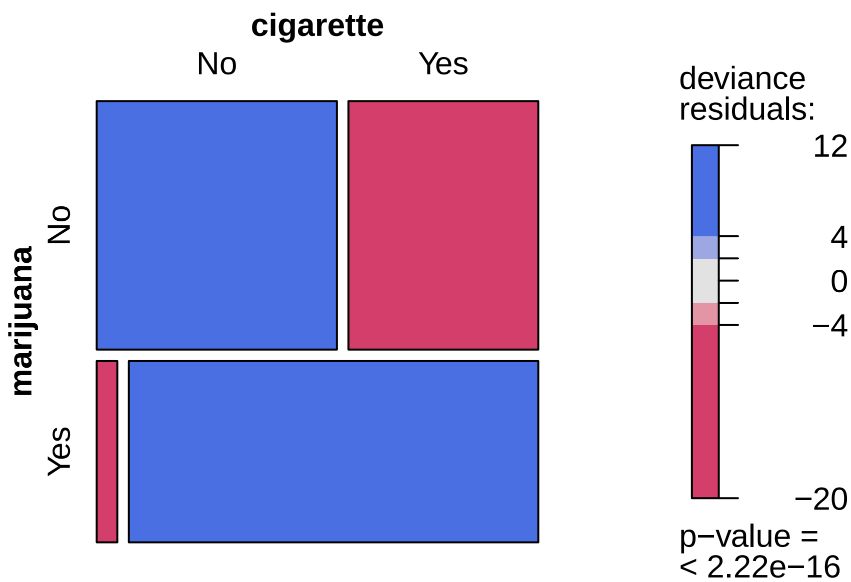
```
# install.packages('vcd') # IF NOT ALREADY INSTALLED ON YOUR PC, THEN UNCOMMENT AND RUN THIS COMMAND
library(vcd)
```

- Check the command `mosaic' in help by typing`?mosaic'

For the I x J case:

- Draw a Mosaic Plot for the marginal contigency table of marijuana cigarette.
- in particular use use mosaic(x,residuals_type="deviance",gp=shading_hcl) where x is the contigency table of interest
- Interpretet the plots

```
mosaic(obs.mc.xtabs,
       residuals_type="deviance",
       gp=shading_hcl)
```




* I observe a substantial deviation from the Independent association.

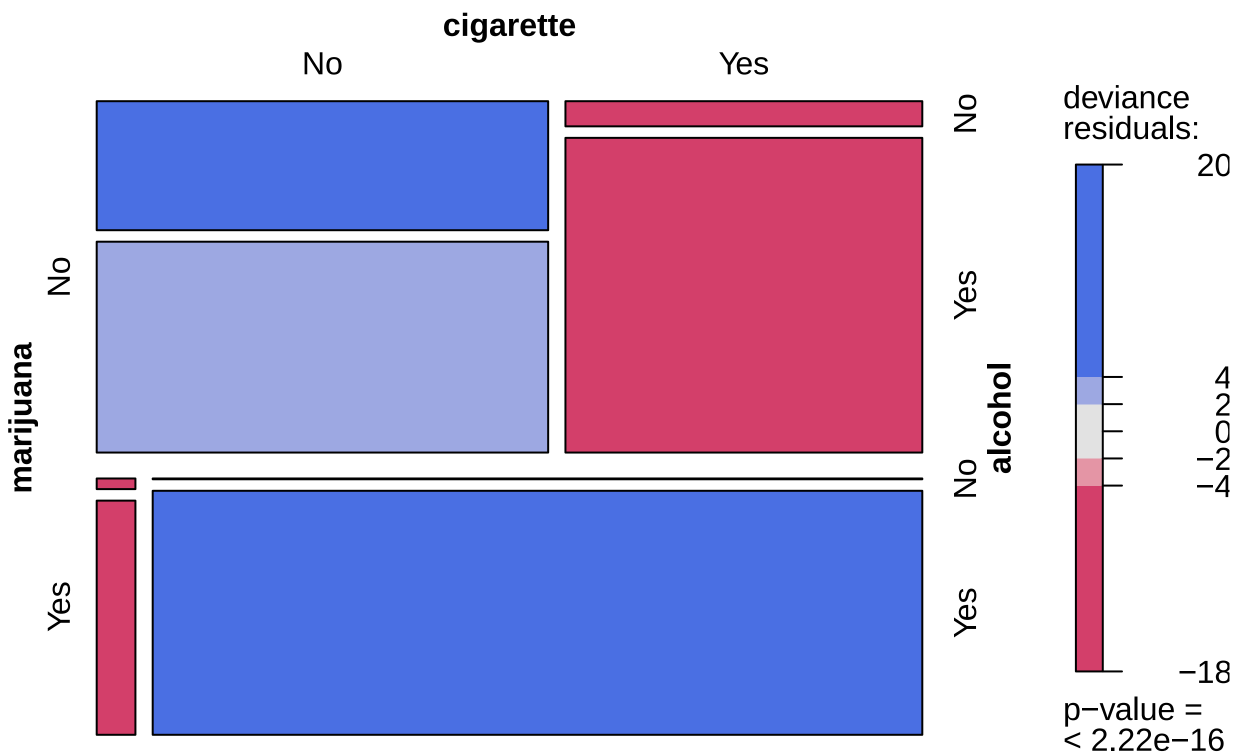
- The p-value of the GoF test based on the deviance for independence model is below the colorbar, and smaller than 0.05; hence I reject the hipothesis at sig. level 5%.

  For the I x J x K case:

  - Draw a Mosaic Plot for the contigency table of marijuana cigarette and alcohol.

  - Interpretet the plots

```
mosaic(obs.xtabs,
       residuals_type="deviance",
       gp=shading_hcl)
```




- I observe a substantial deviation from the Independent association.
- The p-value of the GoF test based on the deviance for independence model is located below the colorbar, and smaller than 0.05; hence I reject the hypothesis at sig. level 5%.

---

## Save me

Generate the document as a Notebook, PDF, Word, or HTML by choosing the relevant option (from the pop-up menu next to the Preview button). Then save your Markdown code by choosing the relevant option (from the task bar menu).

Save the *.Rmd script, so that you can edit it later.