# 5- HYpothesis testing with qualitative variables

Alex Sanchez, Miriam Mota, Ricardo Gonzalo and Santiago Perez-Hoyos

Statistics and Bioinformatics Unit. Vall d'Hebron Institut de Recerca

## Readme

# Introduction

- Categorical variables represent facts that can be better described with *labels* than with numbers.
  - Example: `Sex`, better choose from {`Male` , `Female`} than from: {1,2}.
- Sometimes ordering of labels makes sense, although it is not reasonable to assign numbers to categories:
  - Example: `Tumor stage`: {1,2,3,4}, burt $1 + 2 \neq 3$!!!
- `Sex` is an example of a categorical variable in `nominal` scale
- `Stage` is an example of a categorical variable in `ordinal` scale

```
sex <- factor(c("Female", "Male"))
stage <- factor(1:4, ordered=TRUE)
```

# The analysis of categorical variables

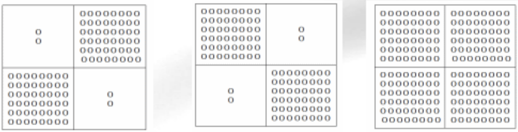
- One variable (tests with proportions)
  - Does the proportion (% affected) match a given value?
  - Is the proportion (% affected) the same in two populations?
- With two variables (chi-square and related)
  - Is there an association between two categorical variables?
  - Is there a relationship between the values of a categorical variable before and after treatment?

# Example

Consider the following study relating smoking and cancer

Load data: "dadescancer.csv"

| | Smoking X=1 | Non smoking X=0 | TOTAL |
|---|---|---|---|
| CANCER Y=1 | 190 | 87 | 277 |
| NO CANCER Y=0 | 60 | 163 | 223 |
| TOTAL | 250 | 250 | 500 |



# Crosstabulating a dataset

```
dadescancer <- read.csv("datasets/dadescancer.csv", strings
```

```
attach(dadescancer)
mytable <-table(cancer, fumar)
mytable
```

```
##            fumar
## cancer      Fuma No fuma
##   Cancer     190      87
##   No cancer   60     163
```

# Crosstabulation (2): Marginal tables

```r
margin.table(mytable, 1) # A frequencies (summed over B)
```

```
## cancer
##    Cancer No cancer
##       277       223
```

```r
margin.table(mytable, 2) # B frequencies (summed over A)
```

```
## fumar
##    Fuma No fuma
##     250     250
```

# Crosstabulation (2): In percentages

```r
prop.table(mytable) # cell percentages
```

```
##            fumar
## cancer       Fuma No fuma
##   Cancer    0.380   0.174
##   No cancer 0.120   0.326
```

```r
prop.table(mytable, 1) # row percentages
```

```
##            fumar
## cancer           Fuma   No fuma
##   Cancer    0.6859206 0.3140794
##   No cancer 0.2690583 0.7309417
```

```r
# prop.table(mytable, 2) # column percentages
```

# Exercices

- With the osteoporosis dataset repeat the crosstabulation done above using
  - Two categorical variables
  - Variable “MENOP” and a newly created variable “catBUA” created by properly categorizing variable BUA.

# One variable: Proportion tests

- According to medical literature, in the period 1950-1980, the proportion of obese individuals (defined by medical criteria: BMI $\geq 30$) was 15% in the population of men over 55 years old.
- A random sample obtained from the same population between 2000 and 2003 showed that, over a total of 723 men older than 55, 142 were obese.
- Considering that the significance level that we use is 5%, can we say that the population of men older than 55 in 2000-2003 had the same proportion of obese cases than that population had in 50'-80'?

# Proportion tests with R

```r
prop.test(x=142, n=723, p=0.15)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  142 out of 723, null probability 0.15
## X-squared = 11.849, df = 1, p-value = 0.0005768
## alternative hypothesis: true p is not equal to 0.15
## 95 percent confidence interval:
##  0.1684325 0.2276606
## sample estimates:
##         p
## 0.1964039
```

# Contingency tables

- A contingency table (a.k.a cross tabulation or cross tab) is a matrix-like table that displays the (multivariate) frequency distribution of the variables.
- It is bidimensional, and classifies all observations according with two qualitative variables (A and B, rows and columns).

| Clasif | $B_1$ | $B_2$ | ... | $B_s$ | Total |
|---|---|---|---|---|---|
| $A_1$ | $n_{11}$ | $n_{12}$ | ... | $n_{1s}$ | $n_{1\bullet}$ |
| $A_2$ | $n_{21}$ | $n_{22}$ | ... | $n_{2s}$ | $n_{2\bullet}$ |
| ... | ... | ... | ... | ... | |
| $A_r$ | $n_{r1}$ | $n_{r2}$ | ... | $n_{rs}$ | $n_{r\bullet}$ |
| Total | $n_{\bullet 1}$ | $n_{\bullet 2}$ | | $n_{\bullet s}$ | $N$ |

# Chi-squared test

**Chi squared independence test**

When the sample comes from a single population with 2 qualitative variables, the aim is to determine if there is relationship between vars:

**Chi squared homogeneity test**

- When each row is a sample from distinct populations (groups, subgroups. . . ), the aim is to determine if both groups have significative differences in that variable

# Chi-squared tests

- When we have:
  - quantitative data,
  - one or more categories,
  - independent observations,
  - adequate sample size ($>$10)
- and our questions are like...
  - *Do the number of individuals or objects that fall in each category differ significantly from the number you would expect?*
  - *Is this difference between the expected and observed due to sampling variation, or is it a real difference?*

# Chi squared.test: Observed vs expected

| Observades | Braf - | Braf + |
|---|---|---|
| Grau 1 | 97 | 5 |
| Grau 2 | 81 | 7 |
| Grau 3 | 32 | 18 |

| Esperades | Braf - | Braf + |
|---|---|---|
| Grau 1 | 89.25 | 12.75 |
| Grau 2 | 77.00 | 11.00 |
| Grau 3 | 43.75 | 6.25 |

# Chi squared tests with R

```
mytable <-table(cancer, fumar)
chisq.test (mytable)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity corre
##
## data:  mytable
## X-squared = 84.214, df = 1, p-value < 2.2e-16
```

Alternatively use Fisher test

```
fisher.test(mytable)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  mytable
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to
```