8- 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

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}, but $1 + 2 \neq 3$!!!
  - Sex is an example of a categorical variable in nominal scale
  - Stage is an example of a categorical variable in ordinal scale

Representing categorical variables in R

- Categorical variables are well represented with *factors*

```
sex <- factor(c("Female", "Male"))
blood_group <- factor(c("A", "B", "AB", "O"))
```

- Besides, factors can be forced to be "ordereded"

```
tumorstage <- factor(1:4, ordered=TRUE)
```

- Be careful with the names of factors, by default, *levels* assigned in alphabetical order.

```
levels(blood_group)
```

```
## [1] "A"  "AB" "B"  "O"
```

Creating factors

- Factors can be created . . .
    - automatically, when reading a file or
        - Not all functions for reading data from file will create a factor!!!
        - Usually levels will be defined from alphabetic order
    - using the factor or the as.factor commands.
        - more flexible

Create factors automatically

- This is achieved by
  - Using the `read.table` or `read.delim` functions for reading
    - Setting the "character variables as.factors" to TRUE
- Example
  - Load the `diabetes` dataset using the `Import Dataset` feature of Rstudio
    - From text (base) (use the file `diabetes.csv`)
    - From text (readr) (use the file `diabetes.csv`)
    - From Excel (use the file `diabetes.xls`)
  - What is the class of the variable `mort`

Exercise 1

- Select one of the datasets that you have worked with during the course
  - diabetes.xls
  - osteoporosis.csv
  - demora.xls
- Read the dataset into R and check that the categorical variables you are interested in are converted into factors.
- Confirm the conversion by summarizing the variables

Exercise 2

- Use the diabetes.sav file and import it into R with the "Import from SPSS" feature.
  - What is the class of the "MORT" variable.
  - Turn it into one factor so that it has the same levels as when you read it using `read.csv`

The analysis of categorical variables

- The analysis of categorical data proceeds as usual:
- Start exploring the data with the tables and graphics
- Proceed to estimation and/or testing if *appropriate*
- Estimation
    - Proportions: Point estimates, confidence intervals, Sample Size
- Testing
    - One variable (tests with proportions)
    - With two variables (chi-square and related)

Types of test with 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

Our goal here would be to determine if there is an association between smoking and cancer.

Crosstabulating a dataset

- Data may come from a table (aggregated) or disagregated in a data file.
- In this case we need to build the table applying "cross-tabulation"

```
dadescancer <- read.csv("datasets/dadescancer.csv",
                        stringsAsFactors = TRUE)
```

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

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

There are many ways to do crosstabulation

```
with(dadescancer, table(cancer, fumar) )
```

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

```
myXtable <- xtabs (~ cancer + fumar, data = dadescancer)
myXtable
```

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

Crosstabulation (2): Marginal tables

Marginal values are important to understand the structure of the data:

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

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

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

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

```
mytable<- addmargins(mytable)
```

Crosstabulation (3): In percentages

Showing tables as percentages is useful for comparisons

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

```
##
##               Fuma No fuma    Sum
##   Cancer     0.0950  0.0435 0.1385
##   No cancer  0.0300  0.0815 0.1115
##   Sum        0.1250  0.1250 0.2500
```

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

```
##
##                Fuma    No fuma      Sum
##   Cancer     0.3429603 0.1570397 0.5000000
##   No cancer  0.1345291 0.3654709 0.5000000
##   Sum        0.2500000 0.2500000 0.5000000
```

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

Exercise 3

- 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 as: 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.

- With a significance level of 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

Alternative "NOT EQUAL". This is set by default.

```
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
```

Alternative "GREATER"

```
prop.test(x=142, n=723, p=0.15, alternative="g")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  142 out of 723, null probability 0.15
## X-squared = 11.849, df = 1, p-value = 0.0002884
## alternative hypothesis: true p is greater than 0.15
## 95 percent confidence interval:
##  0.1725953 1.0000000
## sample estimates:
##         p
## 0.1964039
```

Alternative "LESS THAN"

```
prop.test(x=142, n=723, p=0.15, alternative="l")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  142 out of 723, null probability 0.15
## X-squared = 11.849, df = 1, p-value = 0.9997
## alternative hypothesis: true p is less than 0.15
## 95 percent confidence interval:
##  0.0000000 0.2225404
## sample estimates:
##         p
## 0.1964039
```

Notice that *choosing the wrong alternative may yield unreasonable conclusions*.

Estimation comes with proportion test

- prop.test does **three** distinct calculations
  - A test for the hypothesis $H_0 : p = p_0$ is performed
  - A confidence interval for p is built based on the sample
  - A point estimate for p is also provided.

```
> 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          Hypothesis Test
alternative hypothesis: true p is not equal to 0.15
95 percent confidence interval:
 0.1684325 0.2276606                                      Confidence interval
sample estimates:
        p                                                 Point estimate
0.1964039
```

Exercise 4

- In the osteoporosis dataset.
  - Test the hypothesis that the proportion of women with osteoporosis is higher than 7%
    - In the global population of the study
    - Only in women with osteoporosis
  - Select a sample of size 100 and repeat the test. How do the results change?
  - What sample size should we have taken so that th precision of the confidence intervals would have been at most 3% with a probability of 95%?

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 to two categorical 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

- A *familiy* of tests receiving its name because they all rely on the *Chi-Squared distribution* to compute the test probabilities.

Chi squared independence test

- When the sample comes from a single population with 2 categorical variables, the aim is to determine if there is relationship between them.

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 pair of categories differ significantly from the number you would expect if there was no association?*
  - *Is this difference between the expected and observed due to chance ("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<- with(dadescancer, 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
```

Fisher test. an assumptions-free alternative

Chi-squared test require that sample sizes are "big" and expected frequencies are, at least greater than 5.

Fisher test can be an alternative if these assumptions are not met, especially for two times two tables.

```
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
## 95 percent confidence interval:
##  3.945907 8.936465
## sample estimates:
## odds ratio
```

Some about diagnosis

Diagnosis

FOX

Diagnosis
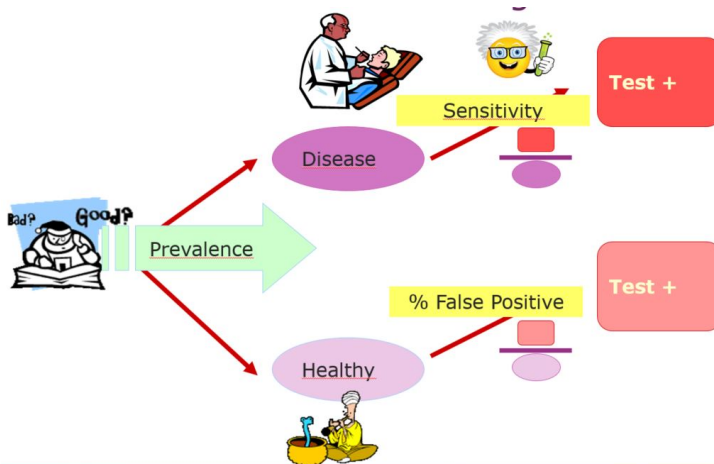
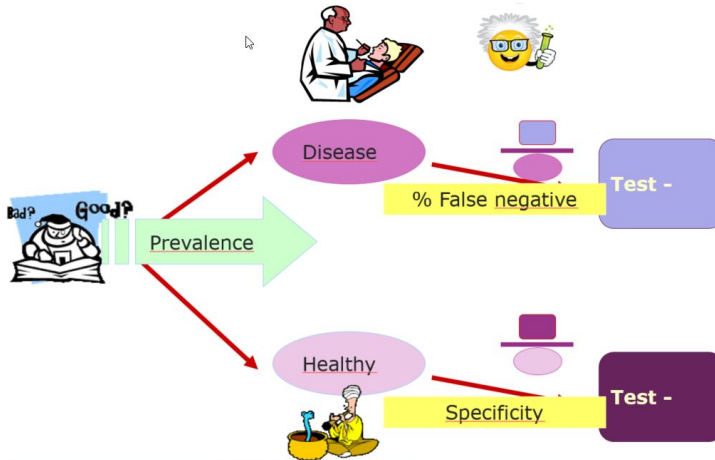Mmm ...let me do some tests and valorate them

Diagnostic test

- Most importan result of medical practice
  - Classification of individuals in healthy or sick
  - Need of reference method or "TRUE"
  - Positive results of test in patients and negative in healthy
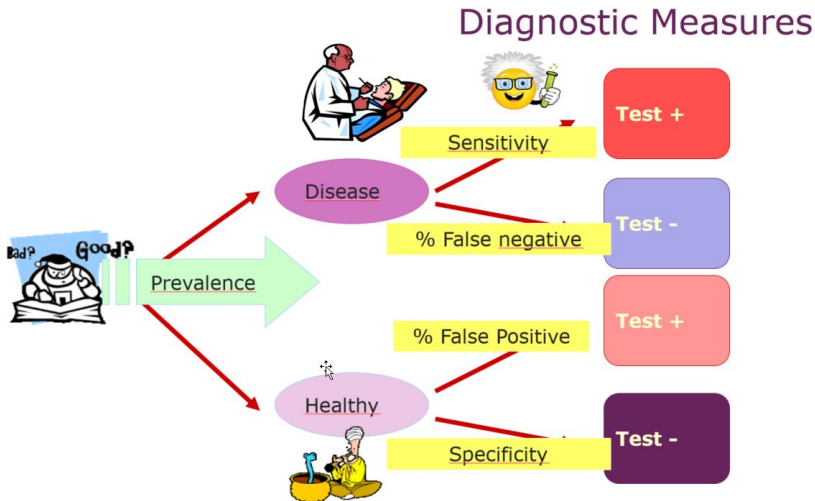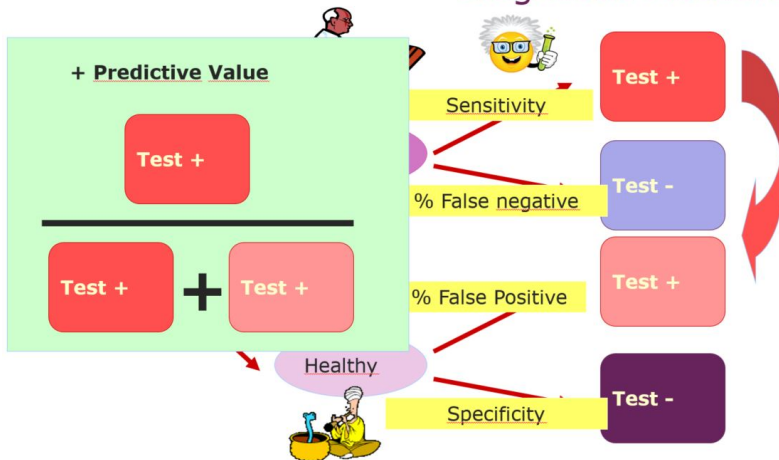
Sensitivity
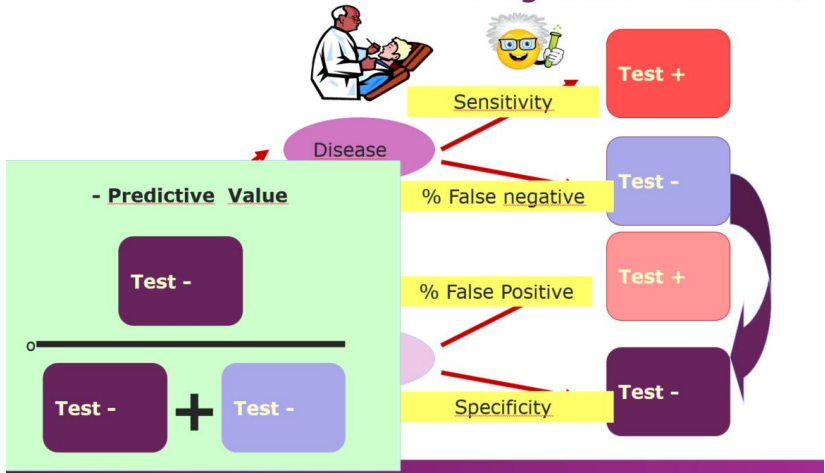
Specificity

Diagnostic Measures

Positive Predictive Value

Negative Predictive Value

Diagnostic meaures in R

```
library(epiR)
table1<- as.table(matrix(c(634,269,487,1251),nrow=2, byrow=
epi.tests(table1)
```

```
##              Outcome +    Outcome -       Total
## Test +            634          269         903
## Test -            487         1251        1738
## Total            1121         1520        2641
##
## Point estimates and 95% CIs:
## --------------------------------------------------------
## Apparent prevalence *                    0.34 (0.32, 0.36)
## True prevalence *                        0.42 (0.41, 0.44)
## Sensitivity *                            0.57 (0.54, 0.59)
## Specificity *                            0.82 (0.80, 0.84)
## Positive predictive value *              0.70 (0.67, 0.73)
## Negative predictive value *              0.72 (0.70, 0.74)
```