Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

8- Hypothesis testing with qualitative variables

Miriam Mota and Santi Pérez

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

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Readme

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Section 1

Introduction. Categorical Variables

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Introduction. Categorical Variables

- 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

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

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"))

tabaco <- factor(c(" Si", "No"))
levels(tabaco)
```

```
## [1] " Si" "No"
```

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

```
levels(blood_group)
```

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

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

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

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

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 `osteoporosis.csv`)
    - From text (readr) (use the file `osteoporosis.csv`)
  - What is the class of the variable `menop`

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Create factors automatically

```
osteo1 <- read.csv("datasets/osteoporosis.csv",sep = "\t",
class(osteo1$menop)
summary(osteo1$menop)
str(osteo1)

library(pacman)
p_load(readr)
osteo2 <- read_delim("datasets/osteoporosis.csv", "\t", esc
class(osteo2$menop)
osteo2$menop <- as.factor(osteo2$menop)
class(osteo2$menop)
summary(osteo2$menop)
str(osteo2)
```

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Exercise 1

- Select `diabetes.xls` datasets
- Read the dataset into R and check that the categorical variables you are interested (mort, tabac, ecg) in are converted into factors.
- Confirm the conversion by summarizing the variables

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

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
- Testing
  - One variable (tests with proportions)
  - With two variables (chi-square and related)

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

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?

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Section 2

Types of test with categorical variables

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Types of test with categorical variables

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis
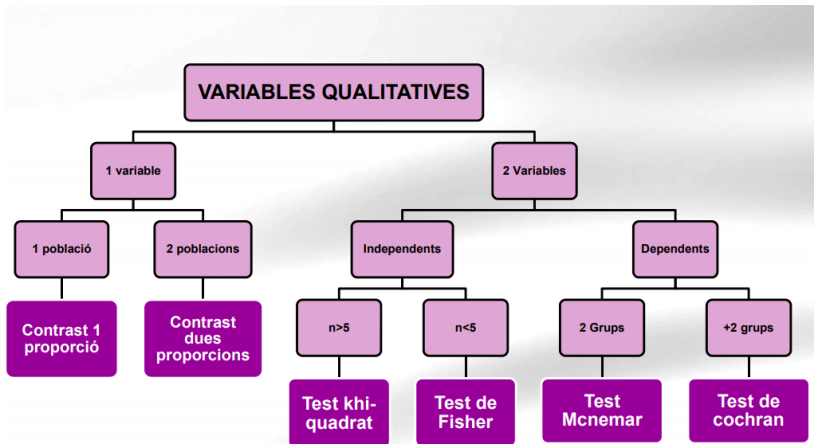
Example

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

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

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, se
head(dadescancer)
```

```
##   id cancer fumar  sexe
## 1  1 Cancer  Fuma Mujer
## 2  2 Cancer  Fuma Mujer
## 3  3 Cancer  Fuma Mujer
## 4  4 Cancer  Fuma Mujer
## 5  5 Cancer  Fuma Mujer
## 6  6 Cancer  Fuma Mujer
```

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

```
##
##           Fuma No fuma
##   Cancer   190      87
```

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Crosstabulation (2): Marginal tables

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

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

```
##
##              Fuma No fuma Sum
##   Cancer      190      87 277
##   No cancer    60     163 223
##   Sum         250     250 500
```

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Crosstabulation (3): In percentages

Showing tables as percentages is useful for comparisons

```
mytable_prop <- prop.table(mytable)
mytable_prop
```

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

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Crosstabulation (3): In percentages

Showing tables as percentages is useful for comparisons

```
gmodels::CrossTable(dadescancer$cancer, dadescancer$fumar, prop.chisq = F)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  500
##
##
##                    | dadescancer$fumar
## dadescancer$cancer |     Fuma |   No fuma | Row Total |
## -------------------|-----------|-----------|-----------|
```

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Plot

Showing plot as percentages is useful for comparisons

```
p_load(dplyr, ggplot2)

# Calcular frecuencias y porcentajes
datos_porcentaje_total <- dadescancer %>%
  group_by(fumar, cancer) %>%
  summarise(n = n()) %>%
  group_by(fumar) %>%
  mutate(porcentaje = n / sum(n) * 100)


ggplot(datos_porcentaje_total, aes(x = fumar, y = porcentaje, fill = cancer)) +
  geom_col(position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(
    title = "Proporción de fumadores por condición de cáncer",
    x = "Fuma",
    y = "Porcentaje",
    fill = "Cancer"
  )
```

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Plot

Showing plot as percentages is useful for comparisons

Proporción de fumadores por condición de cáncer

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis
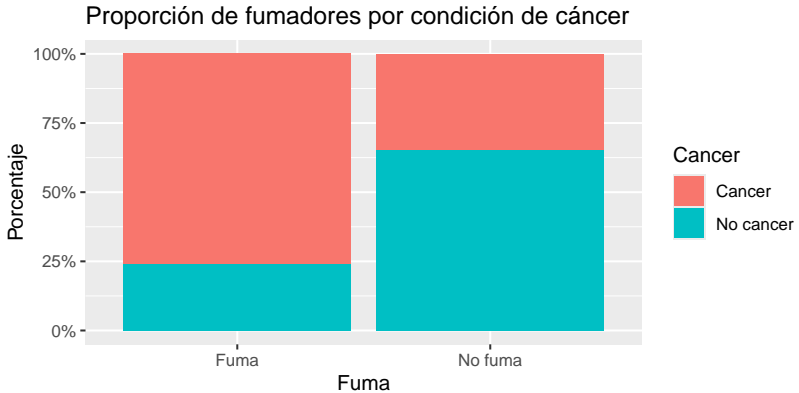
Exercise 2

- With the diabetes dataset repeat the crosstabulation done above using
  - Two categorical variables
  - Variable "mort" and a newly created variable "bmi30" created by properly categorizing variable bmi.

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Section 3

One variable: Proportion tests

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

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'?

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

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
```

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

```
##
##  1-sample proportions test with continuity correction
##
```

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

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
```

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

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*

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

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
alternative hypothesis: true p is not equal to 0.15
95 percent confidence interval:
 0.1684325 0.2276606
sample estimates:
        p
0.1964039
```

Hypothesis Test

Confidence interval

Point estimate

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Exercise 3

- In the diabetes dataset.
  - Test the hypothesis that the proportion of patients with bmi30 is higher than 40%
    - In the global population of the study
    - Only in patients with 'mort' equal "Muerto"

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

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	$n_{\bullet1}$	$n_{\bullet2}$		$n_{\bullet s}$	N

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Section 4

Two variables. Chi-squared test

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Two variables. 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

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Chi-squared tests

- When we have:
  - quantitative data,
  - two 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?*

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

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

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Chi squared tests: Observed vs expected with R

```
require(gmodels)
mytable <- table(dadescancer$cancer, dadescancer$fumar)
CrossTable(mytable,expected = T,prop.chisq = F,prop.c = F,p
```

```
   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
|         N / Table Total |
|-------------------------|
```

Total Observations in Table: 500

	Fuma	No fuma	Row Total
Cancer	190	87	277
	138.500	138.500	
	0.380	0.174	
No cancer	60	163	223
	111.500	111.500	
	0.120	0.326	
Column Total	250	250	500

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Chi squared tests with R

```
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
```

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

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
```

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Exercise 4

- Use the diabetes dataset to study if it can be detected an association between the variables `mort` and `tabac` in the diabetes dataset.
- Do not start with a test but with an appropriate summarization and visualization!

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Mcnemar test

Mcnemar test is used to compare the frequencies of paired samples of dichotomous data

- Ho: There is no significant change in individuals after the treatment
- H1: There is a significant change in individuals after the treatment

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Mcnemar test. Example

SELF * SURGICAL Crosstabulation

Count

		SURGICAL		Total
		Rupture	No Rupture	
SELF	Rupture	69	28	97
	No Rupture	5	63	68
Total		74	91	165

```
.Table <- matrix(c(69,28,5,63), 2, 2, byrow=TRUE)
mcnemar.test(.Table)
```

```
##
##   McNemar's Chi-squared test with continuity correction
##
```

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Cochran Test

Un investigador quiere evaluar si hay diferencia en la proporción de éxitos en 3 tratamientos diferentes aplicados a los mismos 10 pacientes

- $1 =$ éxito,
- $0 =$ falla

```
p_load(DescTools)

datos <- data.frame(
  Tratamiento_A = c(1, 1, 0, 1, 0, 1, 0, 1, 1, 0),
  Tratamiento_B = c(1, 0, 0, 1, 0, 1, 1, 1, 0, 0),
  Tratamiento_C = c(0, 0, 0, 1, 1, 1, 1, 1, 1, 0)
)
```

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Cochran Test

```
# Test de Cochran Q
CochranQTest(as.matrix(datos))
```

```
##
##  Cochran's Q test
##
## data:  y
## Q = 0.4, df = 2, p-value = 0.8187
```

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Section 5

Some about diagnosis

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Some about diagnosis

diagnosis

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Diagnostic test
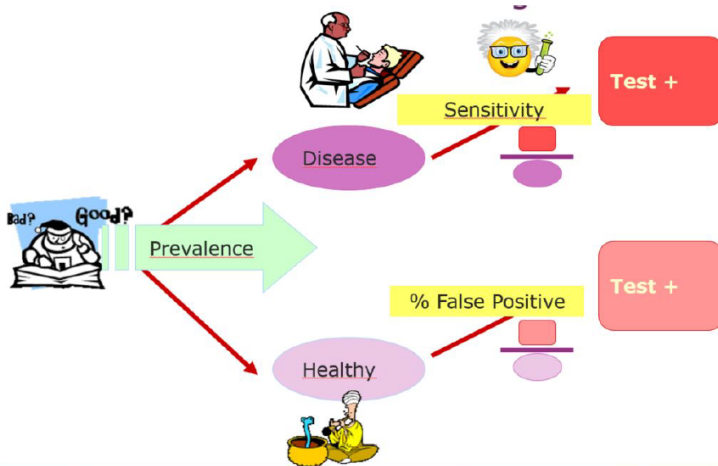
Most importan result of medical practice

- Classification of individuals in healthy or sick
- Need of reference method or "TRUE"
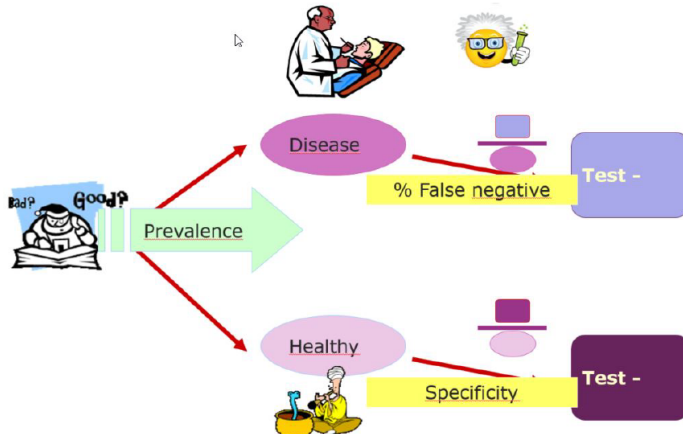- Positive results of test in patients and negative in healthy

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Diagnostic test

		Reference method		
		Sick	Healthy	TOTAL
Diagnostic	Positive	a	b	a+b
Test	Negative	c	d	c+d
	TOTAL	a+c	b+d	a+b+c+d

a = True positives (TP)
b = False positives (FP)
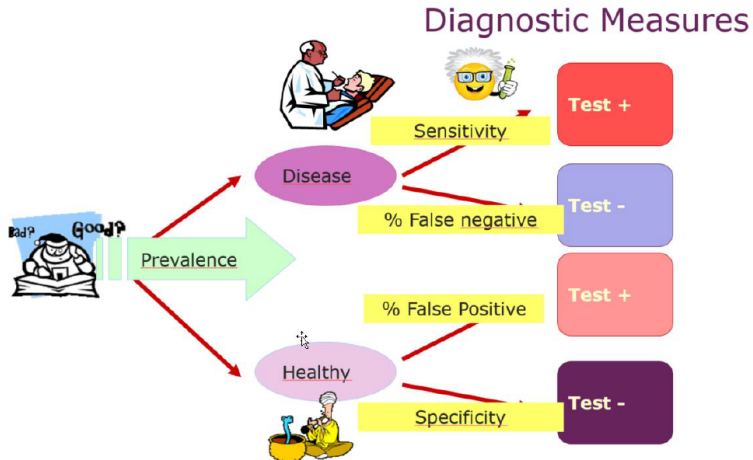c = False negative (FN)
d = True negative (TN)

Introduction. Categorical Variables
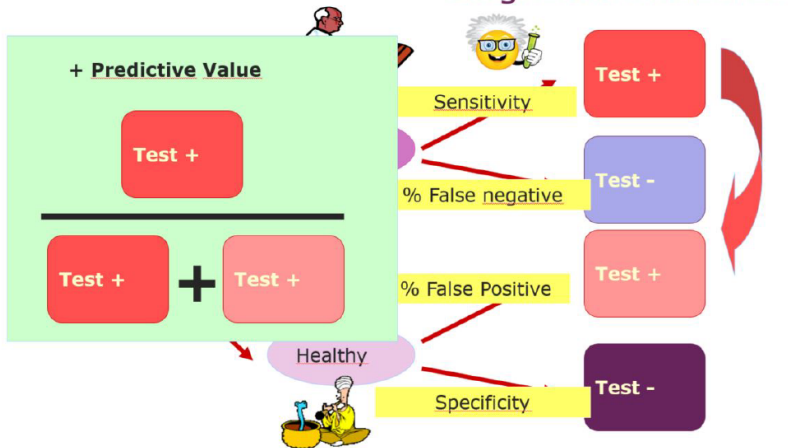Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Sensitivity

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Specificty

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Diagnostic Measures

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Positive Predictive Value

Introduction. Categorical Variables
Types of test with categorical variables
One variable: Proportion tests
Two variables. Chi-squared test
Some about diagnosis

Diagnostic measures in R

```
pacman::p_load(epiR)
table1 <- as.table(matrix(c(634,269,487,1251),nrow=2, byrow=TRUE))
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)
## Positive likelihood ratio             3.20 (2.83, 3.60)
## Negative likelihood ratio             0.53 (0.49, 0.57)
## False T+ proportion for true D- *     0.18 (0.16, 0.20)
## False T  proportion for true D        0.43 (0.41, 0.46)
```