Biostatistics I: Statistical tests for categorical data and R

Eleni-Rosalina Andrinopoulou

Department of Biostatistics, Erasmus Medical Center

e.andrinopoulou@erasmusmc.nl

@erandrinopoulou

McNemar test

Paired categorical data

Is there a difference in the percentage of patients with asthma between the placebo and the drug group (matched data)?

McNemar test

Scenario

Is there a difference in the percentage of patients with asthma between the placebo and the drug group (matched data)?

	Drug (asthma)	Drug (no asthma)	Total
Placebo (asthma)	a	b	a + b
Placebo (no asthma)	c	d	c + d
Total	a + c	b + d	n

Hypothesis

$H_0 : p_a + p_b = p_a + p_c$ and $p_c + p_d = p_d + p_b$

$H_1 : p_a + p_b \neq p_a + p_c$ and $p_c + p_d \neq p_d + p_b$

$H_0 : p_b = p_c$

$H_1 : p_b \neq p_c$

McNemar test: Theory

Test statistic

$$X^2 = \frac{(b-c)^2}{b+c}$$

When the values in the contingency table are fairly small a "correction for continuity" may be applied to the test statistic:

$$X^2 = \frac{(|b-c|-1)^2}{b+c}$$

McNemar test: Theory

Sampling distribution

- χ^2-distribution with *df* = 1
- Critical value and p-value

Type I error

- Normally α = 0.05

Draw conclusions

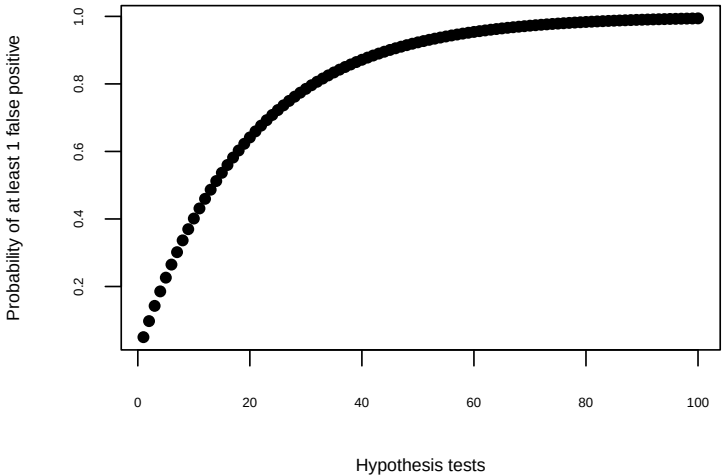
- Compare test statistic (X^2) with the critical value or the p-value with α

Multiple testing

- A single statistical test is rarely assumed
- If we perform *m* independent tests, what is the probability of at least 1 false positive?
  - P(Making an error) = α
  - P(Not making an error) = $1 - \alpha$
  - P(Not making an error in *m* tests) = $(1 - \alpha)^m$
  - P(Making at least 1 error in *m* tests) = $1 - (1 - \alpha)^m$

Multiple testing

Visualize this...



Multiple testing

Methods to adjust for multiple testing:

- **Bonferroni adjustment**: multiply the number of simultaneously tested hypothesis, e.g. $p - value = min(p - value * m, 1)$ or adjust the significant level to $\alpha = \alpha/m$
- **Holm adjustment**: $\alpha_i = \frac{\alpha}{m-i+1}$, where *i* is the order of the hypothesis - we start from the smallest to the largest p-value
- **Hochberg adjustment**: $\alpha_i = \frac{\alpha}{m-i+1}$, where *i* is the order of the hypothesis - we start from the largest p-value

apply Family

Manipulate **vectors** or slices of data from **matrices**, **arrays**, **data frames** and **lists** in a repetitive way

- An aggregating function, like for example the mean, or the sum
- Other transforming or subsetting functions
- Other vectorized functions, which return more complex structures like lists, vectors and matrices

apply Family

apply(), lapply() , sapply(), tapply(), mapply()

But how and when should we use these?

Control Flow

Sometimes we want to perform a particular action/manipulation multiple times and/or on several objects.

To repeat the same action we can specify a loop:

- `for` each element of a vector, a list, …, or
- `while` a particular condition is fulfilled

Sometimes, we may want to execute code only if a certain condition is fulfilled.

To evaluate a condition:

- `if()...else/ifelse()` execute code only if a certain condition is fulfilled.

Functions

What are functions?

- a group of (organized) **R** commands
- a (small) program with flexible (= not pre-specified) input

Almost all commands in **R are functions!**