Discovering Relationships

Eliana Ibrahimi

Biostatistician at the Department of Biology
University of Tirana, Albania

eliana.ibrahimi@fshn.edu.al

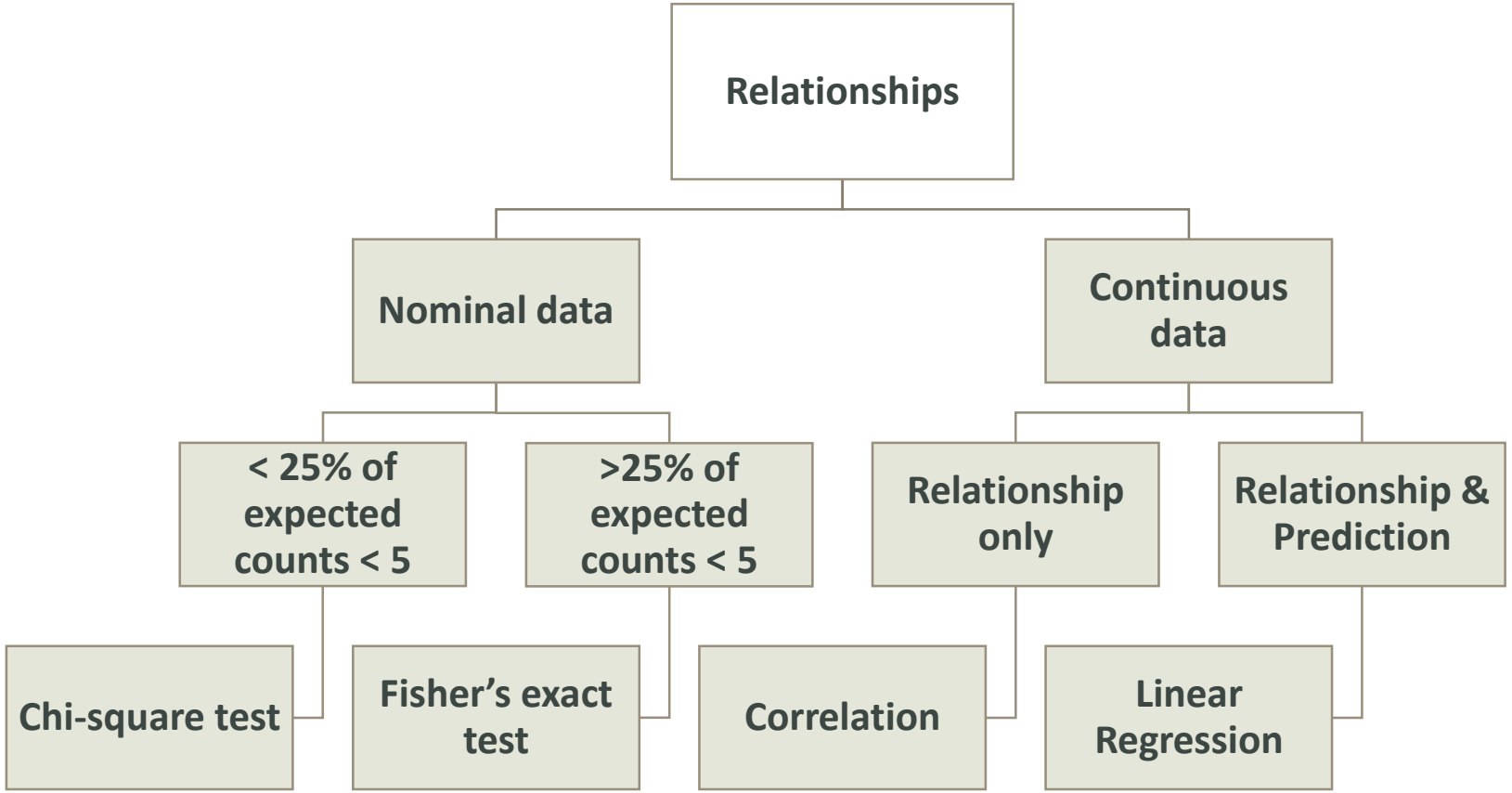
Online Workshop
Hands-on Biostatistics
July, 2020

Overview

Relationships

- Between two nominal variables
  - Chi-square test
  - Fisher’s exact test
- Between two Continuous variables
  - Linear regression
  - Correlation

Which statistical method to perform?



Chi-square test for independence

- The chi-square test for independence, also called Pearson's chi-square test or the chi-square test of association, is used to discover if there is a relationship between two categorical variables.

- **Assumptions**

1. Your two variables should be measured at an ordinal or nominal level (i.e., categorical data).
2. Your two variables should consist of two or more categorical, independent groups.
3. The expected counts should be larger than 5 in more than 75% of cases.

Chi-square test for independence

1. Hypotheses

H_0: Variables are independent

H_1: Variables are related

2. Test statistics

$$X^2 = (O_{11} - E_{11})^2 / E_{11} + (O_{12} - E_{12})^2 / E_{12} + \cdots + (O_{RC} - E_{RC})^2 / E_{RC}$$

- $\chi 2_{(R-1)(C-1)}$ *distribution*

3. Decision

p-value > 0.05 accept H_0

- Variables are independent.

P-value < 0.05 reject H_1

- Variables are related.

Fisher’s exact test

- The expected counts are smaller than 5 in more than 25% of cells?
    - Replace Chi-square with Fisher’s exact test which deals with small samples sizes.

Chi-square & Fisher's exact tests in R

- Let's go to R notebook

Correlation

- The Pearson product-moment correlation coefficient (Pearson's correlation, for short) is a measure of the strength and direction of association that exists between two variables measured on at least an interval scale.

- **Assumptions**

1. Your two variables should be measured at the interval or ratio level (i.e., they are continuous).
2. There is a linear relationship between your two variables. You can check by creating a scatterplot.
3. There should be no significant outliers.
4. Your variables should be approximately normally distributed

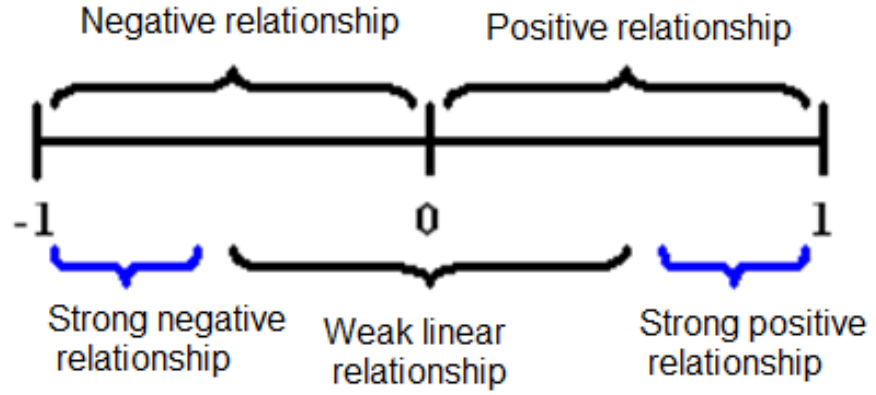
Correlation

- The Pearson product-moment correlation coefficient is calculated as:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

- Is not affected by changes in location or scale in either variable and must lie between −1 and +1.

- **Interpretation**



Cohen (1988):

|r| <0.3 Weak

0.3 ≤|r| < 0.5 Medium

|r| ≥0.5 Strong

Correlation should be significant

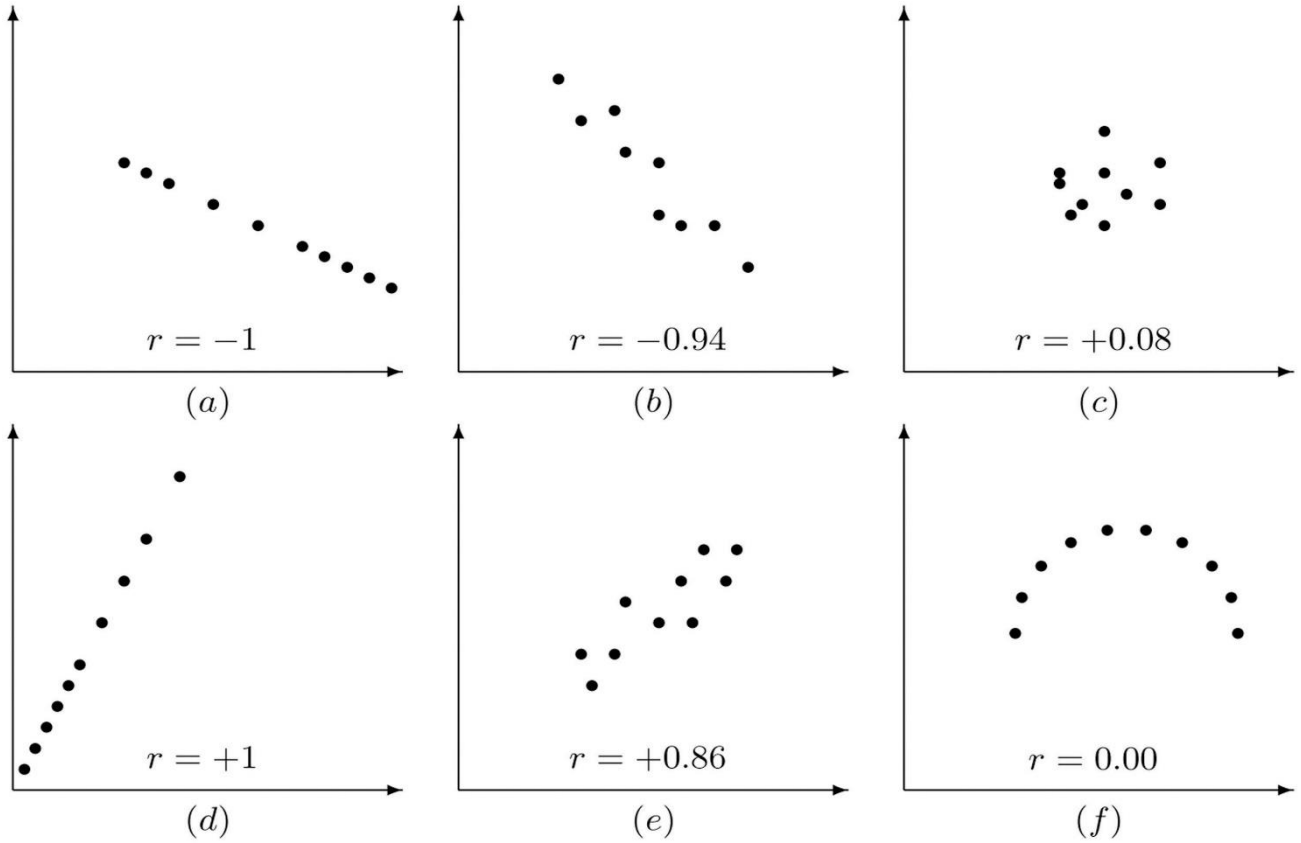
1. Hypotheses

H_0: $\rho = 0$

H_1: $\rho \neq 0$

2. Decision

- p<0.05
- There is a significant correlation

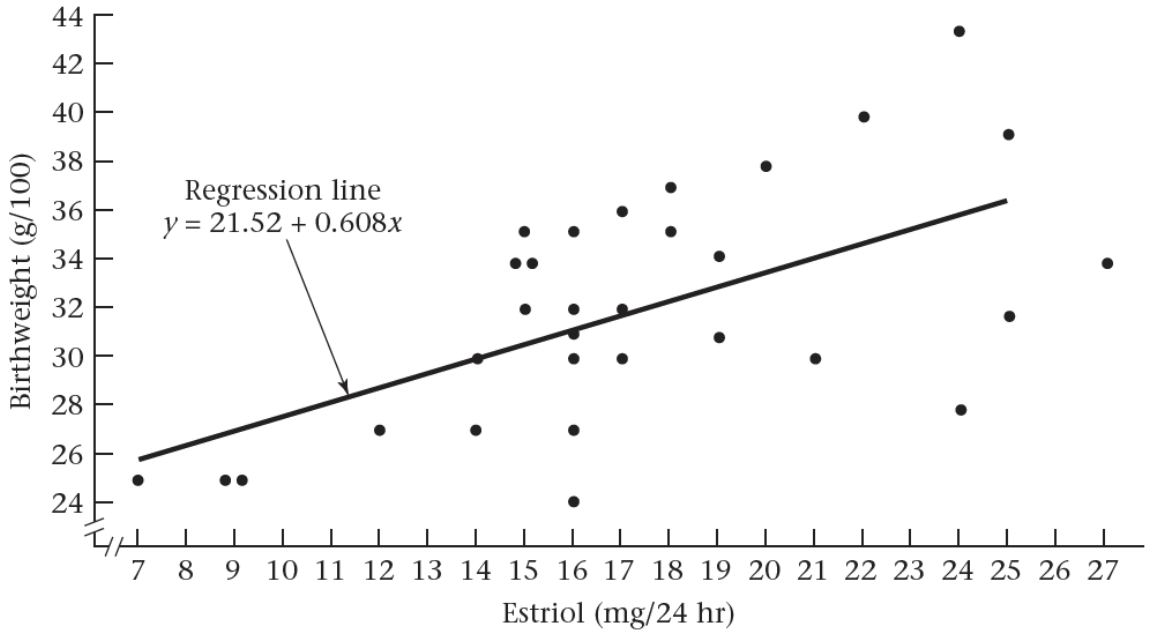


Simple linear regression

- A linear regression is a statistical model that analyzes the relationship between a response variable (often called y) and one or more variables and their interactions (often called x or explanatory variables).

$$y = \alpha + \beta x$$

- Alfa is the intercept and shows the value of Y when x is 0.
- Beta is the slope and shows the change of Y when X changes with 1 unit. Depending on the sign of beta the relationship can be negative or positive.



Simple linear regression

1. Hypotheses for β

Ho: β=0

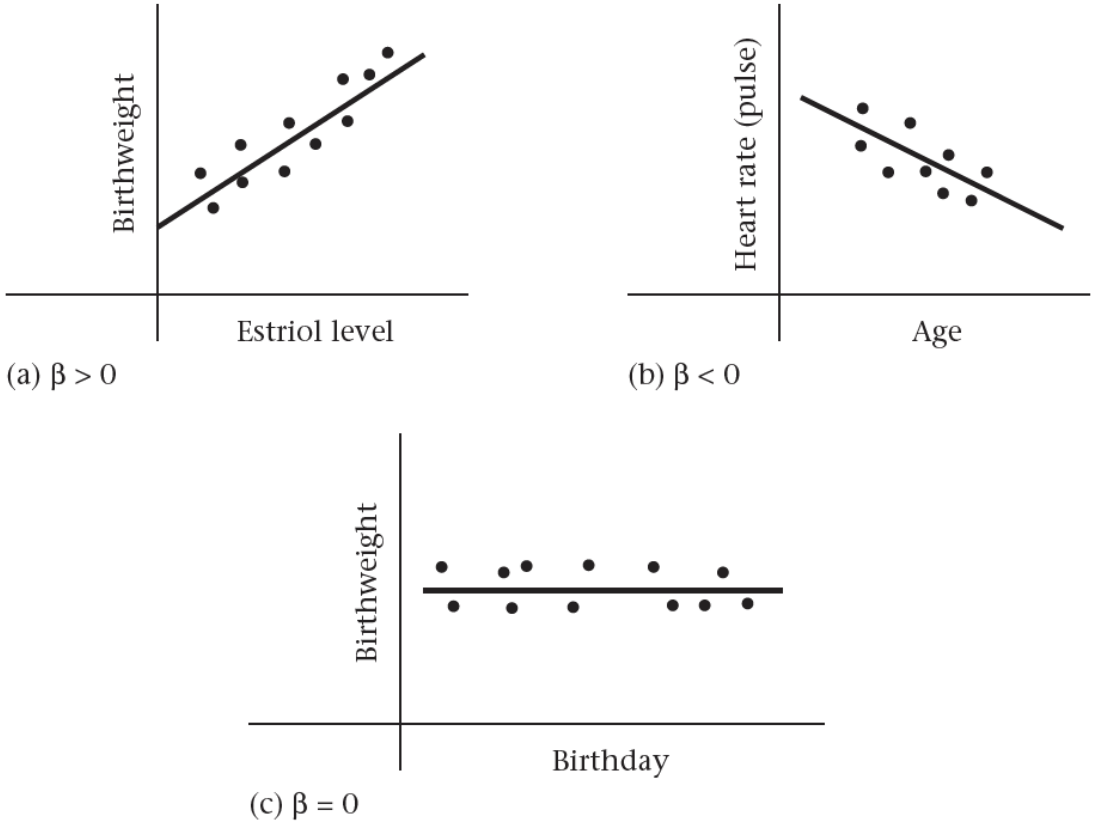
H1: β≠0

2.Test statistics (calculate it in R)

3. Decision

p<0.05 Reject Ho, there is a significant linear relationship between variables.

Interpretation of the regression line for different values of β



Correlation and linear regression in R

- Let’s go to R notebook

References/Useful links

1. *Rosner, Bernard. Fundamentals Of Biostatistics. Cengage Learning, 2011.*
2. Pezzullo, John. Biostatistics For Dummies. Wiley, 2013.
3. https://bolt.mph.ufl.edu/6050-6052
4. http://www.biostathandbook.com/HandbookBioStatThird.pdf