Univariate Statistical Analysis: Discovering Associations

Eliana Ibrahimi

Department of Biology, University of Tirana, Albania

MSB 2023 Training School

October 18-20, 2023, Tirana, Albania

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?

- **Relationships**
  - **Nominal data**
    - **< 25% of expected counts < 5**
      - **Chi-square test**
    - **>25% of expected counts < 5**
      - **Fisher's exact test**
  - **Continuous data**
    - **Relationship only**
      - **Correlation**
    - **Relationship & Prediction**
      - **Linear Regression**

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

- Go to R notebook ‘Univariate Statistical Analysis Part 1’

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

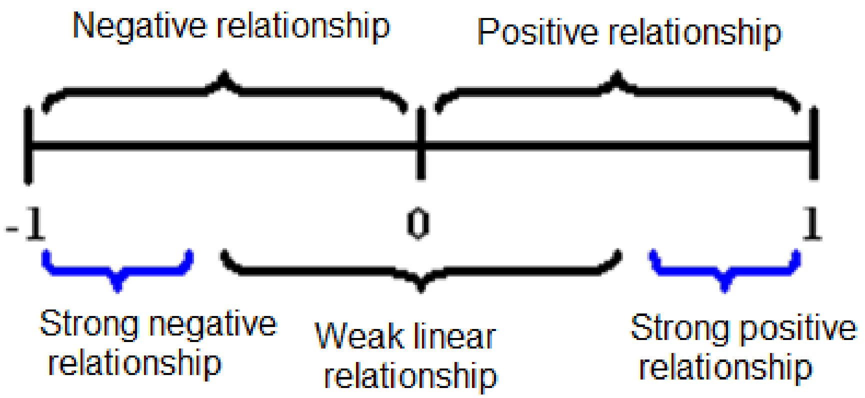
Correlation

- 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

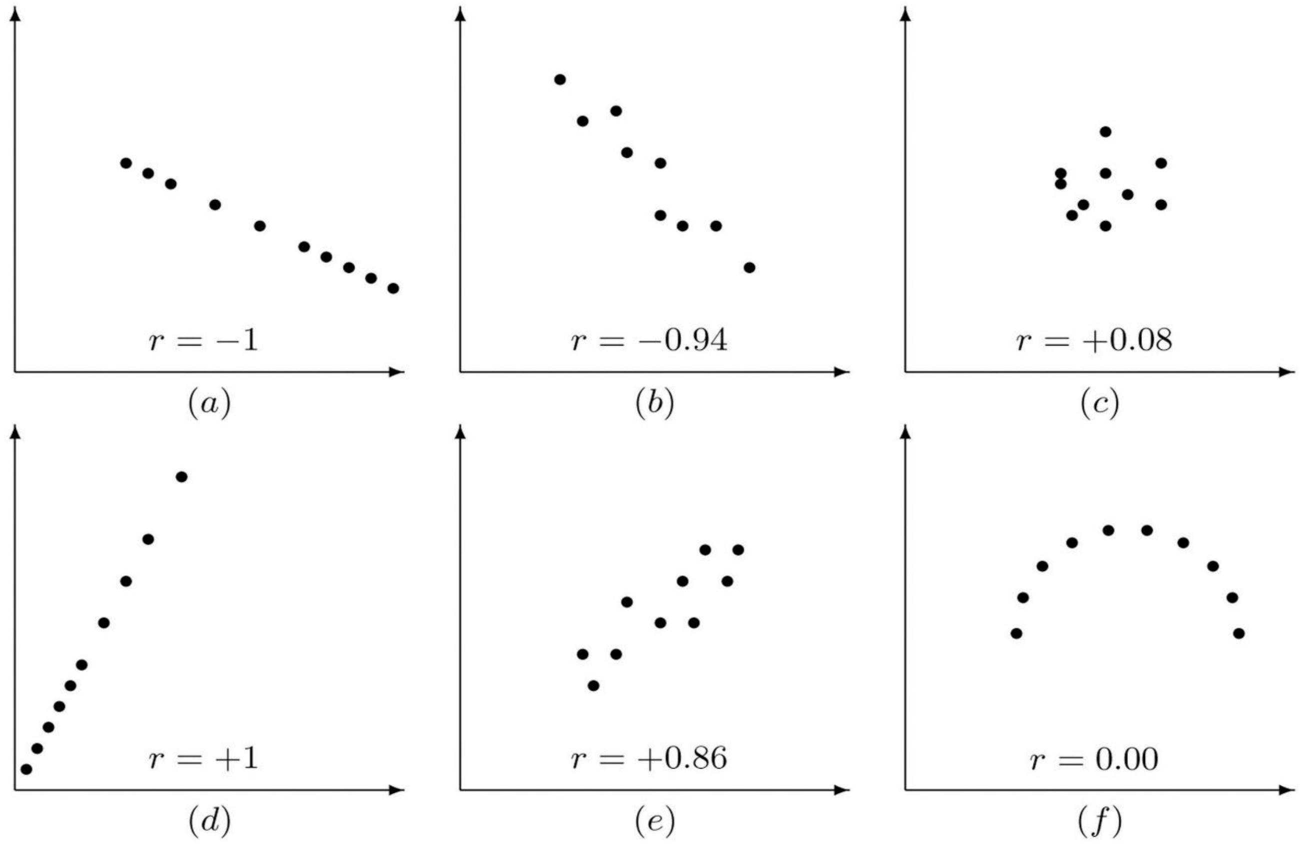
1. Hypotheses

H_0: $\rho = 0$

H_1: $\rho \neq 0$

2. Decision

- $p<0.05$
- There is a significant correlation



Spearman correlation

1. Let ρ be the Spearman's population correlation coefficient, then we can express this test as:

H_0: $\rho = 0$

H_1: $\rho \neq 0$

2. Compute the Spearman correlation coefficient and p-value in R.

$$rho = \frac{\sum(x' - m_{x'})(y_i' - m_{y'})}{\sqrt{\sum(x' - m_{x'})^2 \sum(y' - m_{y'})^2}}$$

Where $x' = rank(x)$ and $y' = rank(y)$.

3. Decision based on p-value

If p<0.05 reject the null hypothesis (Ho) There is an association between variables.

0-0.19 "very weak"
0.20-0.39 "weak"
0.40-0.59 "moderate"
0.60-0.79 "strong"
0.80-1.0 "very strong"

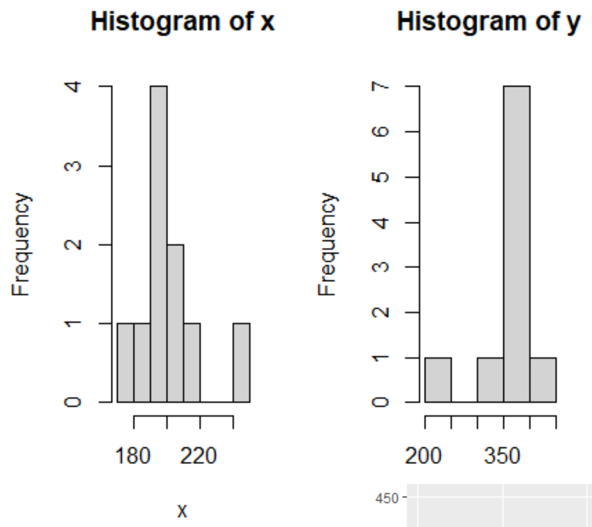
https://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf

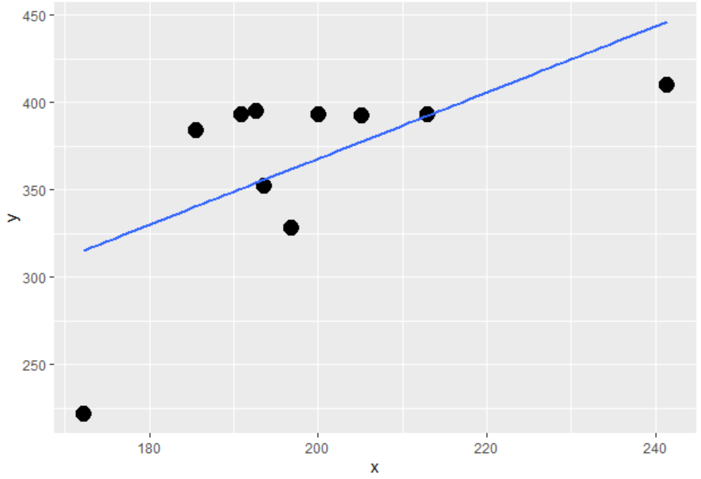
Example 4

Data: We'll use an example data set, which contains the weight of 10 mice before and after a specific treatment.

Research question: Is there a correlation between the mice weight before and after the treatment?

Note: Spearman's correlation coefficient is a statistical measure of the strength of a monotonic relationship between paired data. Read more on monotonic relationships here.





Spearman correlation: Example 4

#In order to calculate the Spearman correlation in R, use the code:

```
cor(x, y, method = c("spearman"))
cor.test(x, y, method=c("spearman"))
```

Output:

```
[1] 0.4666667

        Spearman's rank correlation rho

data:  x and y
S = 88, p-value = 0.1782
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4666667
```

This could be formally reported as follows:

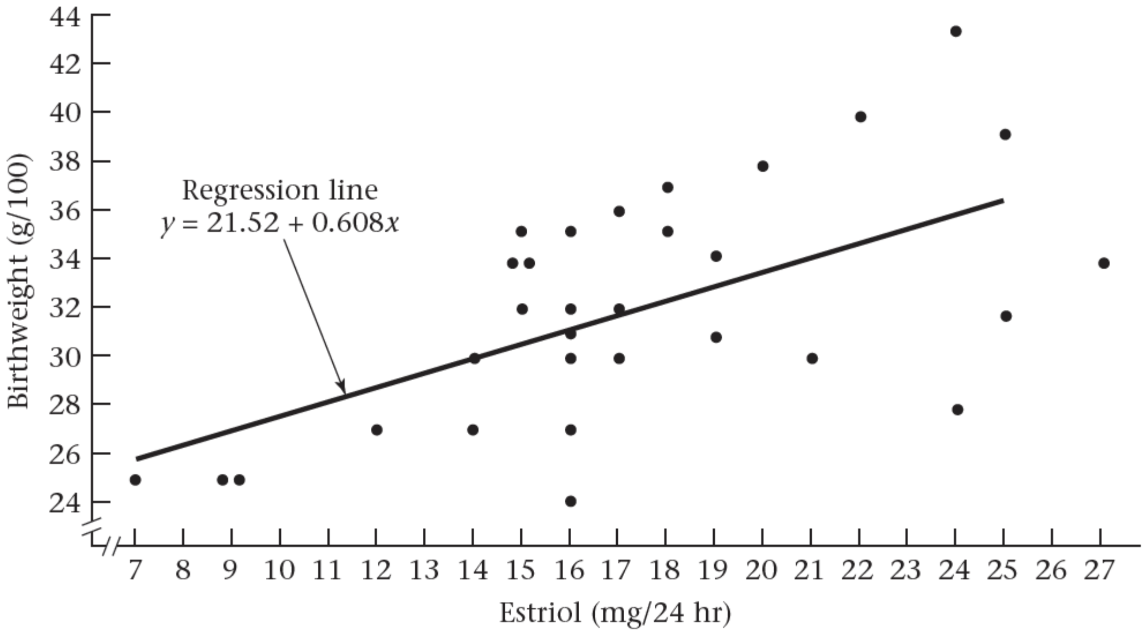
"A Spearman's correlation was run to determine the relationship between values of weight before and after treatment. There was no significant monotonic correlation between weight values (r_s= 0.467, n = 10, p=0.178)."

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.



What is the purpose of fitting a model?

- To explain the relationship between the response and the predictors.
- To predict the response based on the predictors. Often, a good model will do both.

Simple linear regression

1. Hypotheses for β

Ho: β=0

H1: β≠0

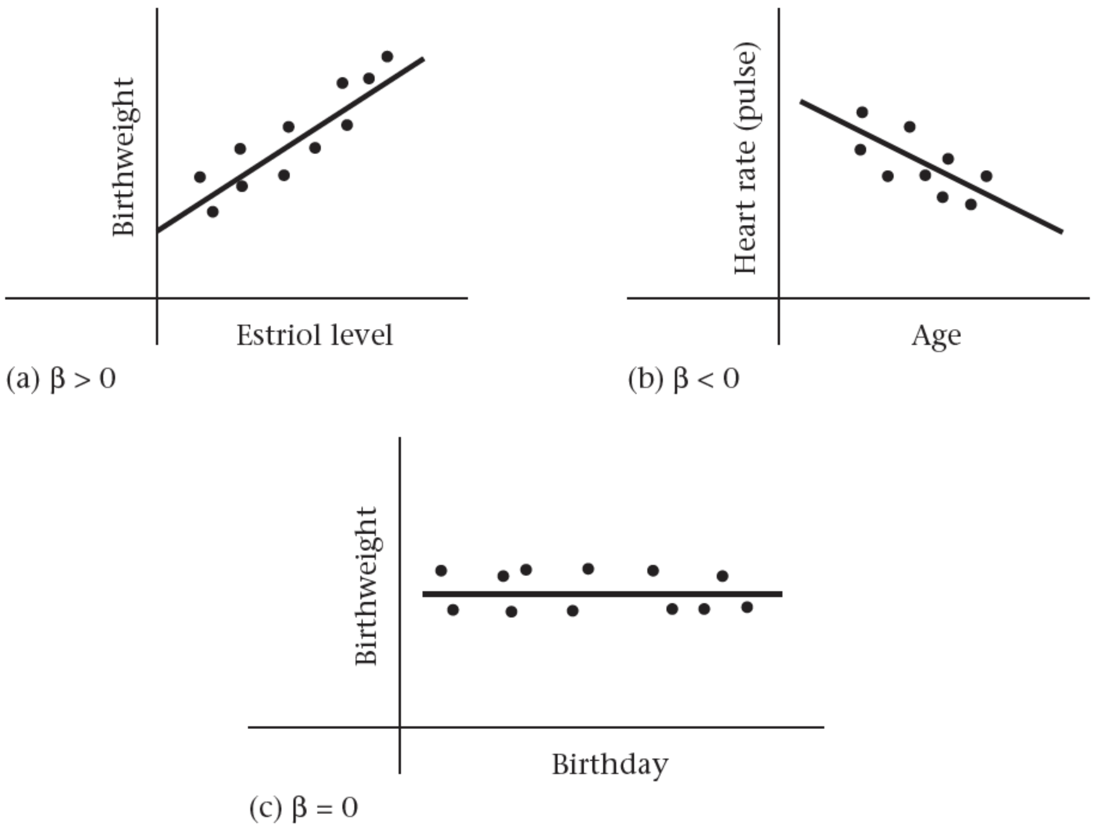
2.Test statistics (calculate it in R)

$$t = \frac{b_j}{s_{b_j}}$$

where b_j is the j[th] regression coefficient and s_{b_j} is the standard error of b_j.

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



Multiple linear regression: Overview

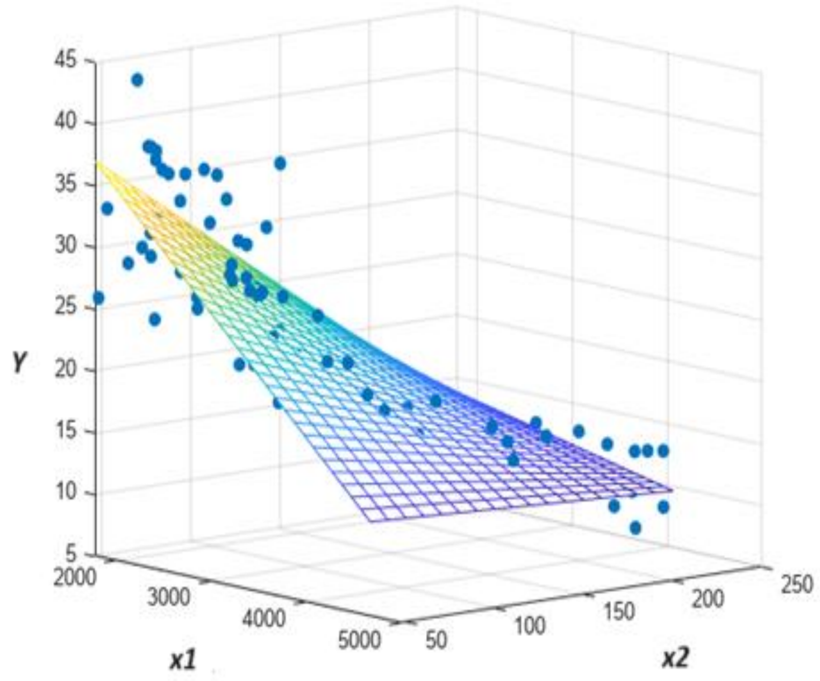
- What is multiple linear regression?
- Exploring the data before fitting multiple linear regression
- Fitting the model
- Checking the assumptions of the model
- Interpreting the output of the model
- Assessing the goodness of fit of the model
- Using the model to make predictions

What is multiple linear regression?

- When there are two or more independent variables used in the regression analysis, the model is not simply linear but a multiple regression model.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon,$$



Significance and interpretation of coefficients

The coefficients can be interpreted similarly to the simple linear regression equation after testing their significance.

If the independent variable X_i increases by one unit and all other predictors are constant, the dependent variable Y increases by β_i.

Checking the assumptions

1. Linear relationship between the dependent and the independent variables.
2. Multicollinearity, no strong correlation between independent variables.
3. Residual values are normally distributed
4. Homoscedasticity assumes that the variance of the residual errors is similar across the value of each independent variable.

Check model performance

- **Coefficient of determination R-Square**

R-squared is the proportion of the variance in the response variable that can be explained by the predictor variables.

Variance of the predicted values

$$R^2 = \frac{s^2_{\hat{y}}}{s^2_y}$$

Variance of the observed values

$$R^2_{adj} = 1 - (1 - R^2) \cdot \frac{n-1}{n-p-1}$$

- **Root mean squared error**

$$\mathrm{RMSE(model, data)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

y_i are the actual values of the response
y_i are the predicted values using the fitted model and the predictors from the data

A case study

We will use data from the Maryland Biological Stream Survey

Dependent variable: number of longnose dace (*Rhinichthys cataractae*) per 75-meter section of stream.

Independent variables (predictors): **area** (in acres) drained by the stream; **dissolved oxygen** (in mg/liter); **maximum depth** (in cm) of the 75-meter segment of stream; **nitrate concentration** (mg/liter); **sulfate concentration** (mg/liter); **water temperature** on the sampling date (in degrees C); and **water hardness** (low, <45 mg equivalent CACO3/L and high, >45).

Longnose dace, *Rhinichthys cataractae*.

Multiple regression step by step in R

Go to Multiple regression R notebook

References/Useful links

1. *Rosner, Bernard. Fundamentals Of Biostatistics. Cengage Learning, 2011.*
2. Pezzullo, John. Biostatistics For Dummies. Wiley, 2013.
3. https://datatab.net/tutorial/linear-regression
4. http://www.biostathandbook.com/HandbookBioStatThird.pdf
5. *https://www.statology.org/multiple-linear-regression-r/*
6. *https://book.stat420.org/model-building.html*
7. http://www.biostathandbook.com/multipleregression.html