ARISTOTLE UNIVERSITY OF THESSALONIKI

FACULTY OF HEALTH SCIENCES - SCHOOL OF MEDICINE

MSc Health Statistics and Data Analytics

Correlation

THEODOROS DIAKONIDIS

email: diakonidis@auth.gr

THESSALONIKI 2021-22

MSc Health Statistics & Data Analytics

Correlation

Definition (General)

- **Correlation** is a bivariate analysis that measures the strength of association between two variables.

Usability

- It can indicate a causal relationship that can be exploited in practice.

Monotonic relationships

We are going to restrict our study to monotonic relationships between two variables.

What is monotonic relationship?

In simple words is one in which, either:

- As the values of one variable increases, so does the value of the other.
- As the values of one variable increases, the other decreases.

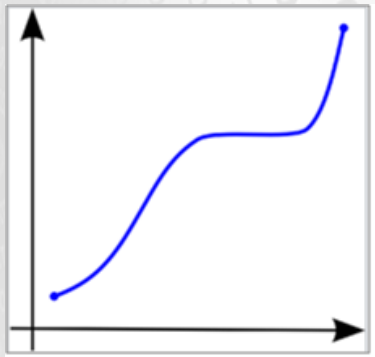


Figure 1 - A monotonically increasing function

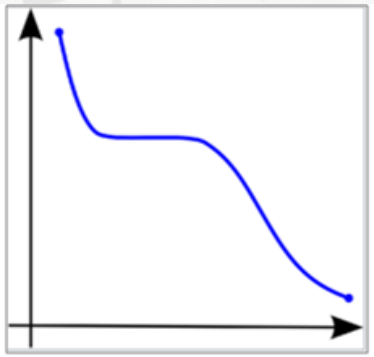


Figure 2 - A monotonically decreasing function

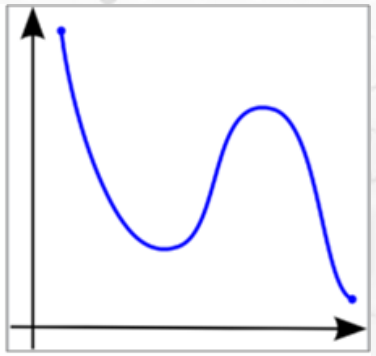
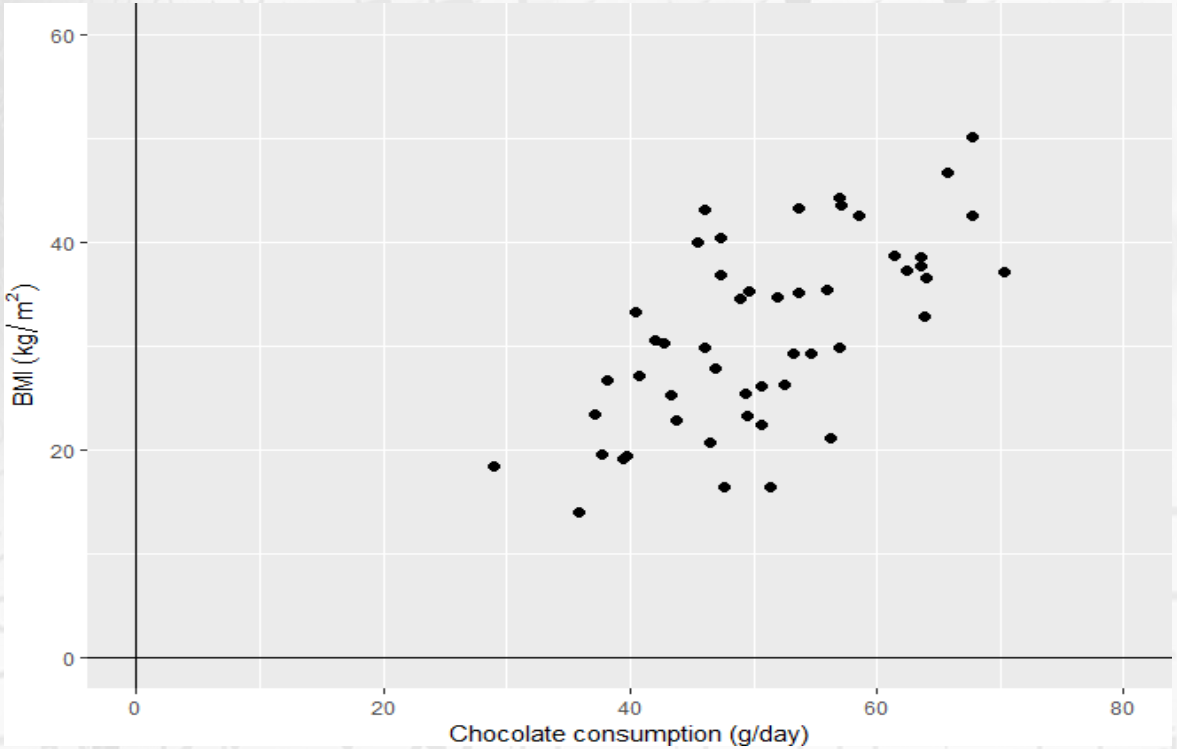


Figure 3 - A function that is not monotonic

Visual inspection (Scatter plot)

a/a	BMI	Chocolate consumption (g/day)
1	33.3	40.4
2	22.9	43.7
3	21.2	56.3
4	30.2	42.7
5	37.3	62.4
6	18.4	28.9
…	…	…
50	26.7	38.2

- A **scatterplot** is a graph that is used to plot the data points for the two variables.
- Each scatterplot has a horizontal axis (*x*-axis) and a vertical axis (*y*-axis). One variable is plotted on each axis.
- Scatterplots are made up of marks; each mark represents one study participant's measures on the variables that are on the *x*-axis and *y*-axis of the scatterplot.



Pearson's correlation (Linear Relationship)

- Also called Pearson's *r* (**dimensionless**)

$$r = \frac{Cov(\mathrm{x}, \mathrm{y})}{\sigma_{\chi}\sigma_{y}}$$

- **Covariance (Cov)** measures the total variation of two random variables from their expected values (mean values).

$$Cov(\mathrm{x}, \mathrm{y}) = \frac{\sum(\mathrm{X}_i - \bar{X})(\mathrm{Y}_i - \bar{Y})}{n}$$

- **Standard deviation (σ_{χ} ,σ_{y})** measure of the amount of variation or dispersion of a set of values.

$$\sigma_{\chi} = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n}} \quad \sigma_{y} = \sqrt{\frac{\sum(\mathrm{Y}_i - \bar{Y})^2}{n}}$$

What is meant by correlation coefficient?

- The correlation coefficient describes how one variable progresses in relation to another.

- A positive correlation indicates that the two move in the same direction, with a +1.0 correlation when they move in tandem.

- A negative correlation coefficient tells you that they instead move in opposite directions. A correlation of zero suggests no correlation at all.

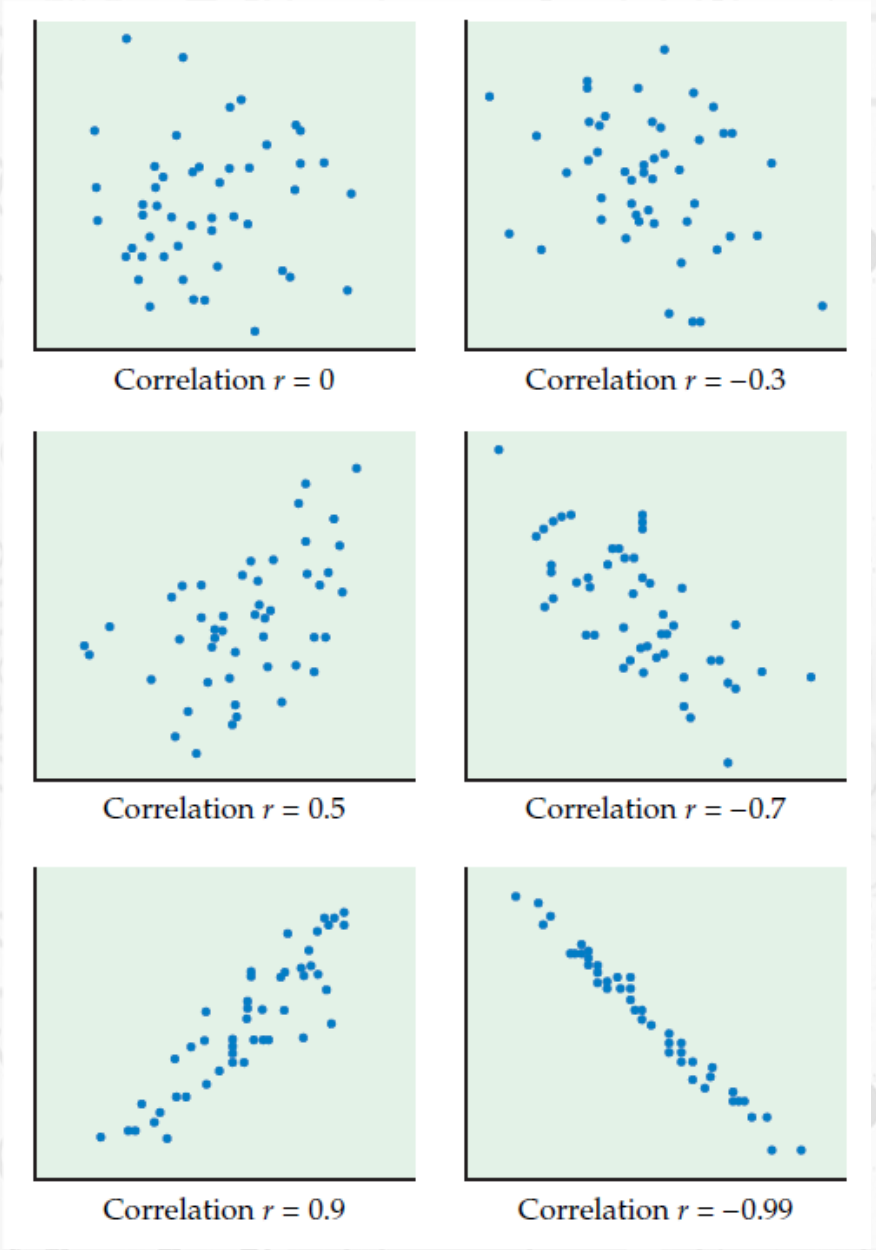
r examples and limits

- **Pure number:** It is independent of the unit of measurement. Values can range from:

$$-1 \leq r \leq +1$$

- **+1** perfect positive relationship
- **-1** perfect negative relationship
- **≈0** no relationship exists

- In the first case variables increase towards the same direction **(as X increases so does Y)** and in the latter case the opposite.

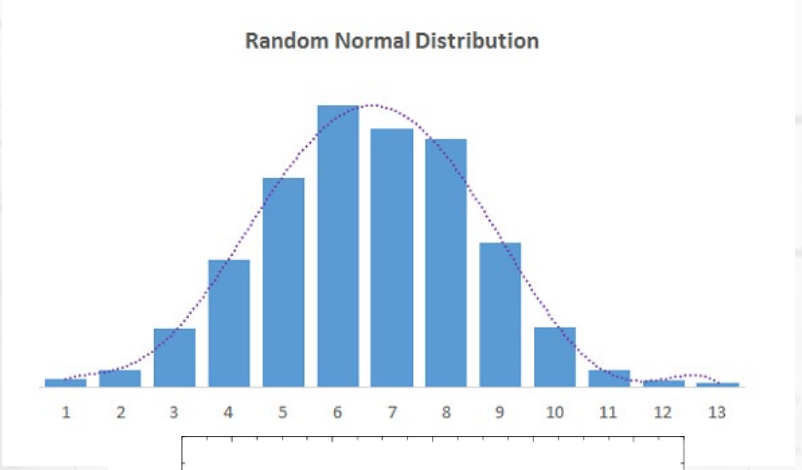


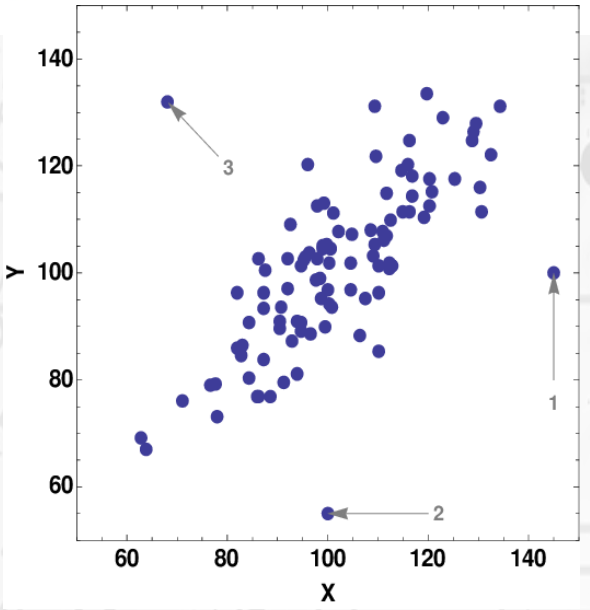
Assumptions of Pearson Correlation r

1. Both variables are normally distributed random values

2. There are no relevant outliers. Extreme outliers may have undue influence.

3. Each pair of x-y values is measured independently from each other pair.





Spearman's ρ

When:

1. The previous assumptions are not met or
2. The scatter plot shows a monotonic non-linear relationship.

- We use Spearman's ρ.
- A non parametric version of Pearson r.
- It can be robust for outliers too.
- What is needed is the relation to be monotonic.

Spearman rank correlation (Monotonic Relationship)

- **Sperman's correlation** (ρ)

$$\rho = \frac{Cov(R(x), R(y))}{\sigma_{R(\chi)}\sigma_{R(y)}}$$

- **Covariance (Cov)** in this case is that of the rank values.
- **Standard deviation ($\sigma_{R(x)}$,$\sigma_{R(y)}$)** are the standard deviations of the rank variables.
- Instead of x and y we put the rank 1,2,3 etc. of the observation R(x),R(y) in comparison to each other.

if all *n* ranks are *distinct integers (no ties in ranks):* $\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$

Where d is the difference between the two ranks of each observation $d_i = R(x_i) - R(y_i)$

Spearman simple calculation

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

S/N	English (mark)	Maths (mark)	Rank (English)	Rank (maths)	d	d^2
1	56	66	9	4	5	25
2	75	70	3	2	1	1
3	45	40	10	10	0	0
4	71	60	4	7	3	9
5	62	65	6	5	1	1
6	64	56	5	9	4	16
7	58	59	8	8	0	0
8	80	77	1	1	0	0
9	76	67	2	3	1	1
10	61	63	7	6	1	1

Interpretation of Spearman's ρ

- It is similar to that of Pearsons r
- The closer to 1 the stronger the monotonic relationship.
- Correlation is an effect size and so we can verbally describe the strength of the correlation using the following guide for the absolute value of ρ :

- .00-.19 "very weak"
- .20-.39 "weak"
- .40-.59 "moderate"
- .60-.79 "strong"
- .80-1.0 "very strong"

Misconceptions

A. Quadratic relation not captured from r or ρ

B. A single outlier can give same number of r with a completely different case, D.

C. A sinus relationship is not captured by r or ρ