University of St.Gallen

EFMD EQUIS ACCREDITED

Methods: Statistics (4,120)

2. Correlation vs. Causality

Spring 2022

Prof. Dr. Roland Füss

Swiss Institute of Banking and Finance (s/bf)

Contents

Learning Objectives

After the lecture, you know how:

- the association between random variables is measured and interpreted by using the concept of **correlation.**
- to distinguish between the concepts of **correlation** and **causality** and to apply them in practice-oriented examples.

Literature

Levine, D.M., K. A. Szabat, and D.F. Stephan. (2016). *Business Statistics: A First Course*, 7th ed., United States: Pearson, **Chapter 3.5**.*

Stinerock, R. (2018). *Statistics with R*. United Kingdom: Sage. **Chapter 1-3**.*

Shira, Joseph (2012). *Statistische Methoden der VWL und BWL*, 4th ed., München et al.: Pearson Studium, **Chapter 1-2**.

Weiers, R. M. (2011). *Introductory Business Statistics*, 7th ed., Canada: Thomson South-Western, **Chapter 3.6**.

*Mandatory literature

I. Correlation vs. Causality

When **pairwise observations** are collected, the data can possibly be constituted in a way such that a **statistical correlation** between the pairs can be calculated.

Correlation does not inevitable mean that a **causal** relationship exists. Thus, in all statistical applications, especially in the examination of characteristics, great attention has to be paid to the coherence in content in order to avoid misuse (**spurious correlation**)!

***Correlation* means a connection between variables, but without a cause-and-effect relation.**

***Causality* means that two variables are connected in a cause-and-effect relationship!**

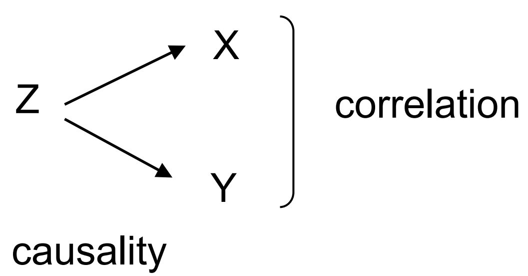
I. Correlation vs. Causality

Investigation A

Finding: There is a connection between shoe size (X) and income (Y).

Conjecture: People with large shoe size earn more.

→ Impact of the cofounder gender (Z)



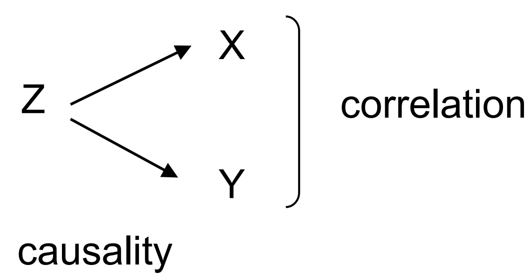
Women have smaller shoe sizes than men. Women have lower incomes than men.

Investigation B

Finding: There is a connection between pariodontosis (X) and heart disease (Y).

Conjecture: People who do not brush their teeth have heart problems.

→ Impact of the cofounder health awareness (Z)



Health-conscious people care for their teeth and take care of their cardiovascular system.

II. Correlation Coefficient

The measurement of the **strength of a statistical relationship** between two metric-scale characteristics is part of correlation analysis.

The **correlation coefficient *(ρ)*** provides a measure of the **strength and direction** of the correlation between two metrically measurable variables, e.g., body height and income, or between the price of oil and the performance of a stock index.

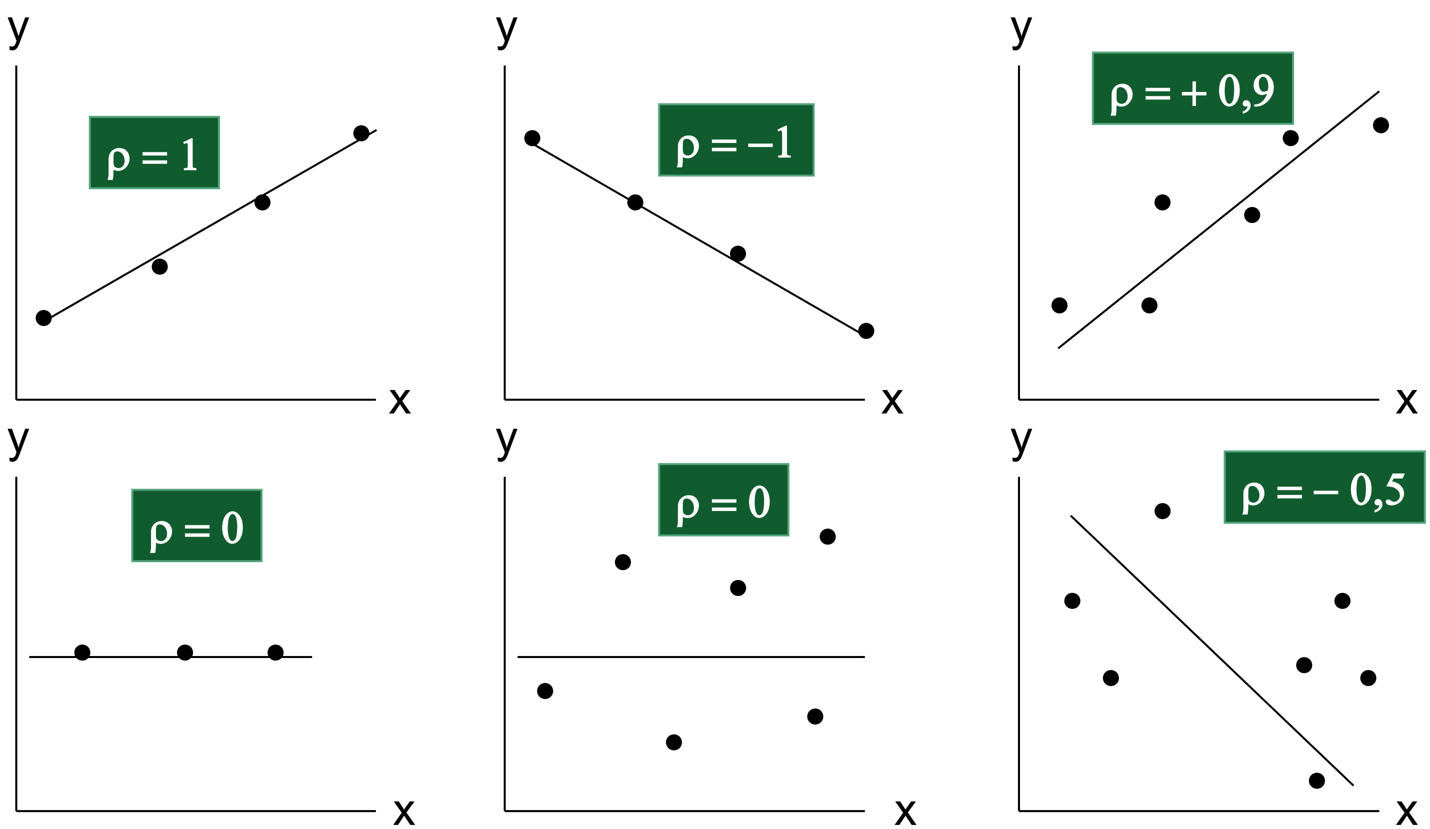
$\boldsymbol{\rho} = -\mathbf{1}$ perfectly negative linear association between the variables (**perfect negative correlation**)

$\boldsymbol{\rho} = \mathbf{0}$ no linear association between the two variables (**uncorrelatedness**)

$\boldsymbol{\rho} = +\mathbf{1}$ perfectly positive linear association between the variables (**perfect positive correlation**)

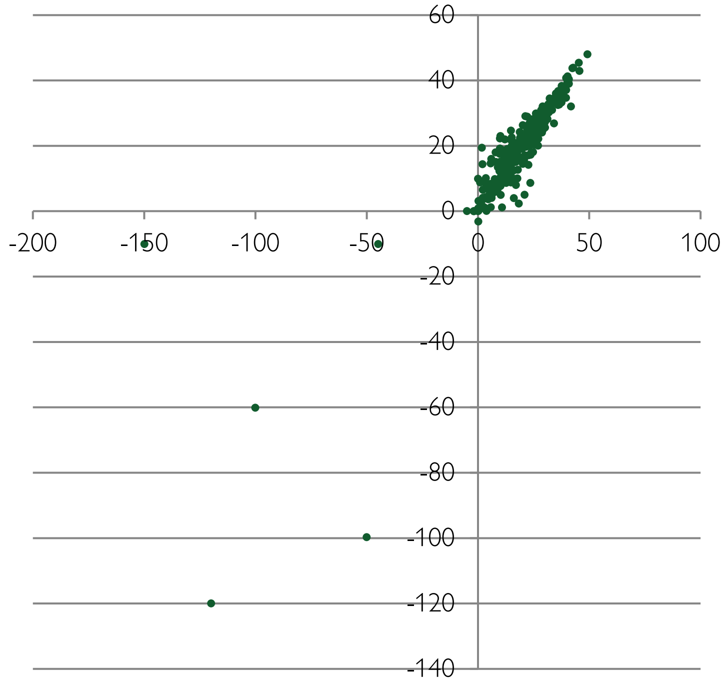
$-\mathbf{1} < \boldsymbol{\rho} < +\mathbf{1}$ the association can more or less be approximated by an upward resp. downward sloping line.

II. Correlation Coefficient

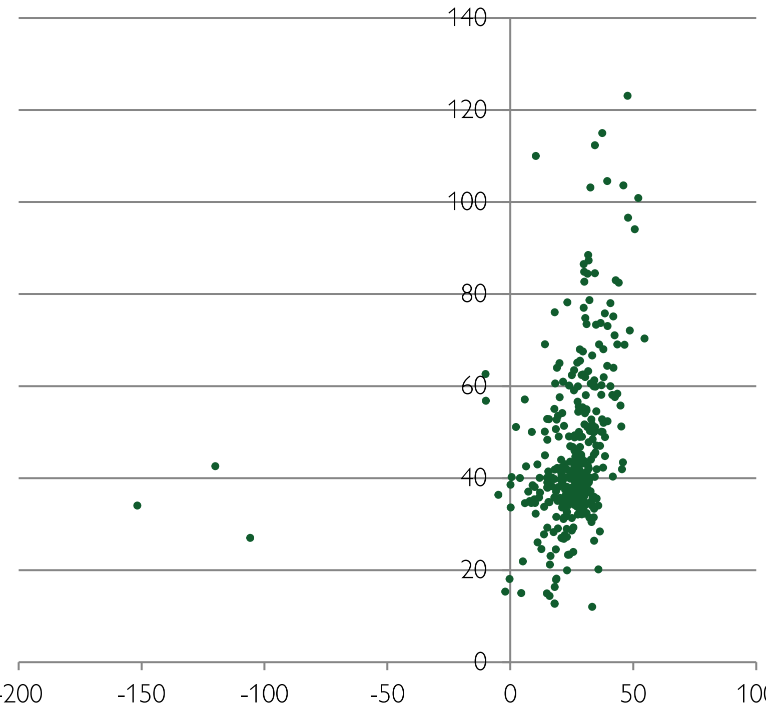


II. Correlation Coefficient (Example)

Correlation = 0.85



Correlation = 0.29



III. Covariance

The **correlation coefficient** is a **standardized measure** (in the interval between -1 and +1) and is therefore independent of the units in which the variables are measured. It is determined by the **covariance**.

The **non-standardized measure** of **covariance** is calculated based on the sum of all products of the mean deviations of both data points.

$$\rho = \frac{\text{covariance}(X, Y)}{\sigma_x \cdot \sigma_y} \qquad \text{covariance}_{x,y} = \frac{\sum_{i=1}^{N} (X_i - \mu_x) \cdot (Y_i - \mu_y)}{N}$$

For **sample data**, the population mean and standard deviation may be replaced by the corresponding sample means and standard deviations:

$$r = \frac{\text{covariance}(x, y)}{s_x s_y} \qquad \text{covariance}_{x,y} = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y})}{n-1}$$

III. Covariance (Example)

i	Size (x_i)	Income (y_i)	$(x_i-\mu_x)$	$(y_i-\mu_y)$	$(x_i-\mu_x)(y_i-\mu_y)$	$(x_i-\mu_x)^2$	$(y_i-\mu_y)^2$
1	185 cm	60000	5	0	0	25	0
2	173 cm	60000	-7	0	0	49	0
3	163 cm	35000	-17	-25000	425000	289	625000000
4	191 cm	90000	11	30000	330000	121	900000000
5	188 cm	55000	8	-5000	-40000	64	25000000
	μ_x=180	***μ_y = 60000***			**715000**	**548**	**1550000000**

Covariance:

$$\text{Covariance}_{x,y} = \frac{715000}{5} = 143000$$

$$\sigma_x = \sqrt{\frac{548}{5}} = 10.5$$

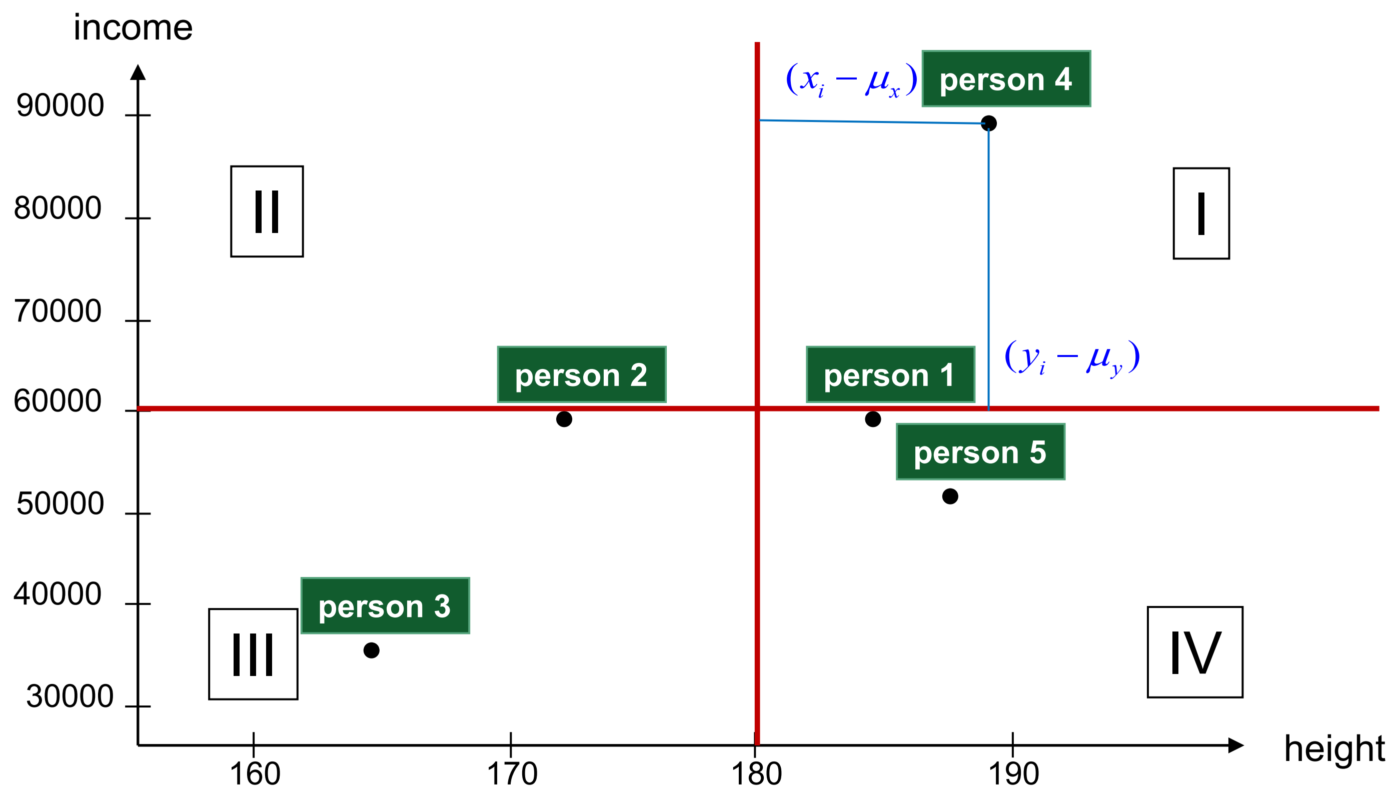
$$\sigma_y = \sqrt{\frac{1550000000}{5}} = 17606.8$$

Correlation Coefficient:

$$\rho = \frac{143000}{10.5 \cdot 17606.8} = 0.77$$

III. Covariance

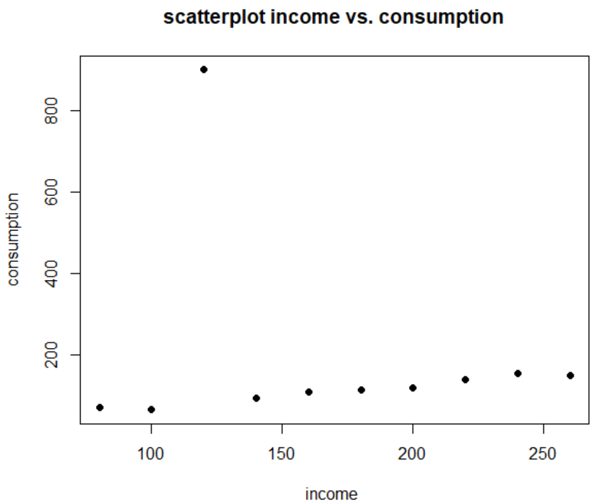
- Points in quadrants **I & III** → **positive contribution** to covariance
- Points in quadrants **II & IV** → **negative contribution** to covariance

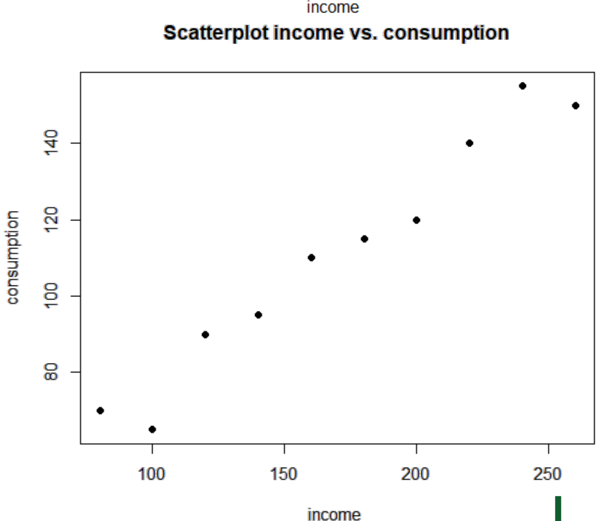


III. Covariance (R-Example)

Open the file "L2-Example_1.R" in R-Studio and reproduce the R-Code.

```
# consider the weekly consumption and weekly income
# of 10 students in a European country
cons<-c(70,65,140,95,150,155,120,900,115,110)
inc<-c(80,100,220,140,260,240,200,120,180,160)
expenses<-data.frame(cons,inc)

# draw a scatter plot and compare income and consumption.
plot(expenses$inc,expenses$cons, main="scatterplot income vs. consumption",
     xlab="income ", ylab="consumption", pch=19)

# the eighth data point is an outlier.
# instead of $90, $900 was  recorded in the data set by mistake.
# correction via index element of the vector:
expenses$cons[8] <- 90

# plot income and consumption against each other again:
plot(expenses$inc,expenses$cons, main="Scatterplot income vs. consumption",
     xlab="income ", ylab="consumption", pch=19)


# covariance
#---------------------------------------
# defining a new covariance function for samples
covariance<-function(x,y) sum((x-mean(x))*(y-mean(y)))/(length(x)-1)
covariance(expenses$inc,expenses$cons)
cov(expenses$inc,expenses$cons) # using the specific function of R


# correlation coefficient
#---------------------------------------
# defining a new function to calculate the correlation coefficient
corrc<-function(x,y) covariance(x,y)/(sd(x)*sd(y))
corrc(expenses$inc,expenses$cons)
cor(expenses$inc,expenses$cons)  # using the specific function of R
```





IV. Portfolio Optimization (R-Example)

Open the file "L2-Example_2.R" in R-Studio! Understand the R-Code and the link to Lectures 1 and 2 of the course "3,120 Corporate Finance".

```
# we want to invest our savings in two out of three shares (IBM, Google and JP Morgan). for this reason,
# we have been monitoring the prices of the three securities since August 2004 and have downloaded the share prices
# at the beginning of each month, the monthly returns as well as the indexed prices. our aim is to find the
# portfolio of two stocks (weighting 50%) that offers the best risk/return trade-off.
# the risk-free interest rate is 0.005.
```

$$Sharpe\ Ratio = \frac{risk\ premium}{standard\ deviation} = \frac{r - r_f}{\sigma}$$

Calculating portfolio returns and portfolio risk

- Calculating the **expected return of a portfolio of stocks** is simple: It is the weighted average of the expected returns of the individual securities in the portfolio.
- The **portfolio risk** (variance) is the sum of all individual **variances** multiplied by their weights squared and all **covariances** multiplied by the weights of both respective stocks.