

Elements of Statistics

Chapter 4:

Measures of association and regression analysis

Ralf Münnich, Jan Pablo Burgard and Florian Ertz

University of Trier
Faculty IV
Economic and Social Statistics Department

Winter term 2021/22

© WiSoStat

The slides are provided as supplementary material by the Economic and Social Statistics Department (WiSoStat) of Faculty IV of the University of Trier for the lecture "Elements of Statistics" in WiSe 2021/22 and the participants are allowed to use them for preparation and reworking. The lecture materials are protected by intellectual property rights. They must not be multiplied, distributed, provided or made publicly accessible - neither fully, nor partially, nor in modified form - without prior written permission. Especially every commercial use is forbidden.

Measures of association for multidimensional distributions

Here: 2 variables with pairs of variates $(x_i; y_i)$

1. Univariate analysis for each variable

$$\bar{x}, \bar{y}, s_x^{*2}, s_y^{*2}, v_x, v_y$$

2. Analysis of the variables' relationship

Problems:

- ▶ Are we able to infer the value of one variable from the value of the other variable?
→ Measurement of (strength of) variables' relationship
- ▶ Is there an *algorithm* (function), governing the variables' relationship?
→ Regression analysis

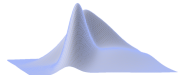
Two-dimensional distributions (1)

| cat. of 1st variable \ cat. of 2nd variable | 1 | ... | k | ... | r | sum |
|---|---------------|-----|---------------|-----|---------------|--------------|
| 1 | n_{11} | ... | n_{1k} | ... | n_{1r} | $n_{1\cdot}$ |
| ... | ... | ... | ... | ... | ... | ... |
| j | n_{j1} | ... | n_{jk} | ... | n_{jr} | $n_{j\cdot}$ |
| ... | ... | ... | ... | ... | ... | ... |
| m | n_{m1} | ... | n_{mk} | ... | n_{mr} | $n_{m\cdot}$ |
| sum | $n_{\cdot 1}$ | ... | $n_{\cdot k}$ | ... | $n_{\cdot r}$ | n |

n_{jk} Joint absolute frequency of j -th value of the first variable and k -th value of the second variable

$n_{j\cdot}$ Absolute frequency of j -th value of the first variable

$n_{\cdot k}$ Absolute frequency of k -th value of the second variable



Two-dimensional distributions (2)

The relative frequencies are given by

$$p_{jk} = \frac{n_{jk}}{n},$$

while

$$p_{k|j} = \frac{n_{jk}}{n_{j.}}$$

are conditional relative frequencies.

Condition: The first variable's realisation is j .

Example 4.1: Unemployment (see Example 2.1):

| Unemployed | [0; 15) | [15; 25) | [25; 45) | [45; 65) | at least 65 | Σ |
|------------|---------|----------|----------|----------|-------------|----------|
| Men | 0 | 124 | 288 | 253 | 1 | 666 |
| Women | 0 | 66 | 235 | 222 | 1 | 524 |
| Σ | 0 | 190 | 523 | 475 | 2 | 1190 |

The absolute marginal distributions were already given in Example 2.1.

The distribution of unemployed females across age groups is given by:

| Unemployed | [0; 15) | [15; 25) | [25; 45) | [45; 65) | at least 65 | Σ |
|------------|---------|----------|----------|----------|-------------|----------|
| Women | 0 | 66 | 235 | 222 | 1 | 524 |

The relative conditional distribution shows the respective share of unemployed women in the different age groups and is reached by dividing the absolute frequencies by the marginal sum (here: 524).

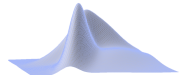
```
load("Example2-8.RData")
round(FQtable["Female",]/FQtable["Female","Sum"],4)
      [0;15) [15;25) [25;45) [45;65) at least 65 Sum
Female      0  0.126  0.4485  0.4237  0.0019    1
```

Empirical distribution function (bivariate case)

For bivariate data, the empirical distribution function is given by

$$F_n(x, y) = \frac{1}{n} \sum_{i=1}^n \mathcal{I}(x_i \leq x \wedge y_i \leq y).$$

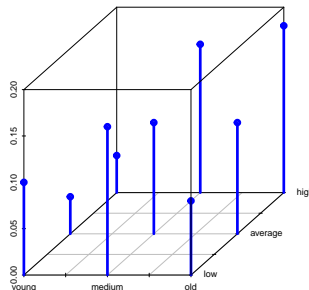
- ▶ for simplicity reasons, the ECDF is displayed in a tabular format
- ▶ ordinal scale is required at least
- ▶ graphical representation is analogous to the univariate case



Example 4.2: Kindergarten (1)

The relationship between age group and the interest in doing handicrafts is investigated in a kindergarten accommodating $n = 50$ children. The results are as follows:

| p_{jk} | young | medium | old |
|----------|-------|--------|------|
| low | 5/50 | 8/50 | 4/50 |
| average | 2/50 | 6/50 | 6/50 |
| high | 2/50 | 8/50 | 9/50 |



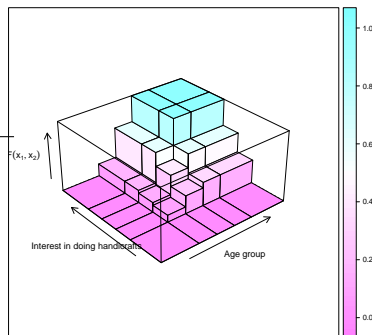
Load data in R:

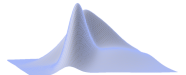
```
load("Example4-2.RData")
```

Example 4.2: Kindergarten (2)

Aggregation of values in both directions up to the relevant cell yields the tabulated empirical distribution function $F_{50}(x, y)$:

| $F_{50}(x, y)$ | young | medium | old |
|----------------|-------|--------|-------|
| low | 5/50 | 13/50 | 17/50 |
| average | 7/50 | 21/50 | 31/50 |
| high | 9/50 | 31/50 | 50/50 |





Example 4.2: Kindergarten (3)

Calculation in R:

```
F_j_k <- t(apply(apply(p_j_k, 2, cumsum), 1, cumsum))  
F_j_k
```

| | young | medium | old |
|---------|-------|--------|------|
| low | 0.10 | 0.26 | 0.34 |
| average | 0.14 | 0.42 | 0.62 |
| high | 0.18 | 0.62 | 1.00 |

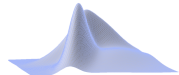
Coefficient of contingency (1)

- ▶ Nominal scale
- ▶ Contingency table (two-dimensional frequencies)

We need a measure which accounts for the relationship between the two variables.

Value $\rightarrow 1$: We can infer the value of one variable from the value of the other variable.

Value $\rightarrow 0$: We cannot even infer a *tendency* for the value of one variable from the value of the other variable.



Example 4.3: Rompers

Relationship between rompers' colours and babies' gender:

| | blue | pink |
|---|------|------|
| m | 10 | 0 |
| f | 0 | 10 |

| | blue | pink |
|---|------|------|
| m | 5 | 5 |
| f | 5 | 5 |

| | blue | pink |
|---|------|------|
| m | 8 | 2 |
| f | 2 | 8 |

Independence of variables

Aim:

Comparison of actual distribution of variables and reference distribution which does not allow any *inference*.

Definition

Two variables are called independent if and only if

$$n_{jk} = \frac{n_{j\cdot} \cdot n_{\cdot k}}{n}$$

holds for all $j = 1, \dots, m$ and $k = 1, \dots, r$.

Problem: The resulting values $\frac{n_{j\cdot} \cdot n_{\cdot k}}{n}$ may not be integers.

Coefficient of contingency (2)

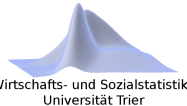
1. Calculation of $n_{jk}^* = \frac{n_{j\cdot} \cdot n_{\cdot k}}{n}$ $j = 1, \dots, m$ $k = 1, \dots, r$.
2. Determination of deviation between actual and theoretical values (independence):

$$\chi^2 = \sum_{j=1}^m \sum_{k=1}^r \frac{(n_{jk} - n_{jk}^*)^2}{n_{jk}^*}$$

3. $K = \sqrt{\frac{\chi^2}{n + \chi^2}}$; $0 \leq K < 1$

4. Standardisation:

$$K_* = \frac{K}{K_{\max}}; \quad K_{\max} = \sqrt{\frac{M-1}{M}}; \quad M = \min(m, r)$$



Coefficient of contingency (3)

K_* is called standardised coefficient of contingency. We have: $0 \leq K_* \leq 1$.

Independence $\Rightarrow K_* = 0$

Perfect relation $\Rightarrow K_* = 1$

Example 4.4: Field of study and gender (1)

We are interested in the relation between field of study and gender.

x : Field **B**usiness administration, **E**conomics, **G**eography

y : Gender m, f

| n_{jk} | m | f | |
|----------|---|---|----|
| B | 2 | 1 | 3 |
| E | 2 | 2 | 4 |
| G | 1 | 2 | 3 |
| | 5 | 5 | 10 |

→

| n_{jk}^* | m | f | |
|------------|-----|-----|----|
| B | 1.5 | 1.5 | 3 |
| E | 2 | 2 | 4 |
| G | 1.5 | 1.5 | 3 |
| | 5 | 5 | 10 |

Example 4.4: Field of Study and gender (2)

Calculations in R:

```
load("Example4-4.RData")
addmargins(n_j_k)
```

| | m | f | Sum |
|-----|---|---|-----|
| B | 2 | 1 | 3 |
| E | 2 | 2 | 4 |
| G | 1 | 2 | 3 |
| Sum | 5 | 5 | 10 |

```
n_j_k_star <- margin.table(n_j_k,1) %*%
               t(margin.table(n_j_k,2))/margin.table(n_j_k)
n_j_k_star
```

| | m | f |
|---|-----|-----|
| B | 1.5 | 1.5 |
| E | 2.0 | 2.0 |
| G | 1.5 | 1.5 |

Example 4.4: Field of Study and gender (3)

$$\begin{aligned}\chi^2 &= \frac{(2 - 1.5)^2}{1.5} + \frac{(1 - 1.5)^2}{1.5} + \frac{(2 - 2)^2}{2} + \frac{(2 - 2)^2}{2} + \\ &\quad \frac{(1 - 1.5)^2}{1.5} + \frac{(2 - 1.5)^2}{1.5} \\ &= \frac{1/2^2}{3/2} \cdot 4 = \frac{2}{3}\end{aligned}$$

Calculation of χ^2 in R:

```
chisq <- summary(n_j_k)$statistic  
chisq
```

```
[1] 0.6666667
```

Note that in R, the object `n_j_k` above has to have the structure table (Checking possible with `str()`).

Example 4.4: Field of Study and gender (4)

$$\Rightarrow K = \sqrt{\frac{(2/3)}{10 + (2/3)}} = \sqrt{\frac{1}{16}} = \frac{1}{4}$$

With $M = 2$ and $K_{max} = \sqrt{\frac{1}{2}}$ we have $K_* = \frac{1}{4} \cdot \sqrt{2} = 0.3536$.

Calculations in R:

```
KK <- sqrt(chisq/(sum(n_j_k) + chisq))
M <- min(dim(n_j_k))
K_max <- sqrt((M-1)/M)
K_star <- KK/K_max
```

KK

[1] 0.25

M

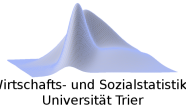
[1] 2

K_max

[1] 0.7071068

K_star

[1] 0.3535534



Pearson's Φ and Cramer's coefficient of contingency

Pearson's Φ -coefficient is defined as:

$$\sqrt{\frac{\chi^2}{n}}.$$

We have $0 \leq \Phi \leq \sqrt{M-1}$, where in case of $\min(m, r) = 2$ we have $M-1 = 1$.

Cramér's coefficient of contingency (Cramér's V) is defined as:

$$V = \sqrt{\frac{\chi^2}{n \cdot (M-1)}}.$$

We have: $0 \leq V \leq 1$. In case of $\min(m, r) = 2$, we have $\Phi = V$.

Both coefficients may only be used for nominal variables.

Rank correlation coefficient of Spearman (1)

- ▶ Ordinal scale
- ▶ Each rank is unique

Additional information compared to coefficient of contingency:

Positive correlation The higher the value of one variable, the higher is the value of the other variable

Negative correlation The higher the value of one variable, the lower is the value of the other variable

Rank correlation coefficient of Spearman (2)

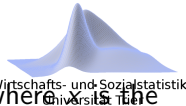
Let x and y have at least ordinal scaling and no duplicated values in x_i and y_i , respectively. The rank correlation coefficient of Spearman is then given by

$$r_{sp} = 1 - \frac{6 \sum_{i=1}^n (\text{Rg}(x_i) - \text{Rg}(y_i))^2}{n(n^2 - 1)}$$

$r_{sp} = +1$: All ranks are identical

$r_{sp} = -1$: All ranks are contrary to each other

In order to determine the rank of an attribute in R we can use the function `rank()` .



Example 4.5: Alpine skiing

Let the results for a combined alpine skiing event be given, where x is the time measured for downhill and y is the time measured for slalom.

| | | | | | |
|-------------------------|---|---|---|---|---|
| $Rg(x_i)$ | 3 | 1 | 5 | 2 | 4 |
| $Rg(y_i)$ | 1 | 3 | 5 | 4 | 2 |
| $(Rg(x_i) - Rg(y_i))^2$ | 4 | 4 | 0 | 4 | 4 |

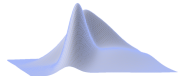
$$r_{sp} = 1 - \frac{6 \cdot 16}{5(25 - 1)} = 1 - \frac{4}{5} = 0.2$$

Calculation of r_{sp} in R:

```
load("Example4-5.RData")
cor_SP <- cor(Rg_x5_5, Rg_y5_5, method = "spearman")
cor_SP
```

```
[1] 0.2
```

Does it matter that the rankings in both disciplines are opposed to the time ranks used here?



Tied ranks

If there are ties, we may replace ranks for identical values by a mean rank of the observations affected. Then we can use:

$$r_{sp} = \frac{\sum_{i=1}^n \text{Rk}(x_i) \cdot \text{Rk}(y_i) - \frac{1}{n} \sum_{i=1}^n \text{Rk}(x_i) \sum_{i=1}^n \text{Rk}(y_i)}{\sqrt{\sum_{i=1}^n \text{Rk}(x_i)^2 - \left(\frac{1}{n} \sum_{i=1}^n \text{Rk}(x_i)\right)^2} \cdot \sqrt{\sum_{i=1}^n \text{Rk}(y_i)^2 - \left(\frac{1}{n} \sum_{i=1}^n \text{Rk}(y_i)\right)^2}}.$$

This matches the correlation coefficient of Bravais-Pearson for the ranks of the observations (instead of their values).

For contingency tables we use:

$$r_{sp} = \frac{\sum_{j=1}^m \sum_{k=1}^r \text{Rk}(x_j) \text{Rk}(y_k) n_{jk} - \frac{1}{n} \sum_{j=1}^m \text{Rk}(x_j) n_{j.} \sum_{k=1}^r \text{Rk}(y_k) n_{.k}}{\sqrt{\left(\sum_{j=1}^m \text{Rk}(x_j)^2 n_{j.} - \left(\frac{1}{n} \sum_{j=1}^m \text{Rk}(x_j) n_{j.}\right)^2\right) \cdot \left(\sum_{k=1}^r \text{Rk}(y_k)^2 n_{.k} - \left(\frac{1}{n} \sum_{k=1}^r \text{Rk}(y_k) n_{.k}\right)^2\right)}}.$$

Example 4.6: Scholarships (1)

Two reviewers had to give their assessment of $n = 50$ students applying for a scholarship. In the final round, the following ratings could be awarded: *excellent* (A), *very good* (B) and *good* (C). The following table contains the results:

| | | Reviewer II | | |
|------------|---|-------------|----|----|
| | | A | B | C |
| Reviewer I | A | 3 | 2 | 0 |
| | B | 1 | 12 | 2 |
| | C | 0 | 4 | 26 |

Due to the large number of ties, a specification of the mean ranks is required at first.

For reviewer I, we have:

Rating A: Rank 1 – 5, 5 times a mean rank of 3

Rating B: Rank 6 – 20, 15 times a mean rank of 13

Rating C: Rank 21 – 50, 30 times a mean rank of 35.5

Example 5.6: Scholarships (2)

Analogously, we get the mean ranks 2.5, 13.5 and 36.5 for Reviewer II.

Calculation in R:

```
load("Example4-6.RData")  
Rg_x5_6_mean <- c(3.0, 13.0, 35.5)  
Rg_y5_6_mean <- c(2.5, 13.5, 36.5)
```

Finally, using the formula for contingency tables, we get:

$$r_{\text{sp}} = \frac{38797.5 - \frac{1}{50} \cdot 1275 \cdot 1275}{\sqrt{7875 \cdot 8096}} = \frac{6285}{7984.735} = 0.7871.$$

Example 4.6: Scholarships (3)

Calculation of r_{sp} in R:

```
n <- sum(n_j_k)

Rx_Ry_sum <- sum(n_j_k[1, ]*Rg_x5_6_mean[1]*Rg_y5_6_mean) +
  sum(n_j_k[2, ]*Rg_x5_6_mean[2]*Rg_y5_6_mean) +
  sum(n_j_k[3, ]*Rg_x5_6_mean[3]*Rg_y5_6_mean)

Rx_sum <- sum(n_j_k * Rg_x5_6_mean)
Ry_sum <- sum(t(n_j_k) * Rg_y5_6_mean)
Rx_2 <- sum(Rg_x5_6_mean^2 * margin.table(n_j_k,1))
Ry_2 <- sum(Rg_y5_6_mean^2 * margin.table(n_j_k,2))

cor_SP <- (Rx_Ry_sum - 1/n * Rx_sum * Ry_sum) /
  (sqrt(Rx_2 - 1/n * Rx_sum^2) *
   sqrt(Ry_2 - 1/n * Ry_sum^2))

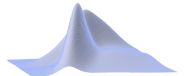
round(cor_SP, 4)

[1] 0.7871
```

Example 4.6: Scholarships (4)

Alternatively, we could use individual data on ranks and the first formula to reach the same result ($Rk(x_i); Rk(y_i)$):

| | | |
|---------------------------------|----------------------------------|-----------------------------------|
| $\underbrace{(3; 2.5)}_{3x}$ | $\underbrace{(3; 13.5)}_{2x}$ | $\underbrace{(3; 36.5)}_{0x}$ |
| $\underbrace{(13; 2.5)}_{1x}$ | $\underbrace{(13; 13.5)}_{12x}$ | $\underbrace{(13; 36.5)}_{2x}$ |
| $\underbrace{(35.5; 2.5)}_{0x}$ | $\underbrace{(35.5; 13.5)}_{4x}$ | $\underbrace{(35.5; 36.5)}_{26x}$ |



Preliminary remarks on τ and γ measures

We define the following two relationship patterns for two pairs of values $(x_i; y_i)$ and $(x_j; y_j)$:

Concordant pair: We have $x_i < x_j$ and $y_i < y_j$ or $x_i > x_j$ and $y_i > y_j$, so that the comparison of the components of the pairs is unidirectional.

Discordant pair: We have $x_i < x_j$ and $y_i > y_j$ or $x_i > x_j$ and $y_i < y_j$, so that the comparison of the components of the pairs is counterdirectional.

The number of concordant and discordant pairs is labelled n_c and n_d , respectively.

Additionally, we may have to take ties into account. T_x is the number of ties of the first variable and T_y is the number of ties of the second variable.

Kendall's τ and Goodman and Kruskal's γ

We have:

$$\tau_a = \frac{n_c - n_d}{\frac{1}{2} \cdot n \cdot (n - 1)}$$

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_c + n_d + T_x) \cdot (n_c + n_d + T_y)}}$$

$$\tau_c = \frac{n_c - n_d}{\frac{1}{2} \cdot n^2 \cdot \frac{M-1}{M}} = \frac{2M \cdot (n_c - n_d)}{n^2 \cdot (M - 1)}$$

$$\gamma = \frac{n_c - n_d}{n_c + n_d}$$

τ_a requires contingency tables without ties. τ_b is commonly used, but only takes on the value 1 for quadratic contingency tables. τ_c accounts for differing numbers of rows and columns.

Example 4.7: see Ex. 4.6 (1)

In order to calculate the τ and *gamma* measures, we have to find the concordant and discordant pairs as well as the ties in X and Y . We get:

$$n_c = 3(12 + 2 + 4 + 26) + 2(2 + 26) + 1(4 + 26) + 12 \cdot 26 = 530$$

$$n_d = 2(1 + 0) + 0 + 12 \cdot 0 + 2(0 + 4) = 10$$

Calculation of n_c and n_d in R:

```
load("Example4-6.RData")
```

```
nc <- sum(n_j_k[1,1]*n_j_k[2:3,2:3], n_j_k[1,2]*n_j_k[2:3,3],  
          n_j_k[2,1]*n_j_k[3,2:3], n_j_k[2,2]*n_j_k[3,3])
```

```
nd <- sum(n_j_k[3,1]*n_j_k[1:2,2:3], n_j_k[2,1]*n_j_k[1,2:3],  
          n_j_k[3,2]*n_j_k[1:2,3], n_j_k[3,1]*n_j_k[2,2])
```

```
nc
```

```
[1] 530
```

```
nd
```

```
[1] 10
```

Example 4.7: see Ex. 4.6 (2)

For the ties, we get:

$$T_x = 3(2 + 0) + 1(12 + 2) + 0 + 0 + 12 \cdot 2 + 4 \cdot 26 = 148$$

$$T_y = 3(1 + 0) + 2(12 + 4) + 0 + 0 + 12 \cdot 4 + 2 \cdot 26 = 135$$

Calculation of T_x and T_y in R:

```
Tx <- sum(n_j_k[,1]*n_j_k[,2:3], n_j_k[,2]*n_j_k[,3])  
Ty <- sum(t(n_j_k)[,1] * t(n_j_k)[,2:3],  
          t(n_j_k)[,2] * t(n_j_k)[,3])
```

```
Tx
```

```
[1] 148
```

```
Ty
```

```
[1] 135
```

Example 4.7: see Ex. 4.6 (3)

We finally reach:

$$\tau_a = \frac{530-10}{\frac{1}{2} \cdot 50 \cdot (50-1)} = 0.4245$$

$$\tau_b = \frac{530-10}{\sqrt{(530+10+148) \cdot (530+10+135)}} = 0.7631$$

$$\tau_c = \frac{530-10}{\frac{1}{2} \cdot 50^2 \cdot \frac{3-1}{3}} = 0.624$$

$$\gamma = \frac{530-10}{530+10} = 0.9630$$

In Example 4.6, the result was $r_{sp} = 0.7871$.

Example 4.7: see Ex. 4.6 (4)

Calculation of τ_a , τ_b , τ_c and γ in R:

```
n <- sum(n_j_k)
M <- min(dim(n_j_k))

tau_a <- (nc - nd) / (1/2 * n * (n - 1))
tau_b <- (nc - nd) / sqrt((nc + nd + Tx) * (nc + nd + Ty))
tau_c <- 2 * M * (nc - nd) / (n^2 * (M - 1))
gamma <- (nc - nd) / (nc + nd)
```

```
round(tau_a, 4)
```

```
[1] 0.4245
```

```
round(tau_b, 4)
```

```
[1] 0.7631
```

```
round(tau_c, 4)
```

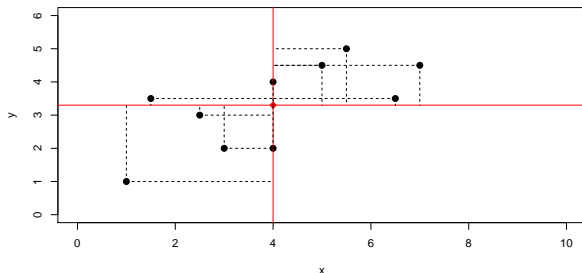
```
[1] 0.624
```

```
round(gamma, 4)
```

```
[1] 0.963
```

Correlation and covariance

Correlation vs. linear regression

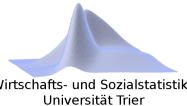


- ▶ Positive correlation (see figure)
- ▶ Negative correlation
- ▶ Standardisation

The covariance of two metric variables x and y is given by

$$s_{xy}^* = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

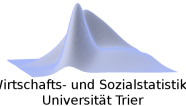
Some kind of standardisation is needed!



Correlation coefficient of Bravais-Pearson (1)

For metrically scaled variables x and y with positive variances of x and y , the correlation coefficient of Bravais-Pearson is defined as

$$r_{xy} = \frac{s_{xy}^*}{s_x^* \cdot s_y^*} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \cdot \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}} .$$



Correlation coefficient of Bravais-Pearson (2)

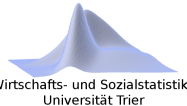
Properties:

1. $-1 \leq r_{xy} \leq 1$
2. Special cases: $r_{xy} = -1; 0; +1$
3. Transformation

$$u_i = a_0 + a_1 \cdot x_i; v_i = b_0 + b_1 \cdot y_i$$

$$r_{uv} = \text{sgn}(a_1 \cdot b_1) \cdot r_{xy}$$

Frequently, only the algebraic sign ($r_{xy} \geq 0$) is of interest in economics.



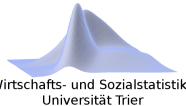
Problems of the correlation coefficient (1)

Non-linear relationships

Let $x_i = -2; -1; 1; 2$ and $y_i = x_i^2$. The resulting correlation coefficient is

$$r_{xy} = 0 \quad .$$

There is a quadratic relationship in the data, which is not comprehended by the correlation coefficient!



Problems of the correlation coefficient (2)

Correlation and causality

Interpretation of $r_{xy} = 0.98$:

A statistical interrelation does not necessarily indicate a theoretical interrelation.

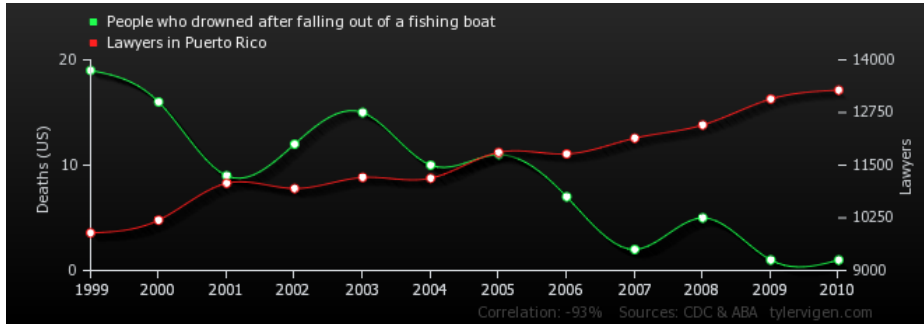
- ▶ Presidential elections / Superbowl in the USA

(<http://www.theguardian.com/sport/blog/2012/feb/01/super-bowl-ology-science-impotence-2012>)

- ▶ Beer consumption / Count of unemployed people per month

→ Spurious correlation

Spurious correlation



Source:

Tyler Vigen – spurious correlations (2019).

http://tylervigen.com/view_correlation?id=30074

Example 4.8: Correlation coefficient (1)

The following table contains information on $n = 10$ units and two variables:

| i | x_i | y_i | x_i^2 | y_i^2 | $x_i \cdot y_i$ |
|----------|-------|-------|---------|---------|-----------------|
| 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 1.5 | 3.5 | 2.25 | 12.25 | 5.25 |
| 3 | 2.5 | 3 | 6.25 | 9 | 7.5 |
| 4 | 3 | 2 | 9 | 4 | 6 |
| 5 | 4 | 2 | 16 | 4 | 8 |
| 6 | 4 | 4 | 16 | 16 | 16 |
| 7 | 5 | 4.5 | 25 | 20.25 | 22.5 |
| 8 | 5.5 | 5 | 30.25 | 25 | 27.5 |
| 9 | 6.5 | 3.5 | 42.25 | 12.25 | 22.75 |
| 10 | 7 | 4.5 | 49 | 20.25 | 31.5 |
| Σ | 40 | 33 | 197 | 124 | 148 |

Example 4.8: Correlation coefficient (2)

Data input in R:

```
x4_8 <- c(1, 1.5, 2.5, 3, 4, 4, 5, 5.5, 6.5, 7)  
y4_8 <- c(1, 3.5, 3.2, 2, 4, 4.5, 5, 3.5, 4.5)
```

At first, the univariate measures needed are computed ...

$$\bar{x} = 40/10 = 4$$

$$\bar{y} = 33/10 = 3.3$$

$$s_x^{*2} = \frac{1}{10} \cdot 197 - 4^2 = 3.7$$

$$s_y^{*2} = \frac{1}{10} \cdot 124 - 3.3^2 = 1.51$$

Example 4.8: Correlation coefficient (3)

... thereupon, these are combined in order to determine a result for the covariance and correlation coefficient, respectively:

$$s_{xy}^* = \frac{1}{10} \cdot 148 - 4 \cdot 3.3 = 1.6$$
$$r_{xy} = \frac{1.6}{\sqrt{3.7 \cdot 1.51}} = 0.6769$$

Calculation of s_{xy}^* and r_{xy} in R:

```
n <- length(x4_8)
```

```
x_y_cov <- (n-1)/n * cov(x4_8, y4_8)
```

```
cor_BP <- cor(x4_8, y4_8, method = "pearson")
```

```
x_y_cov
```

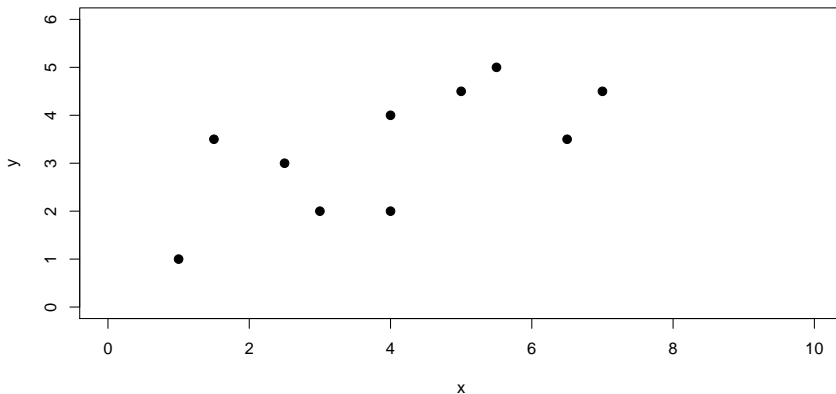
```
[1] 1.6
```

```
round(cor_BP, 4)
```

```
[1] 0.6769
```

Example 4.8: Correlation coefficient (4)

Correlation vs. linear regression

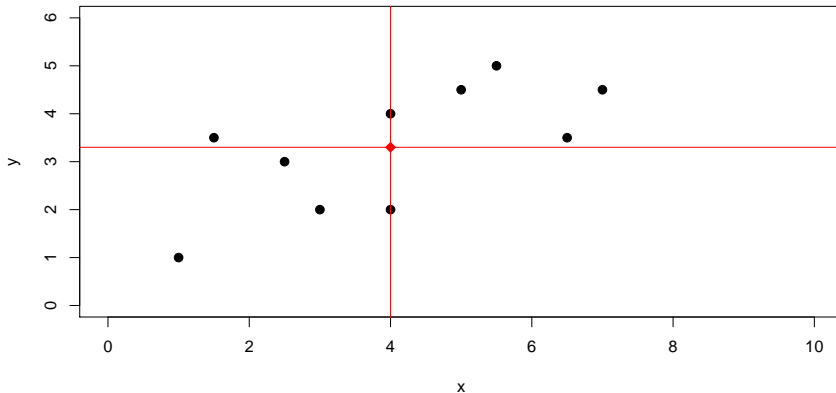


In R:

```
plot(x4_8,y4_8,xlim=c(0,10),ylim=c(0,6),xlab="x",ylab="y",  
type="p",pch=16)
```

Example 4.8: Correlation coefficient (5)

Correlation vs. linear regression

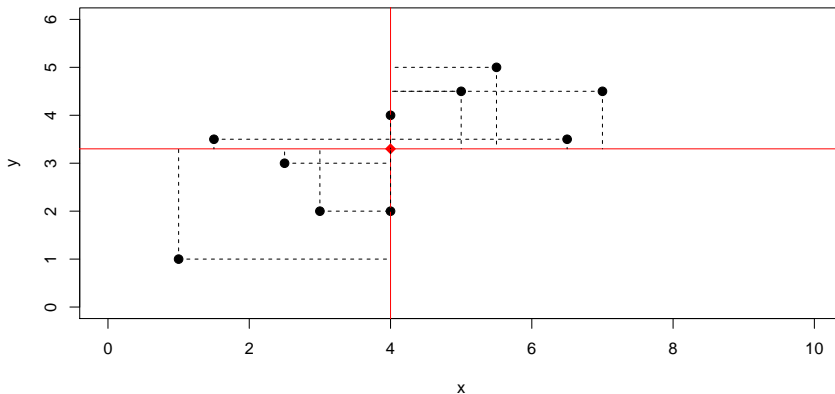


In R:

```
abline(v = mean(x4_8), col = "red")  
abline(h = mean(y4_8), col = "red")  
points(x = mean(x4_8), y = mean(y4_8), col = "red", pch=19)
```

Example 4.8: Correlation coefficient (6)

Correlation vs. linear regression



Linear regression

x and y are continuous metric variables.

x is the so-called independent variable.

y is the so-called dependent variable.

We are looking for a relationship:

$$y = f(x).$$

→ Dependency analysis

Linear relationships are of primary interest: $y = a + b \cdot x$.

Problem: The observations typically do not lie on a line. Why is that the case?

Simple linear regression

We assume a linear model:

$$Y = \alpha + \beta \cdot X \quad .$$

There might be more than one value of y that is corresponding to a certain value of x (random error). We use capital letters when talking about models.

We would like to determine the parameters a and b of

$$\hat{y}_i = a + b \cdot x_i \quad ,$$

where \hat{y}_i is the vertical projection of y_i to the regression line. $e_i = y_i - \hat{y}_i$ is the residual corresponding to observation x_i . The method of ordinary least squares (OLS) determines estimates for the parameters a and b :

$$Z(a, b) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (a + b \cdot x_i))^2 \rightarrow \min$$

Solution to minimisation problem

Using the normal equations

$$n \cdot a + b \cdot \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \quad (1\text{st normal equation})$$

$$a \cdot \sum_{i=1}^n x_i + b \cdot \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i \cdot y_i \quad (2\text{nd normal equation})$$

we get

$$b = \frac{\sum_{i=1}^n x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = \frac{s_{xy}^*}{s_x^{*2}}$$

and $a = \bar{y} - b \cdot \bar{x}$. Finally, for the sample regression line we have:

$$\hat{y} = \bar{y} + \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot (x - \bar{x})$$

Coefficient of determination and properties of OLS

The measure $r_{xy}^2 = \frac{s_{\hat{y}}^2}{s_y^2}$ is called coefficient of determination. For simple linear regression, it equals the squared correlation coefficient of Bravais-Pearson (for x and y). The special cases of $r_{xy}^2 = 0$ and $r_{xy}^2 = 1$ are particularly interesting.

Properties:

1. Centre of gravity: (\bar{x}, \bar{y}) is a point on the sample regression line.
2. The residuals cancel each other out: $\sum_{i=1}^n e_i = \sum_{i=1}^n (y_i - \hat{y}_i) = 0$.
3. $b = r_{xy} \cdot \frac{s_y^*}{s_x^*} \quad \left(= \frac{s_{xy}^*}{s_x^* \cdot s_y^*} \cdot \frac{s_y^*}{s_x^*} = \frac{s_{xy}^*}{s_x^{*2}} \right)$

Example 4.9: see Ex. 4.8 (1)

We get:

$$b = \frac{1.6}{3.7} = 0.6769 \cdot \sqrt{\frac{1.51}{3.7}} = 0.4324$$

$$a = 3.3 - 0.4324 \cdot 4 = 1.5703$$

$$\hat{y}(8) = 1.5703 + 0.4324 \cdot 8 = 5.0297$$

$$r^2 = 0.6769^2 = 0.4582$$

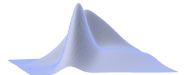
$$\hat{y}(9) = 1.5703 + 0.4324 \cdot 9 = 5.4622$$

Calculations in R:

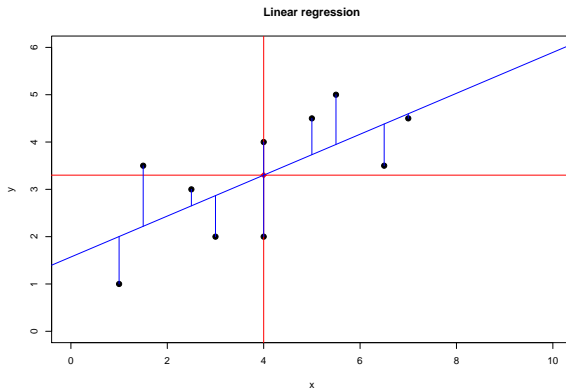
```
reg_mod <- lm(y4_8 ~ x4_8)

a <- summary(reg_mod)$coeff[1,1]
b <- summary(reg_mod)$coeff[2,1]
r_2 <- summary(reg_mod)$r.squared
y_hat_8 <- a + b * 8
y_hat_9 <- a + b * 9
```

| <code>round(a, 4)</code> | <code>round(b, 4)</code> | <code>round(r_2, 4)</code> | <code>y_hat_8</code> | <code>y_hat_9</code> |
|--------------------------|--------------------------|----------------------------|----------------------|----------------------|
| [1] 1.5703 | [1] 0.4324 | [1] 0.4582 | [1] 5.02973 | [1] 5.462162 |



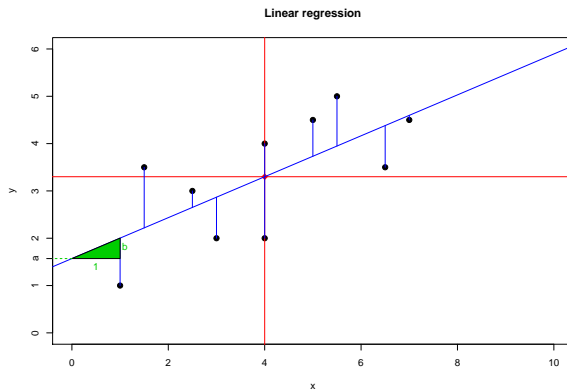
Example 4.9: see Ex. 4.8 (2)

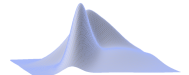


In R:

```
plot(x4_8, y4_8, xlim=c(0,10), ylim=c(0,6), type="p", pch=16)
abline(v = mean(x4_8), col = "red")
abline(h = mean(y4_8), col = "red")
points(x = mean(x4_8), y = mean(y4_8), col = "red", pch=19)
abline(reg_mod, col="blue")
```

Example 4.9: see Ex. 4.8 (3)





Example 4.9: see Ex. 4.8 (4)

