Fachhochschule
Dortmund
University of Applied Sciences and Arts

Data Science

06: Bivariate characteristics

Klaus Kaiser
WiSe 2024 / 2035

Data Science

Recap Frequencies

A histogram is one way to present frequencies of continuous / metric data. For this, the data is classified into classes (see one of the previous slides) with class boundaries $c_0, \ldots, c_k$. Then, the j^{th} class, for $j = 1, \ldots, k$, is represented by a box starting from c_{j-1} to c_j (width equals to $d_j = c_j - c_{j-1}$) and height

$$g_j := \frac{f_j}{d_j} = \frac{f_j}{c_j - c_{j-1}}.$$

- The area of the box equals to the frequency of the class:

$$d_j \cdot g_j = d_j \frac{f_j}{d_j} = f_j$$

- The complete area of all boxes of the histogram equals to 1.
- If every class width has the same size, the height g_j is proportional to the relative frequency f_j.

For an ordinal or continuous variable X with observations $x_1 \dots x_n$ with different values $a_1, \dots, a_l$. We define the **absolute cumulative frequency** as

$$H(x) = \sum_{a_j \leq x} h(a_j)$$

and the **empirical distribution function** $F_n : \mathbb{R} \to [0, 1]$ as

$$F_n(x) = \sum_{a_j \leq x} f(a_j) = \frac{H(x)}{n}.$$

Data Science

Recap Motivation

To describe a set of observations it could be useful to reduce in on one or few characteristic values - or **central tendencies**.

- Which value is the most common?
- Which value is the one in the middle?
- Which value is the averaged one?

Depending on the observations and question different values could be interesting.

Data Science
Recap Overview

Types of data

	nominal	ordinal	metric
Mode	✓	✓	(✓)
Median / quartile	✗	✓	✓
Arithmetic mean	✗	✗	✓

- Using the mode value for metric data is not recommended

Outlier

	Mode	Median / quartile	Arithmetic mean
Robustness	✓	✓	✗

Data Science

Recap Box plots

Data Science

Recap Range

Let $x_1, \ldots, x_n$ be observations of a metric variable X. Then, we define the range r as the difference between the minimal $x_{(1)}$ and the maximal $x_{(n)}$ value:

$$r = x_{(n)} - x_{(1)}$$

- r only depends on the minimal and maximal value. All further information on $x_1, \ldots, x_n$ are lost.
- r is not robust concerning outlier!

Data Science

Recap Quartile range

Let $x_1, \ldots, x_n$ be observations of a metric variable X. Then, we define the quartile range (also interquartile range) qd as the difference between the upper quartile $x_{0.75}$ and lower quartile $x_{0.25}$ value:

$$qd = x_{0.75} - x_{0.25}$$

- qd is robust concerning outlier
- 50% of the "central" observations are given between $x_{0.25}$ and $x_{0.75}$

Data Science

Today

Empirical variance and **bivariate characteristics**

Statistical dispersion

Let $x_1, \ldots, x_n$ be observations of a metric variable X and $\bar{x}$ the corresponding arithmetic mean. Then we define the **empirical variance** of $x_1, \ldots, x_n$ as the mean squared deviation given by

$$\tilde{s}^2 = \frac{1}{n}\left[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\right] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Furthermore, we define

$$\tilde{s} = \sqrt{\tilde{s}^2}$$

as the **empirical standard deviation**.

- The suffix *empirical* is used to differentiate it from the variance of a random variable (later!). The word shows, that the value was computed on concrete data.
- Often (especially in software products), the sampling variance, given by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

  is preferred. Note that the difference is small for large n.
- The empirical variance can also be computed by

$$\tilde{s}^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

Data Science

Statistical dispersion Quartile range

Task: What is the empirical variance in the following example?

Daily temperature in °C at 10 o'clock (metric)

11.2 13.3 14.1 13.7 12.2 11.3 9.9

step	result
$\bar{x}$	12.24
$x_i - \bar{x}$	-1.04, 1.06, 1.86, 1.46, -0.04, -0.94, -2.34
$(x_i - \bar{x})^2$	1.09, 1.12, 3.45, 2.12, 0.00, 0.89, 5.49
$\sum_{i=1}^{n} (x_i - \bar{x})^2$	14.16
$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$	2.02

Steiners theorem

Let $x_1, \ldots, x_n \in \mathbb{R}$, $a \in \mathbb{R}$. Then, there holds

$$\sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - a)^2$$

Let $x_1, \ldots, x_n$ be observations of a metric variable X and $\tilde{s}_X^2$ the corresponding empirical variance. Furthermore, let $\tilde{s}_Y^2$ be the empirical variance of the variable Y which is given by the linear transformation

$$y_i = ax_i + b \text{ for } i \in \{1, \ldots, n\}$$

with $a, b \in \mathbb{R}$. Then, there holds:

$$\tilde{s}_Y^2 = a^2 \tilde{s}_X^2.$$

Data Science

Statistical dispersion Example

data	**mean**	**median**	0.25**-quartile**	0.75**-quantile**
Federal states	5169455.13	3262270.5	2035008.5	6641274.75
Large cities	330394.33	187119	128619	313420.5

data	**range**	**quartile range**	**empirical variance**	**empirical standard deviation**
Federal states	17197285	4606266.25	22795310650229.98	4774443.49
Large cities	3682192	184801.5	229138903806.70	478684.56

Bivariate characteristics

Data Science

Bivariate characteristics

So far: Describing one variable with plots, frequencies, tendencies ... But, we have already seen that data often consists of multiple variables.

Name	Location	Growth
Berlin	North	growing strongly
⋮	⋮	⋮
Dortmund	North	growing
⋮	⋮	⋮

If two variables are paired, i.e. they stem from the same observation, we call them **bivariate**. The more general case is **multivariate data**, i.e. multiple variables which stem from the same observation.

Data Science

Bivariate characteristics

Is there a ”connection” or **correlation** between these variables? If x then y?

Task: Can you image variables which have a correlation?

Data Science

Bivariate characteristics

Is there a "connection" or **correlation** between these variables? If x then y?

Task: Can you image variables which have a correlation?

Examples:

- A larger person might have a larger shoe size
- A faster car might have a longer braking distance
- A younger house might be more expansive
- ...

How can we formalize such a correlation?

Describing correlation of two variables with a dimension number (correlation coefficient). There are different ways to compute a correlation coefficient depending on the type of variables.

- Nominal characteristics: Measures of association
- Ordinal characteristics: Rank correlation coefficients
- Metric characteristics: Correlation coefficients

We the bivariate variable (X, Y), which consists of two variables X and Y with $n \in \mathbb{N}$ observations $x_1, \ldots x_n$ and $y_1, \ldots, y_n$.

- Every entry of X equals to one of $l_X \in \mathbb{N}$ with $l_X \leq n$ different values $a_1, \ldots, a_l$.
- Every entry of Y equals to one of $l_Y \in \mathbb{N}$ with $l_Y \leq n$ different values $b_1, \ldots, b_{l_Y}$.

(X,Y)	**observation**	(x_1, y_1)	(x_2, y_2)	$\ldots$	(x_{n-1}, y_{n-1})	(x_n, y_n)
	value	(a_1, b_1)	(a_1, b_2)	$\ldots$	(a_{l_X-1}, b_{l_Y})	(a_{l_X}, b_{l_Y})
X	**observation**	x_1	x_2	$\ldots$	x_{n-1}	x_n
	value	a_1	a_1	$\ldots$	a_{l_X-1}	a_{l_X}
Y	**observation**	y_1	y_2	$\ldots$	y_{n-1}	y_n
	value	b_1	b_2	$\ldots$	b_{l_Y}	b_{l_Y}

Bivariate characteristics

Measure of association

Data Science

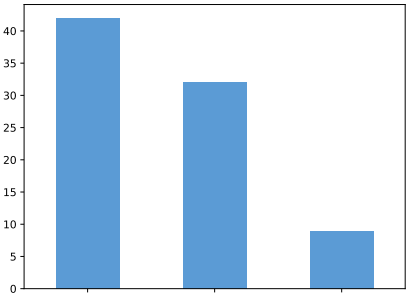
Bivariate characteristics

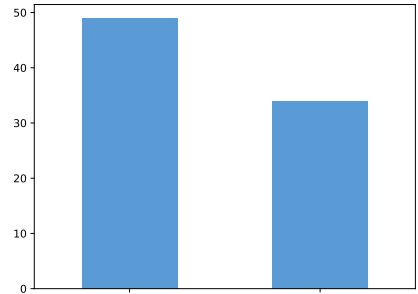
	Name	Population	Area	Location	Growth
1	Berlin	3782202	891.12	North	growing strongly
2	Hamburg	1910160	755.09	North	growing strongly
⋮	⋮	⋮	⋮	⋮	⋮
8	Leipzig	619879	297.80	South	growing strongly
9	Dortmund	595471	280.71	North	growing
⋮	⋮	⋮	⋮	⋮	⋮

Table: List of the largest cities (83 cities with more than 100000 inhabitants) of Germany adapted from[1]. Added columns location (North, South) and Growth (shrinking, growing, growing strongly)

Data Science

Bivariate characteristics





- $h(growing) = 42$, $h(growingstrongly) = 32$, $h(shrinking) = 9$
- $h(North) = 49$, $h(South) = 34$

Task: Can we see a correlation between growth and location?

Data Science

Bivariate characteristics

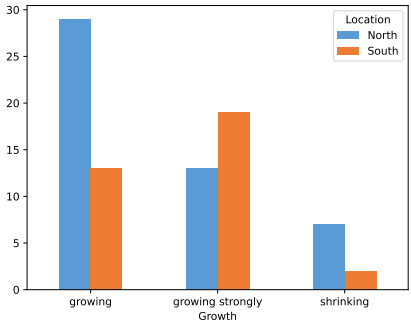
Consider both variables simultaneously, i.e. computing frequencies of every combination of the values.

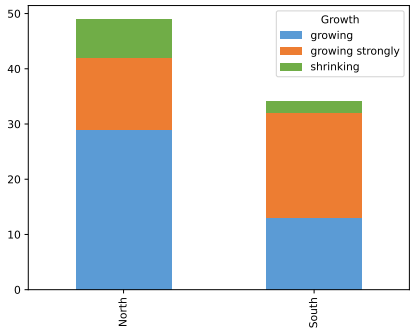
"How many cities are growing and located in the North?"

Location	Growth	Absolute Frequency
North	growing	29
North	growing strongly	13
North	shrinking	7
South	growing	13
South	growing strongly	19
South	shrinking	2

Data Science

Bivariate characteristics





Task: Can we see a correlation between growth and location?

Data Science

Bivariate characteristics

We define

- the absolute combined frequency of value (a_i, b_j) as

$$h_{ij} = h(a_i, b_j) = \sum_{o=1}^{n} (x_o = a_i \wedge y_o = b_j)$$

- the marginal absolute frequency of value a_i as $h_{i\bullet} = h(a_i, \bullet) = \sum_{o=1}^{n} (x_o = a_i)$
- the marginal absolute frequency of value b_j as $h_{\bullet j} = h(\bullet, b_j) = \sum_{o=1}^{n} (y_o = b_j)$

Data Science

Bivariate characteristics

Similarly, we define

- the relative combined frequency of value (a_i, b_j) as $f_{ij} = f(a_i, b_j) = \frac{h(a_i, b_j)}{n}$
- the marginal relative frequency of value a_i as $f_{i\bullet} = f(a_i, \bullet) = \frac{h(a_i, \bullet)}{n}$
- the marginal relative frequency of value b_j as $f_{\bullet j} = f(\bullet, b_j) = \frac{h(\bullet, b_j)}{n}$

Note that the marginal frequency equals to the frequency of the variable itself.

Data Science

Bivariate characteristics

Location	Growth	h_{ij}	f_{ij}
North	growing	29	0.35
North	growing strongly	13	0.16
North	shrinking	7	0.08
South	growing	13	0.16
South	growing strongly	19	0.23
South	shrinking	2	0.02

Location	$h_{i\bullet}$	$f_{i\bullet}$
North	49	0.59
South	43	0.41

Growth	$h_{\bullet j}$	$f_{\bullet j}$
growing	42	0.51
growing strongly	32	0.39
shrinking	9	0.11

Data Science

Bivariate characteristics Contingency table

The two-dimensional frequency distribution of nominal variables is often represented by a **contingency table**. A $l_X - l_Y$**-contingency** table consists of l_X-rows, for each value of the first variable one, and l_Y-columns, for each value of the second variable.

X \ Y	b_1	b_2	$\dots$	b_{l_Y}
a_1	h_{11}	h_{12}	$\dots$	h_{1l_Y}
a_2	h_{21}	h_{22}	$\dots$	h_{2l_Y}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_{l_X}	h_{l_X1}	h_{l_X2}	$\dots$	$h_{l_Xl_Y}$

- $l_X - l_Y$-contingency table with absolute frequencies

X \ Y	b_1	b_2	$\dots$	b_{l_Y}
a_1	f_{11}	f_{12}	$\dots$	f_{1l_Y}
a_2	f_{21}	f_{22}	$\dots$	f_{2l_Y}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_{l_X}	f_{l_X1}	f_{l_X2}	$\dots$	$f_{l_Xl_Y}$

- $l_X - l_Y$-contingency table with relative frequencies

The $2 - 2$ contingency table is also called **fourfold table.**

X \ Y	b_1	b_2
a_1	h_{11}	h_{12}
a_2	h_{21}	h_{22}

- $l_X - l_Y$-contingency table with absolute frequencies

X \ Y	b_1	b_2
a_1	f_{11}	f_{12}
a_2	f_{21}	f_{22}

- $l_X - l_Y$-contingency table with relative frequencies

Data Science

Bivariate characteristics Contingency table

Growth \ Location	North	South	$\sum$
shrinking	7	2	9
growing	29	13	42
growing strongly	13	19	32
$\sum$	49	34	83

Growth \ Location	North	South	$\sum$
shrinking	0.08	0.02	0.11
growing	0.35	0.16	0.51
growing strongly	0.16	0.23	0.39
$\sum$	0.59	0.41	1

Data Science

Bivariate characteristics Contingency table

Growth \ Location	North	South	$\sum$
shrinking	0.08	0.02	0.11
growing	0.35	0.16	0.51
growing strongly	0.16	0.23	0.39
$\sum$	0.59	0.41	1

- 35% of the cities are in the North and growing
- 23% of the cities are in the South and strongly growing

Relative combined frequencies do not provide direct indication of the relationship between values / characteristics.

What is the frequency of shrinking, growing and growing strongly under the condition that we consider a city in the North or South?

North

Growth	absolute	relative
shrinking	7	0.14
growing	29	0.59
growing strongly	13	0.27

South

Growth	absolute	relative
shrinking	2	0.06
growing	14	0.41
growing strongly	19	0.56

- Let $f_{\bullet j} > 0$. Then we define the **conditional frequency** of $X = x_i$ under the condition $Y = y_j$ as

$$f_{X=x_i|Y=y_j} = \frac{f_{ij}}{f_{\bullet j}} \text{ for } i \in \{1, \ldots, l_X\}$$

- Let $f_{i\bullet} > 0$. Then we define the **conditional frequency** of $Y = y_j$ under the condition $X = x_i$ as

$$f_{Y=y_j|X=x_i} = \frac{f_{ij}}{f_{i\bullet}} \text{ for } j \in \{1, \ldots, l_Y\}$$

- Note the notation: $AA|BB$ denotes AA holds under the condition that BB holds - in short: AA holds given BB.
- $F_{Y=y_1|X=x_i}, \ldots, f_{Y=y_{l_Y}|X=x_i}$ is called **conditional frequency distribution** of Y under the condition $X = x_i$

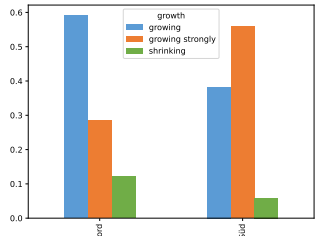
What is the conditional frequency of *growth* given *location*?

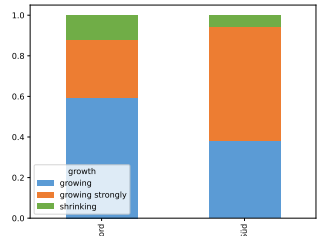
growth \ location	growing	growing strongly	shrinking	$\sum$
North	0.59	0.29	0.12	1
South	0.38	0.56	0.06	1

Task: What associations can we observe?

Data Science

Bivariate characteristics Contingency table





- A large city in the South is more likely to grow strongly than a large city in the North
- A large city in the North is more likely to shrink than a large city in the south
- ...

Conditional frequency gives us a first insight on correlations - how would they look like if there is no correlation?

- The conditional frequency distributions of $X|Y = y_{j_1}$ and $X|Y = y_{j_2}$ for $j_1 \in \{1, \ldots, l_Y\}$ and $j_2 \in \{1, \ldots, l_Y\}$ equal, if for the relative conditional frequencies holds:

$$f_{X|Y=y_{j_1}} = f_{X|Y=y_{j_2}}$$

for all $i = 1, \ldots, l_X$.

The variable X is **empirically independent** of variable Y if all conditional frequency distributions of $X|Y = y_j$ for all $j = 1, \ldots, l_Y$ are equal.

- X is empirically independent of $Y \Leftrightarrow$ Y is empirically independent of X
- X is empirically independent of $Y \Leftrightarrow h_{ij} = \frac{h_{i\bullet} h_{\bullet j}}{n}$

How to describe the relation between two variables? What are the requirements on such a **measure of association**?

- For independent variables the measure should be zero
- For ”fully dependent” variables the measure should be one

The term ”fully dependent” is difficult to be defined. Generally, it should mean that the distribution of variable Y can be completely derived by only knowing the distribution of variable X. Note that this is only possible for a quadratic table.

Idea to construct a measore of association

Compare the given contingency table with one, which fulfills the empirical independence given the same marginal distribution.

The χ^2**-coefficient** is given by

$$\chi^2 = \sum_{i=1}^{l_X} \sum_{j=1}^{l_Y} \frac{\left(h_{ij} - \tilde{h}_{ij}\right)^2}{\tilde{h}_{ij}} \text{ with } \tilde{h}_{ij} = \frac{h_{i\bullet} h_{\bullet j}}{n} \text{ and } \chi^2 \in [0, \infty)$$

- χ^2 is zero if the two variables are empirically independent.
- A large χ^2 indicates a correlation between the two variables - but what is large?

Data Science

Bivariate characteristics Contingency table

Growth \ Location	North	South	$\sum$
shrinking	7	2	9
growing	29	13	42
growing strongly	13	19	32
$\sum$	49	34	83

$$\chi^2 = \frac{(7-5.31)^2}{5.31} + \cdots + \frac{(19-13.11)^2}{13.11}$$
$$= 7.53$$

Is there a correlation between growth and location? At least we can say, that the two variables are not completely independent?

Measures of association

- The **Pearson contingency coefficient** K_P is given by

$$K_p = \sqrt{\frac{\chi^2}{\chi^2 + n}} \text{ with } K_P \in \left[0, \sqrt{\frac{\min(l_X - 1, l_Y - 1)}{\min(l_X, l_Y)}}\right]$$

- The **corrected Pearson contingency coefficient** K_P^* is given by

$$K_P^* = \frac{K_P}{\max K_P} = \sqrt{\frac{\chi^2}{\chi^2 + n}} \cdot \sqrt{\frac{\min(l_X, l_Y)}{\min(l_X - 1, l_Y - 1)}} \text{ with } K_P^* \in [0, 1]$$

Data Science

Bivariate characteristics Contingency table

Growth \ Location	North	South	$\sum$
shrinking	7	2	9
growing	29	13	42
growing strongly	13	19	32
$\sum$	49	34	83

$$\chi^2 = \frac{(7-5.31)^2}{5.31} + \cdots + \frac{(19-13.11)^2}{13.11} = 7.53$$

$$K_P^* = \sqrt{\frac{\chi^2}{\chi^2+n}} \cdot \sqrt{\frac{\min(l_X, l_Y)}{\min(l_X-1, l_Y-1)}} = \sqrt{\frac{7.53}{7.53+83}} \cdot \sqrt{2} = 0.41$$

There seems to be a weak or medium association between location and growth.

The value $0 \leq K_P^* \leq 1$ can be interpreted in different ways. In the following one "general" interpretation of the value is given.

value	**interpretation**
$K_P^* = 0$	empirically independent
$K_P^* \in (0, 0.3]$	weak correlation
$K_P^* \in (0.3, 0.7]$	medium correlation
$K_P^* \in (0.7, 1]$	strong correlation

Bivariate characteristics

Correlation Coefficient

Data Science

Bivariate characteristics

Given two metric variables X and Y - is there a measure for (linear) correlation?

Linear correlation?

Does there hold $y_i = \alpha x_i + \beta$, with $\alpha, \beta \in \mathbb{R}$ and $\alpha \neq 0$? - in statistics this equation will not hold for one choice of α, β. Thus, we need a measure to identify how strong the correlation is!

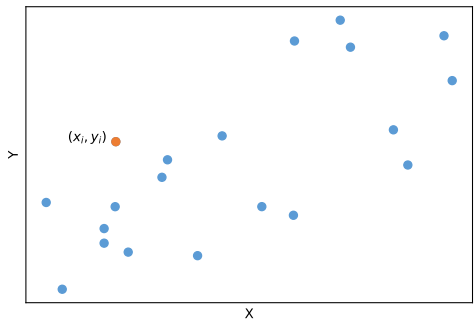
Of interest?

- **Form** of the correlation (e.g. linear, quadratic, ...)
- **Direction** of the correlation (positive, negative)
- **strength** of the correlation (strong, medium, weak)

Data Science

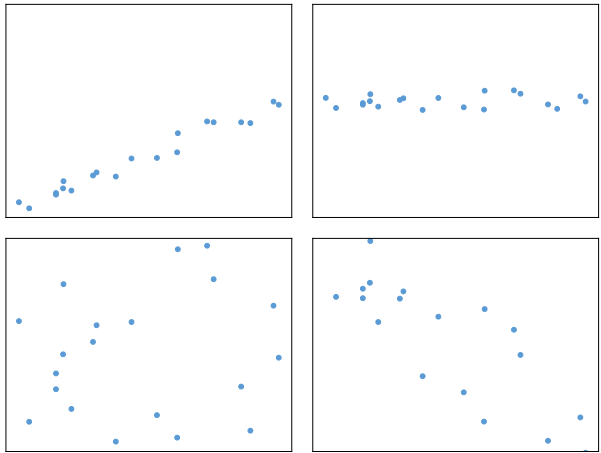
Bivariate characteristics

In a **scatter plot**, every variable of (X, Y) is associated with one axis of Cartesian coordinates. Every observation (x_i, y_i) for $i = 1, \ldots, n$ is marked, e.g. with a cross or a point.



Data Science

Bivariate characteristics

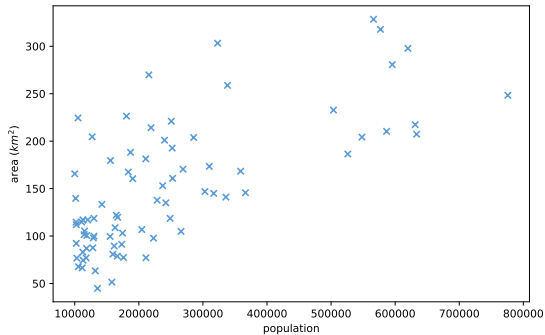


Task: Is there a correlation or not?

Data Science

Bivariate characteristics

Name	Population	Area
Frankfurt a. M.	773068	248.31
Stuttgart	632865	207.33
⋮	⋮	⋮
Dortmund	595471	280.71
⋮	⋮	⋮

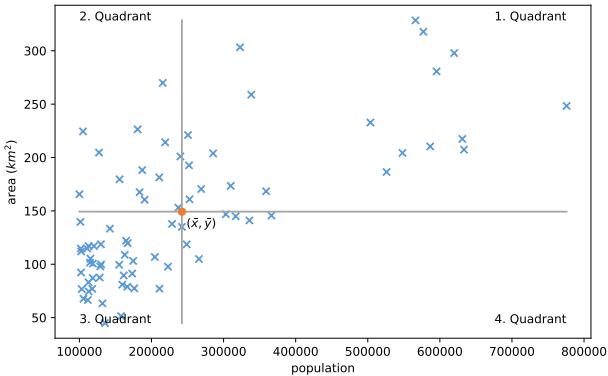


Task: Is there are correlation between population and area? Of which type is this correlation?

Data Science

Bivariate characteristics

Compare values with the arithmetic mean - separate the scatter plot into 4 quadrants.



1. Quadrant	$x_i > \bar{x}$	$y_i > \bar{y}$
2. Quadrant	$x_i < \bar{x}$	$y_i > \bar{y}$
3. Quadrant	$x_i < \bar{x}$	$y_i < \bar{y}$
4. Quadrant	$x_i > \bar{x}$	$y_i < \bar{y}$

Data Science

Bivariate characteristics

Values in the first and third quadrant indicate positive correlation.

- ”large” x-values $\leftrightarrow$ ”large” y-values
- ”small” x-values $\leftrightarrow$ ”small” y-values

Values in the second and fourth quadrant indicate negative correlation.

- ”large” x-values $\leftrightarrow$ ”small” y-values
- ”small” x-values $\leftrightarrow$ ”large” y-values

The **empirical covariance** of two variables X and Y is defined by

$$\tilde{s}_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

where $\bar{x}$ and $\bar{y}$ are the corresponding arithmetic means.

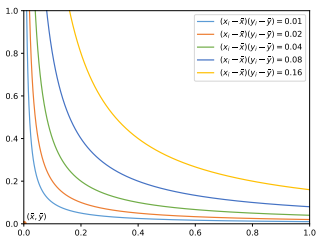
- The empirical variance of a variable X is a special case of the empirical covariance, i.e. $\tilde{s}_X = \tilde{s}_{XX}$

Data Science

Bivariate characteristics

The term $(x_i - \bar{x})(y_i - \bar{y})$ indicates in which quadrants the values (x_i, y_i) lie:

- $(x_i - \bar{x})(y_i - \bar{y}) > 0$: The values lie in the first or third quadrant
- $(x_i - \bar{x})(y_i - \bar{y}) < 0$: The values lie in the second or fourth quadrant

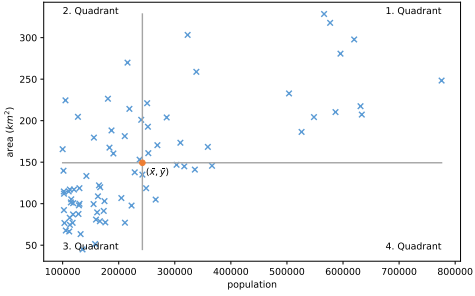


- The absolute value of the term $(x_i - \bar{x})(y_i - \bar{y})$ indicates how ”deep” the value lie in the corresponding quadrant.
- The sum then indicates in which quadrants the values are most likely

Data Science

Bivariate characteristics

name	population	area	$y_i - \bar{y}$	$x_i - \bar{x}$	$(x_i - \bar{x})(y_i - \bar{y})$
Frankfurt am Main	775790	248.31	99.02	533604.73	52840107.48
⋮	⋮	⋮	⋮	⋮	⋮
Dortmund	595471	280.71	131.42	353285.73	46430510.53
⋮	⋮	⋮	⋮	⋮	⋮



$$\tilde{s}_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 594781797.95$$

Positive direction, but is 594781797.95 a strong, weak or medium correlation?

Data Science

Bivariate characteristics

The **Pearson correlation** coefficient r_{XY}, also **empirical correlation**, for two variables X and Y is given by

$$r_{XY} = \frac{\tilde{s}_{XY}}{\tilde{s}_X \tilde{s}_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $\tilde{s}_X$ and $\tilde{s}_Y$ are the empirical standard deviations of the variable X and Y.

- Correlation coefficient is a dimensionless number
- Correlation coefficient only measures a linear correlation between the two variables

Data Science

Bivariate characteristics

The value $-1 \leq r_{XY} \leq 1$ gives a glimpse on a possible linear correlation of two variables.

value	description
$r_{XY} = 1$	All values are lying exactly on one row with positive gradient
$r_{XY} = -1$	All values are lying exactly on one row with negative gradient
$r_{XY} = 0$	X and Y are not linear correlated
$r_{XY} > 0$	X and Y are positive linear correlated
$r_{XY} < 0$	X and Y are negative linear correlated

Data Science

Bivariate characteristics

The value r_{XY} can be interpreted in different ways. In the following one "possible" interpretation of the value is given.

value	interpretation
$r_{XY} = 0$	empirically independent
$r_{XY} \in (-0.3, 0) \cup (0, 0.3)$	weak correlation
$r_{XY} \in (-0.5, 0.3] \cup [0.3, 0.5)$	medium correlation
$r_{XY} \in [-1, -0.5] \cup [0.5, 1]$	strong correlation

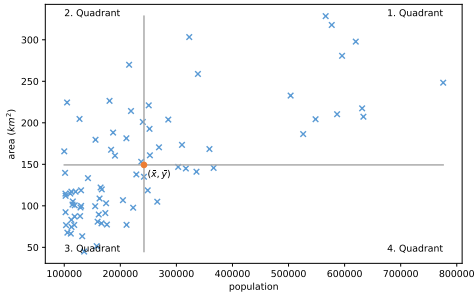
Data Science

Bivariate characteristics

Let’s take a look

Data Science

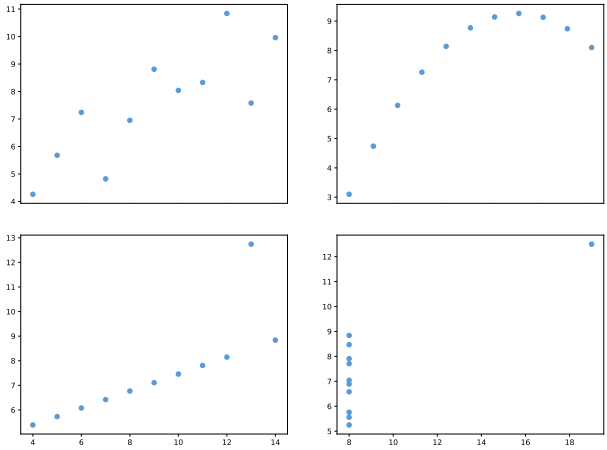
Bivariate characteristics



The resulting correlation coefficient: $r_{XY} = 0.70$, thus we observe a **strong correlation**.

Data Science

Bivariate characteristics



The **Anscombe's quartet** describe four datasets, where all have (nearly) the same descriptive statistics, but are completely different.

Take also the scatter plot into account to verify a possible correlation in the data!

Bivariate characteristics

Ordinal data

Data Science

Bivariate characteristics

What about two ordinal variables?

- Ordinal variables are ”between” nominal and metric!

Ordinal variables can always be seen as nominal variables by ignoring the order. Thus, computing **measures of association** is possible and useful.

- For example if the variables have few values which have a large frequency
- The variable **Growth** in our example is ordinal since there is a natural order:

$$shrinking < growing < growingstrongly$$

Data Science

Bivariate characteristics

Alternative: Mapping ordinal variables to metric variables to compute a correlation coefficient.

The **rank correlation** can be computed to measure an association between two ordinal variables. This is done by ”ranking” the values, corresponding to their order. Then a correlation coefficient for a metric variable can be used.

- Consider the correlation coefficient r_{XY} and replace the observations x_i and y_i with the corresponding ranks $R(x_i)$ and $R(y_i)$, for $i = 1, \ldots, n$.
  - For example for test-scores and marks

Summary & Outlook

Data Science

Summary & Outlook: Summary

- You can compute different statistical deviations
- You can analyze bivariate characteristics concerning their correlation
- You are able to compute different correlation coefficients and understand their meaning

Data Science

Summary & Outlook: Outlook

Start: Probability theory

References

Data Science

Summary & Outlook: Endnotes

[1] https://de.wikipedia.org/wiki/Liste_der_größten_deutschen_Städte

Data Science

Summary & Outlook: Acknowledgement

Parts of the lecture base on the lecture ”Statistics” (FH Dortmund)
by
Prof. Dr. Sonja Kuhnt and Prof. Dr. Nadja Bauer.