UNIVERSITÄT
LEIPZIG

10-207-0003: Introduction to Stochastics

Multivariate Problems, Correlation

14.05.2025, Leipzig

Dr. Ing. Andreas Niekler

Computational
Humanities
UNIVERSITÄT LEIPZIG

UNIVERSITÄT
LEIPZIG

SYLLABUS

1. Empirical research and scale levels
2. Univariate description and exploration of data
3. Graphical representation of characteristics / Explorative data analysis
4. Measures of data distribution
5. ***Multivariate Problems, Correlation***
6. Regression
7. Probability distributions
8. Central Limit Theorem
9. Confidences
10. Statistical testing
11. Linear Regression
12. Correlation and covariance
13. Logistic regression
14. Bayes theorem

Additional: Entropy, Mutual Information, Maximum Likelihood Estimator, Mathy Stuff

MULTIVARIATE DATA

- We have defined graphical and numerical description of a single characteristic.
- The analysis of one-dimensional characteristics is usually only a first (but important) step in describing the available data.
- Especially in the humanities, we are interested in analyzing relationships between several characteristics.
- Examples of typical research questions:
  - Does gender influence earned income?
  - Is there a relationship between social class affiliation and inclination towards education?
  - Is there a relationship between cultural popularity and media presence?
  - Is readability influenced by text properties?
  - Etc.

MULTIVARIATE DATA

- To answer such questions and similar ones, several characteristics are collected jointly for each statistical unit of a population $G = \{u_1, \dots, u_n\}$.

- Remark: If one has, for example, the characteristics X, Y, Z, one can also consider the pair (X, Y) or the triple (X, Y, Z) as a whole, such that one then speaks of a two-dimensional (bivariate) or three-dimensional (trivariate) characteristic.

- Subsequently, we will focus primarily on the analysis and description of two characteristics, X and Y, i.e., we consider a raw list with data points $(x_1, y_1), \dots, (x_n, y_n)$.

MULTIVARIATE DATA

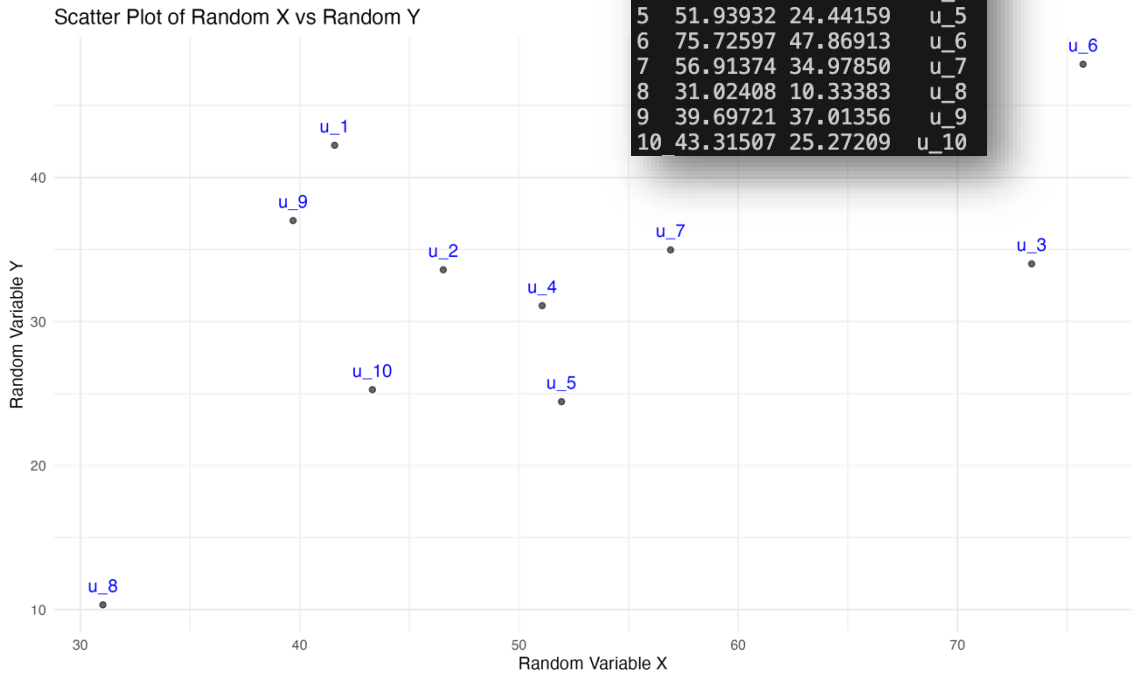
- To describe the relationship between quantitative variables, we will learn about:
  - Scatter plots for the graphical investigation of a relationship between two quantitative variables
  - The correlation coefficient as a measure that quantifies the strength (and also direction) of the linear relationship between two quantitative characteristics.
  - OLS regression for determining a so-called regression line, which describes or represents a (possible) linear relationship between two quantitative variables both graphically and numerically

SCATTER PLOT

- A scatter plot describes the relationship between two quantitative variables X and Y in the form of a Cartesian point diagram:
  - The values of X appear on the horizontal axis, while the values of Y appear on the vertical axis.
  - Each statistical unit u_i is represented by a point in the diagram, whose position is determined by the corresponding raw values x_i and y_i.

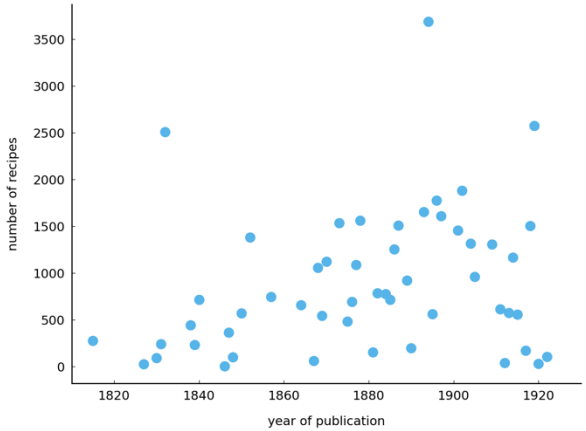
SCATTER PLOT

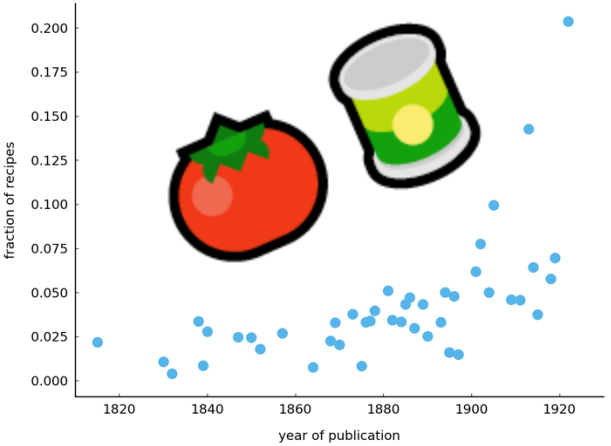
	x	y	label
1	41.59287	42.24082	u_1
2	46.54734	33.59814	u_2
3	73.38062	34.00771	u_3
4	51.05763	31.10683	u_4
5	51.93932	24.44159	u_5
6	75.72597	47.86913	u_6
7	56.91374	34.97850	u_7
8	31.02408	10.33383	u_8
9	39.69721	37.01356	u_9
10	43.31507	25.27209	u_10



EXAMPLES FOR SCATTERPLOTS

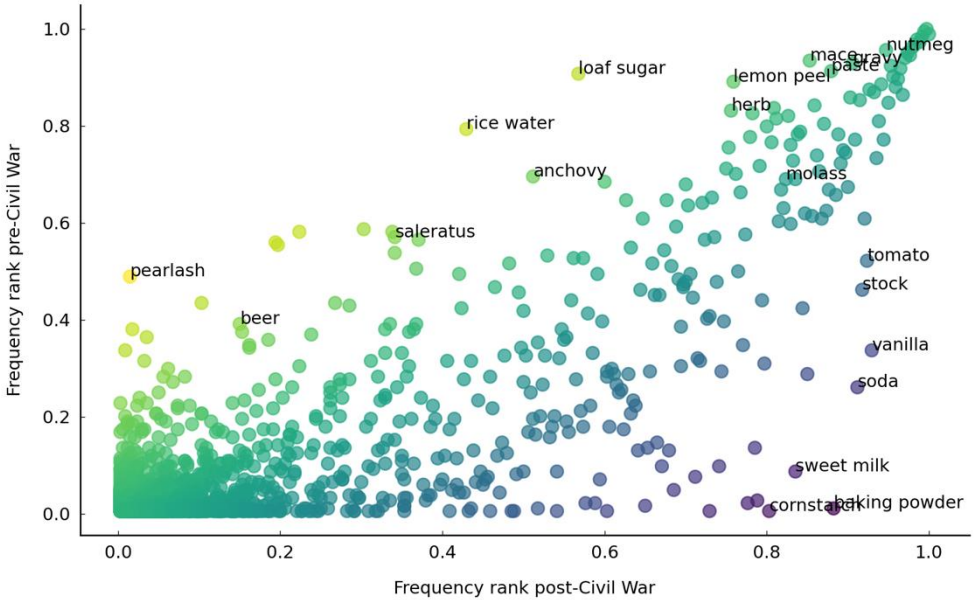
Feeding America: The Historic American Cookbook dataset: Karsdorp, F., Kestemont, M., & Riddell, A. (2021). Humanities Data Analysis: Case Studies with Python. Princeton University Press.





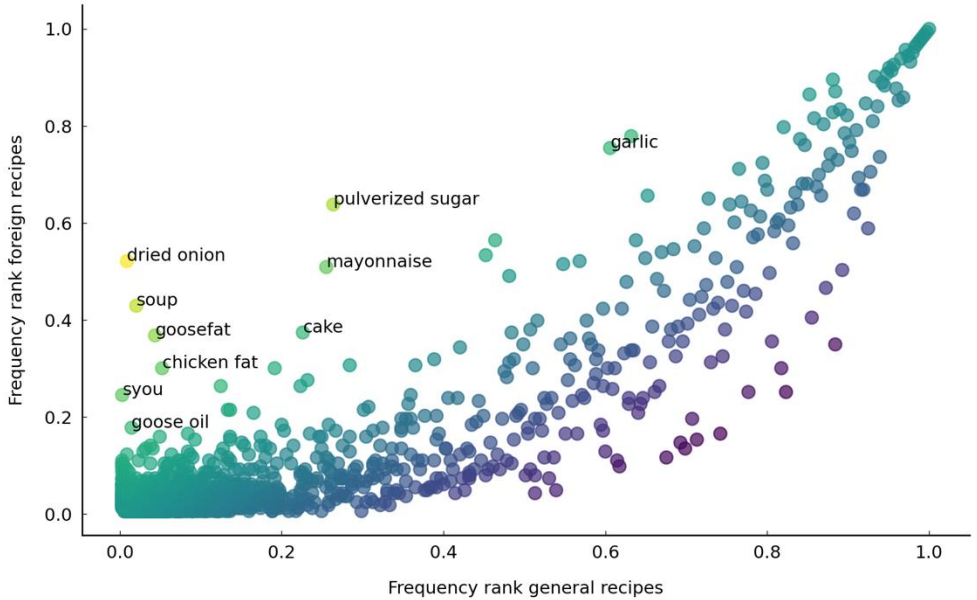
EXAMPLES FOR SCATTERPLOTS

Feeding America: The Historic American Cookbook dataset: Karsdorp, F., Kestemont, M., & Riddell, A. (2021). Humanities Data Analysis: Case Studies with Python. Princeton University Press.



EXAMPLES FOR SCATTERPLOTS

Feeding America: The Historic American Cookbook dataset: Karsdorp, F., Kestemont, M., & Riddell, A. (2021). Humanities Data Analysis: Case Studies with Python. Princeton University Press.



GROUPING OF SIMILAR VALUES

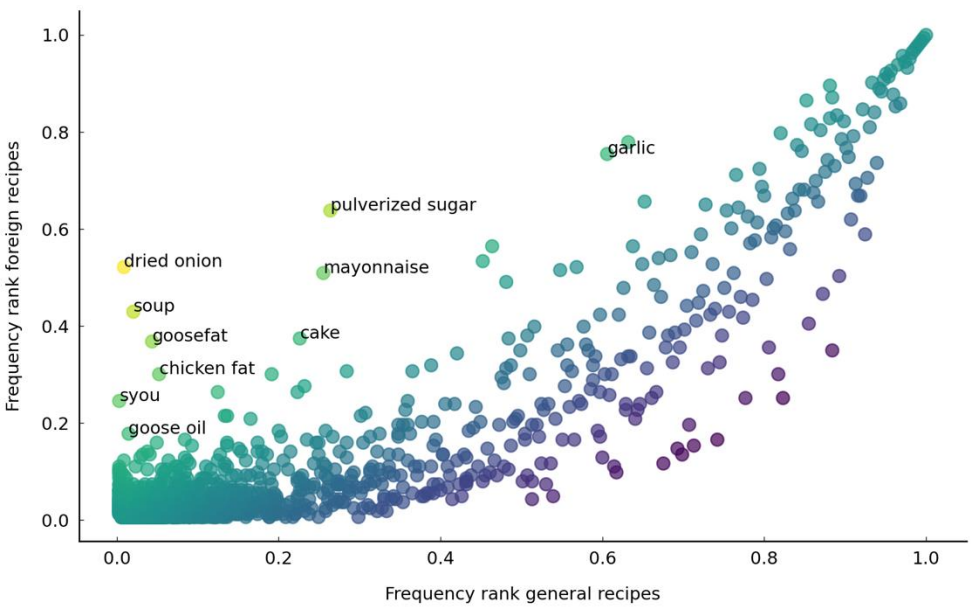
Feeding America: The Historic American Cookbook dataset: Karsdorp, F., Kestemont, M., & Riddell, A. (2021). Humanities Data Analysis: Case Studies with Python. Princeton University Press.

DISPLAYING ADDITIONAL FEATURES IN SCATTER PLOTS

- Scatter plots primarily serve to **display two metric features** but can include additional features through suitable representation.
- **Additional Categorical Feature in the Scatter Plot**: To include an additional categorical feature in a scatter plot, one can use, for example, different color codings or symbols.
- **Additional Metric Feature in the Scatter Plot:** To display an additional metric feature in a scatter plot, one can vary the size of the points or utilize the time dimension.
- Remark: Although it is in principle possible with the help of software to create so-called 3D plots, allowing three metric features to be displayed simultaneously, the **use of such graphics is generally discouraged**.

USING COLORS FOR ADDITIONAL FEATURES

Feeding America: The Historic American Cookbook dataset: Karsdorp, F., Kestemont, M., & Riddell, A. (2021). Humanities Data Analysis: Case Studies with Python. Princeton University Press.



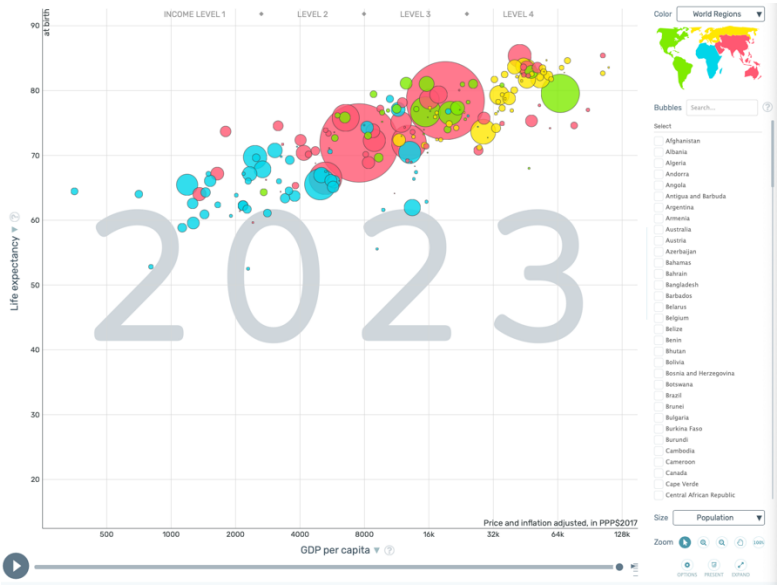
More common in general receipes

More common in foreign receipes

GAP MINDER

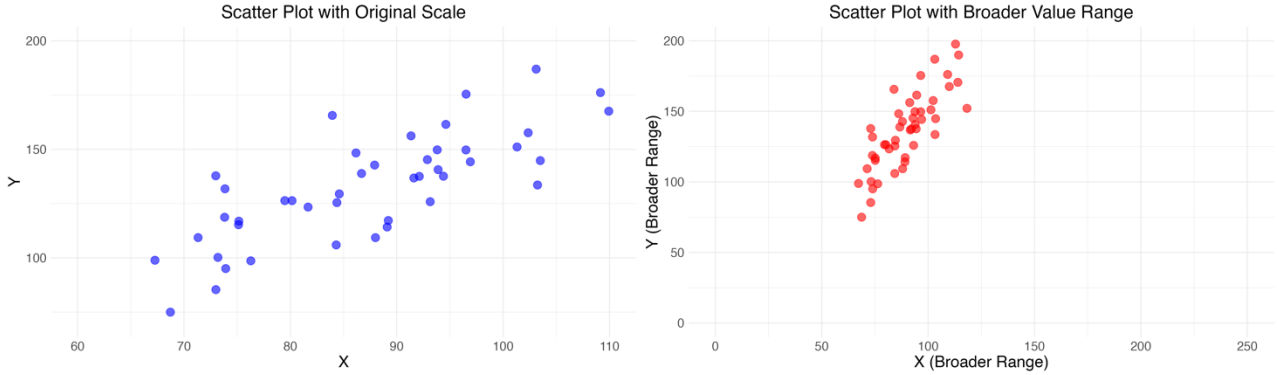
Source: Free material from www.gapminder.org

- Gapminder developed a very successful tool for visualizing data on the state of our world's development.
- Also note the very entertaining TED talk by Hans Rosling and the book "Factfulness" by Hans Rosling, Anna Rosling Rönnlund, and Ola Rosling is also highly recommended reading.
  - Also highly recommended: Our World in Data by Max Roser et al.



DEPENDENCY

- The **presence and direction of a relationship** between two quantitative features can usually be recognized quite well from a scatter plot.
- However, when it comes to **assessing the strength of the relationship**, scatter plots can sometimes be a bit misleading.
- The same data in two differently scaled scatter plots:



CORRELATION COEFFICIENT (BRAVAIS AND PEARSON)

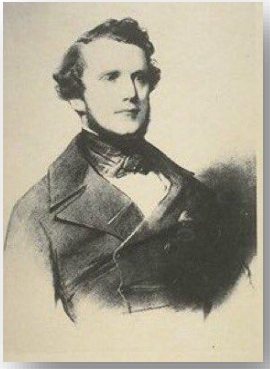
👉 The (linear) correlation or the so-called correlation coefficient r measures the strength and direction of the linear relationship between two variables X and Y and is defined as:

$$r = r_{XY} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_X}\right)\left(\frac{y_i - \bar{y}}{s_Y}\right),$$

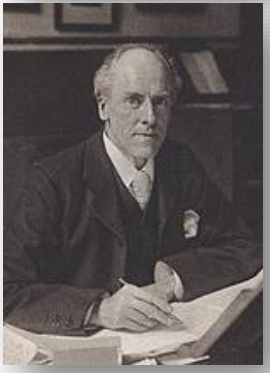
where $\bar{x}$ and $\bar{y}$ are the means, and s_X and s_Y are the standard deviations of the two variables X and Y.

- One speaks of a positive (linear) correlation if $r > 0$.
- Analogously, one speaks of a negative (linear) correlation if $r < 0$.
- For $r \approx 0$, one says that X and Y are uncorrelated.

CORRELATION COEFFICIENT (BRAVAIS AND PEARSON)



- **Auguste Bravais:** A French physicist and statistician. In the mid-19th century (around 1846), Bravais did significant early mathematical work related to the concept of correlation, particularly in the context of bivariate distributions. He developed ideas that contributed to the measure of linear association between two variables, essentially laying some of the mathematical groundwork for the product-moment correlation.



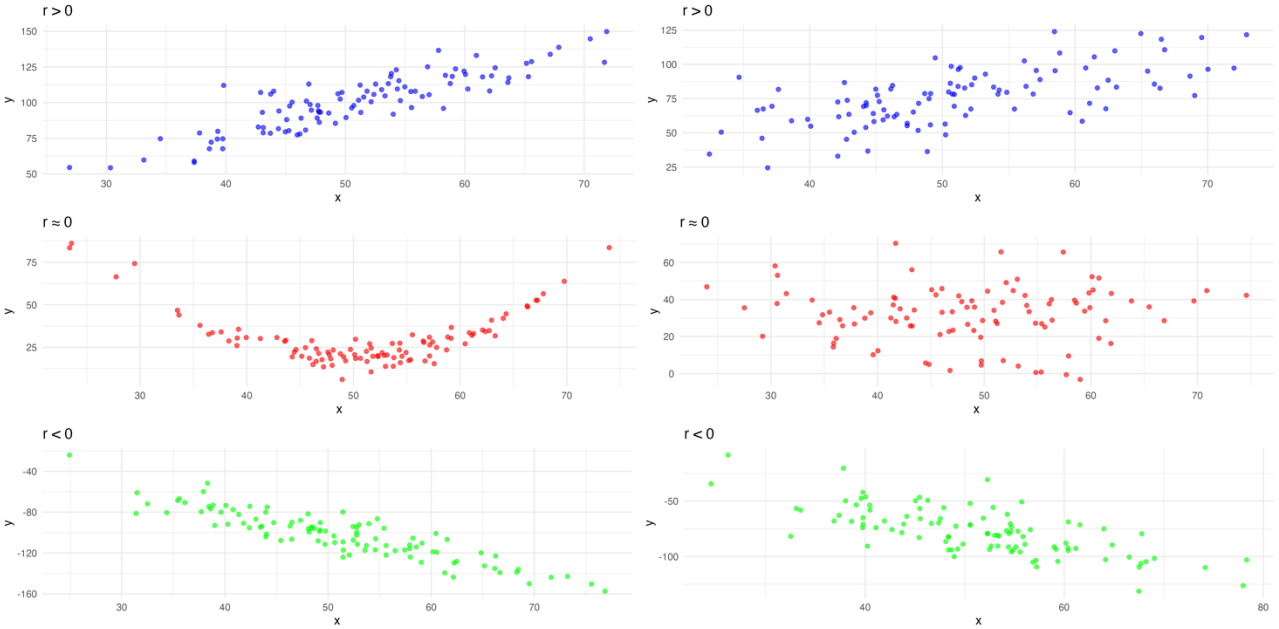
- **Karl Pearson:** A highly influential English mathematician and biostatistician. Building on the earlier work of Francis Galton and Auguste Bravais, Pearson rigorously formalized and popularized the product-moment correlation coefficient around the turn of the 20th century. He provided the mathematical definition and formula that is most widely used today and extensively applied it in various scientific fields..

CORRELATION COEFFICIENT (BRAVAIS AND PEARSON)

$$r = r_{XY} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_X}\right)\left(\frac{y_i - \bar{y}}{s_Y}\right)$$

- The formula looks complicated, but:
- $\left(\frac{x_i-\bar{x}}{s_X}\right)$ and $\left(\frac{y_i-\bar{y}}{s_Y}\right)$ are the standardized values (z-scores) of the unit i for both features X and Y.
- $\left(\frac{x_i-\bar{x}}{s_X}\right)\left(\frac{y_i-\bar{y}}{s_Y}\right)\begin{cases} > 0 \; if \; x_i > \bar{x} \; and \; y_i > \bar{y} \; or \; x_i < \bar{x} \; and \; y_i < \bar{y} \\ = 0 \; if \; x_i = \bar{x} \; pr \; y_i = \bar{y} \\ < 0 \; x_i < \bar{x} \; and \; y_i > \bar{y} \; or \; x_i > \bar{x} \; and \; y_i < \bar{y} \end{cases}$
- All these products are summed up, so that an overall positive relationship leads to a positive r, and conversely, an overall negative relationship leads to a negative r.

CORRELATION COEFFICIENT (BRAVAIS AND PEARSON)



CORRELATION COEFFICIENT (EFFICIANT COMPUTATION)

- With the help of the z-values $z_{X,i}$ and $z_{Y,i}$ we can understand the definition of the coefficient more intuitively:

$$r = \frac{1}{n}\sum_{i=1}^{n} z_{X,i} \cdot z_{Y,i}$$

- However, it can also be shown:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}$$

- Attention: Both formulas are thus valid and always yield the same result, with one formula providing an intuitive definition and the other allowing for efficient calculation!

PROPERTIES CORRELATION COEFFICIENT

- The correlation coefficient is symmetric with respect to X and Y, i.e., $r = r_{XY} = r_{YX}$.

- The correlation coefficient is invariant under positive linear transformations, i.e., if $\tilde{X} = a + bX$ with $b > 0$, then $r = r_{XY} = r_{\tilde{X}Y}$.

- The correlation is always between -1 and 1, i.e., $r \in [-1,1]$.

PROPERTIES CORRELATION COEFFICIENT

- With a correlation of -1, there is a perfect negative linear relationship, i.e., $r = -1 \leftrightarrow Y = a + bX$ with $b < 0$.

- With a correlation of 1, there is a perfect positive linear relationship, i.e., $r = 1 \leftrightarrow Y = a + bX$ with $b > 0$.

- **Remark:** Usually, there is no perfect linear relationship, but in the next lecture, we will examine more closely how well a relationship between X and Y can still be approximately described by $Y = a + bX$ or $X = \tilde{a} + \tilde{b}Y$ $\rightarrow$ ***Simple Linear Regression***.

PROPERTIES CORRELATION COEFFICIENT

- For further assessment of the strength of a correlation, we could follow Cohen's recommendation for the social sciences:

$\|r\|$	0.1	0.3	0.5
Interpretation	Small	Medium	strong

- However, these schematic values are to be understood more as an intersubjective convention of a scientific discipline.
  - For example, in physics, a correlation of 0.9 in the context of a high-precision measurement for a linear relationship would be rather poor, whereas in the social sciences, a correlation of 0.3 can already be quite acceptable.
- Occasionally, the following assessment is also used:
  - $|r| \leq 0.5$: weak correlation.
  - $0.5 < |r| \leq 0.8$: medium correlation.
  - $|r| > 0.8$: strong correlation.

PROPERTIES CORRELATION COEFFICIENT

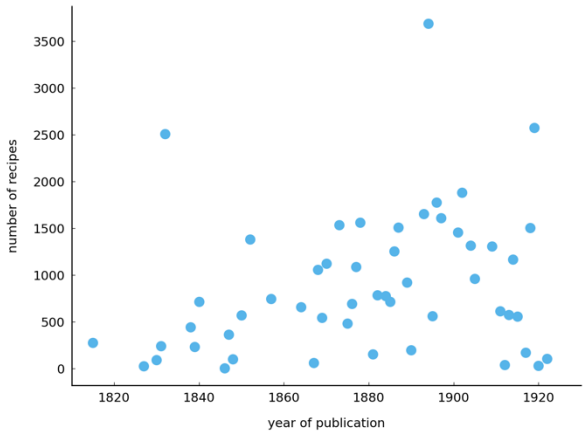
- For further assessment of the strength of a correlation, we could follow Cohen's recommendation for the social sciences:

$\|r\|$	0.1	0.3	0.5
Interpretation	Small	Medium	strong

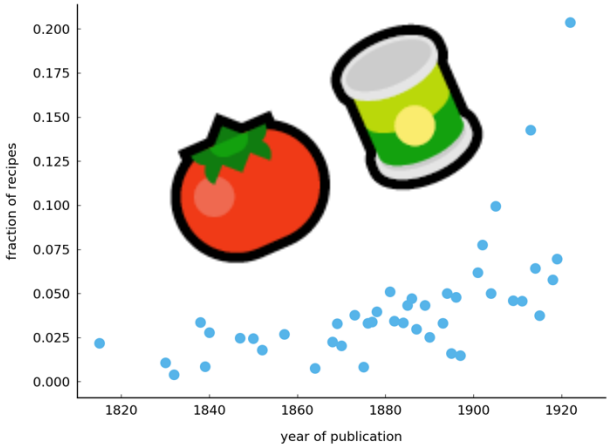
- However, these schematic values are to be understood more as an intersubjective convention of a scientific discipline.
  - For example, in physics, a correlation of 0.9 in the context of a high-precision measurement for a linear relationship would be rather poor, whereas in the social sciences, a correlation of 0.3 can already be quite acceptable.
- Occasionally, the following assessment is also used:
  - $|r| \leq 0.5$: weak correlation.
  - $0.5 < |r| \leq 0.8$: medium correlation.
  - $|r| > 0.8$: strong correlation.

EXAMPLES FOR SCATTERPLOTS

Feeding America: The Historic American Cookbook dataset: Karsdorp, F., Kestemont, M., & Riddell, A. (2021). Humanities Data Analysis: Case Studies with Python. Princeton University Press.



0.27



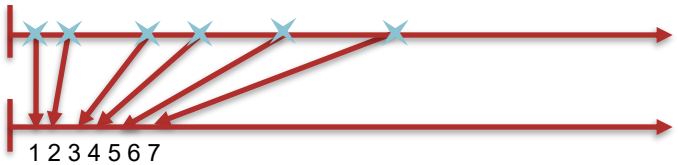
0.48

ADDITIONAL CORRELATION COEFFICIENTS

- The correlation coefficient can thus only **be meaningfully applied to two quantitative variables at the metric scale level** - with three significant exceptions.
- If Pearson's correlation coefficient is calculated for two dichotomous, 0-1 coded features or alternative features X and Y, it is also referred to as the so-called **Point Correlation Coefficient** → Φ coefficient.
- If Pearson's correlation coefficient is applied to one alternative feature X and one quantitative feature Y, it is also referred to as the so-called **Point-biserial correlation**.
- Pearson's correlation coefficient can also be applied to two **ordinal scale features** with the help of a clever transformation of the data → ***Spearman's rank correlation coefficient.***

ADDITIONAL CORRELATION COEFFICIENTS

- We consider a bivariate feature (X,Y), where X and Y are only on an ordinal scale but have many different values.
- On an ordinal scale, distances are not meaningfully interpretable, i.e $(\bar{x}, \bar{y})$would be arbitrary numbers, just like the values $(x_i - \bar{x}), (y_i - \bar{y})$ themselves, which is why the correlation coefficient (Bravais/Pearson) cannot be applied in this way.
- However, the raw values can be rank-transformed:



- When rank-transforming the raw values, one must distinguish two cases: Do so-called ties exist or not?
  - Ties mean whether certain raw values x_i and y_i appear multiple times in the respective original list.
  - No ties: 1, 54, 7, 4
  - One tie: **5**, 3, **5**, 2, 6, 7, 4

ADDITIONAL CORRELATION COEFFICIENTS

- When there are no ties, instead of calculating with the original list $(x_i, y_i)_{i=1,\dots,n}$, one calculates with the rank-transformed original list $\big(rg(x_i), rg(y_i)\big)_{i=1,\dots,n}$, where $rg(x_i) = j :\Leftrightarrow x_i = x(j)$.
- The rank $rg(x_i)$ is thus the order number that x_i occupies in the ordered original list $x(1) < x(2) < \cdots < x(n)$ (analogously for $rg(y_i)$).
- The smallest observation is thus assigned the value 1, the second smallest the value 2, and so on, the largest the value n.
- Example:

x_i	1	7	2	5.3	16
$Rg(x_i)$	1	4	2	3	5

ADDITIONAL CORRELATION COEFFICIENTS

- When ties are present, i.e., if several units have the same value for variable X or variable Y, the average value of the ranks in question is used.
- Example:

x_i	1	7	7	3	10
Possible Rank	1	3 or 4	3 or 4	2	5
$Rg(x_i)$	1	3.5	3.5	2	5

RANK CORRELATION (SPEARMAN)

- The ranks of a variable X defined in this way represent a transformation $X \rightarrow rg(X)$ in the context of a population, based on the empirical values.

👉 **Spearman's Rank Correlation Coefficient:** The so-called Spearman's rank correlation coefficient is defined as

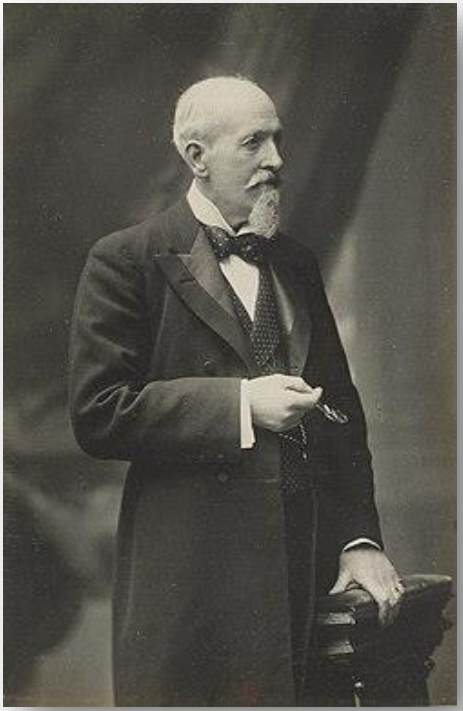
$$r_s = r_{rg(X)rg(Y)},$$

i.e., as the Bravais-Pearson correlation for the two rank-transformed features $rg(X)$ and $rg(Y)$.

- The value of the rank correlation coefficient can thus be interpreted in complete analogy to the Pearson correlation coefficient.

CHARLES EDWARD SPEARMAN

- **Charles Spearman** (1863–1945) was a pioneering English psychologist and statistician. He is best known for his work in psychometrics and for developing Spearman's rank correlation coefficient (r_s or ρ), a non-parametric measure used to assess the monotonic relationship between two ranked variables. Spearman's work laid foundational stones in both the fields of statistics, particularly in dealing with ranked data, and the study of human intelligence.



RANK CORRELATION (SPEARMAN)

- In general, it can be shown that

$$r_s = \frac{\sum_{i=1}^{n} rg(x_i) * rg(y_i) - n\left(\frac{n+1}{2}\right)^2}{\sqrt{\sum_{i=1}^{n} rg(x_i) - n\left(\frac{n+1}{2}\right)^2}\sqrt{rg(y_i) - n\left(\frac{n+1}{2}\right)^2}}$$

- If there are no ties present the formula gets easy:

$$r_s = 1 - 6 \cdot \frac{\sum_{i=1}^{n}\left(rg(x_i) - rg(y_i)\right)^2}{n(n^2-1)}$$

EXAMPLE

- Two linguists, Dr. Miller and Dr. Smith, assess the complexity of 10 different texts on a scale from 1 (very simple) to 100 (very complex). This example includes ties in their ratings.
- Calculate the Spearman rank correlation coefficient for features X and Y with
  - X Rating by Dr. Miller
  - Y Rating by Dr. Smith
- Ratings by the two assessors:

Person i	1	2	3	4	5	6	7	8	9	10
x_i	25	40	60	15	75	80	40	50	90	25
y_i	30	35	55	20	70	85	40	35	95	30
$rg(x_i)$	2.5	4.5	7	1	8	9	4.5	6	10	2.5
$rg(y_i)$	2.5	4.5	7	1	8	9	6	4.5	10	2.5

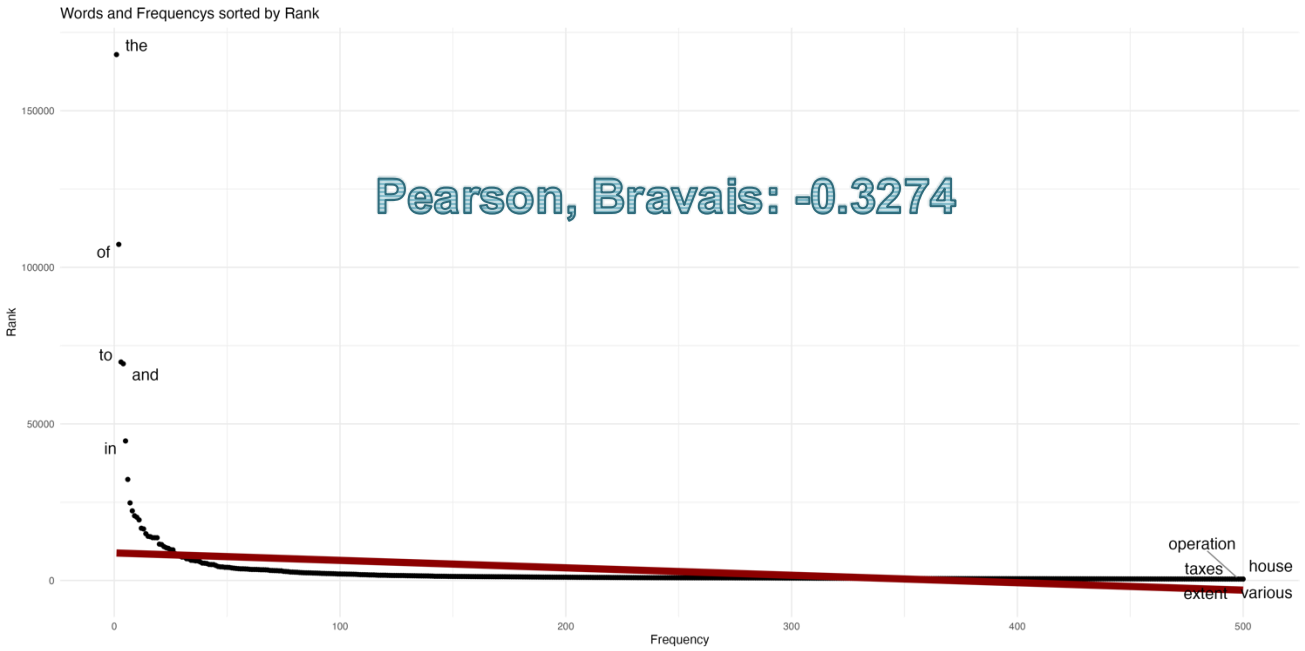
EXAMPLE

- Two linguists, Dr. Miller and Dr. Smith, assess the complexity of 10 different texts on a scale from 1 (very simple) to 100 (very complex). This example includes ties in their ratings:
- Spearman's Rank Correlation Coefficient: 0.97273
- **Interpretation:** While the absolute ratings of the individual assessors are subjectively influenced, they nevertheless show high intersubjective consistency (inter-rater reliability). It is predominantly true that the higher the rating of one assessor, the higher the rating of the other, so that at least the measurement of the order (but not the absolute classification) appears reliable.

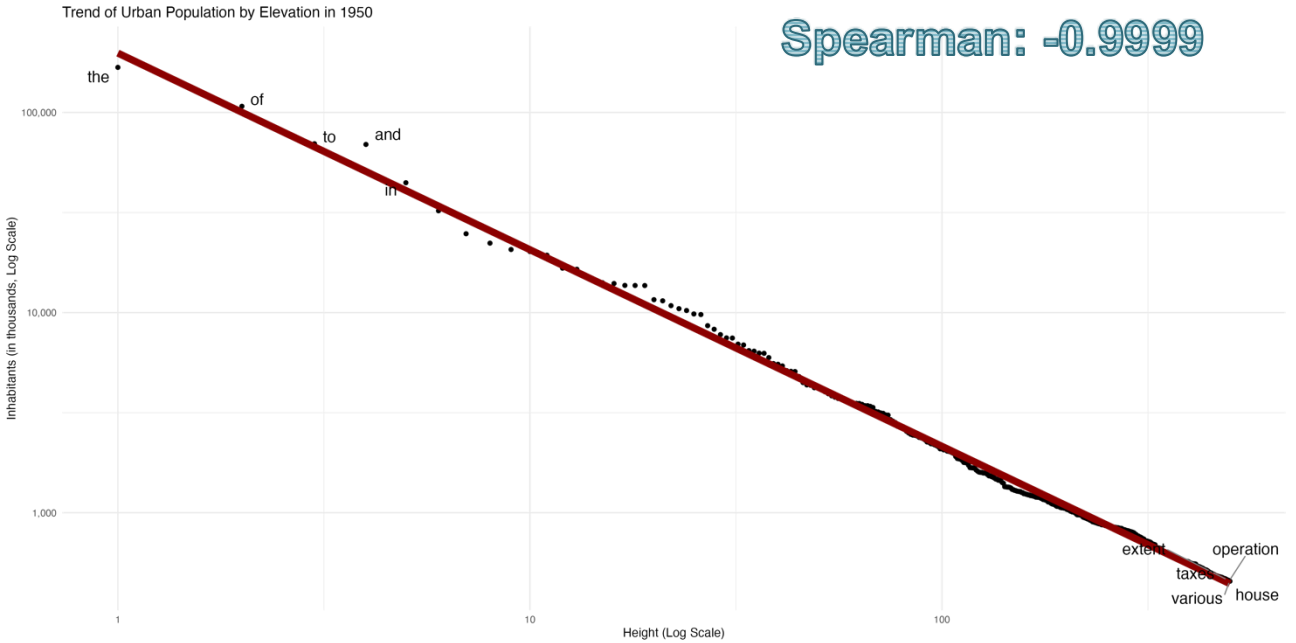
ADDITIONAL USE OF SPEARMAN CORRELATION

- One can now also meaningfully use the rank correlation coefficient for metric data.
- In contrast to the correlation coefficient according to Bravais and Pearson, Spearman's version then measures the strength not of a linear relationship, but more generally of a possible monotonic relationship.
  - A strong linear relationship of the form $Y = a + bX$ is a special case of a monotonic relationship.
  - For example, a relationship of the form $Y = a + bX^3$ would no longer be linear with respect to X and Y, but it would be monotonic.

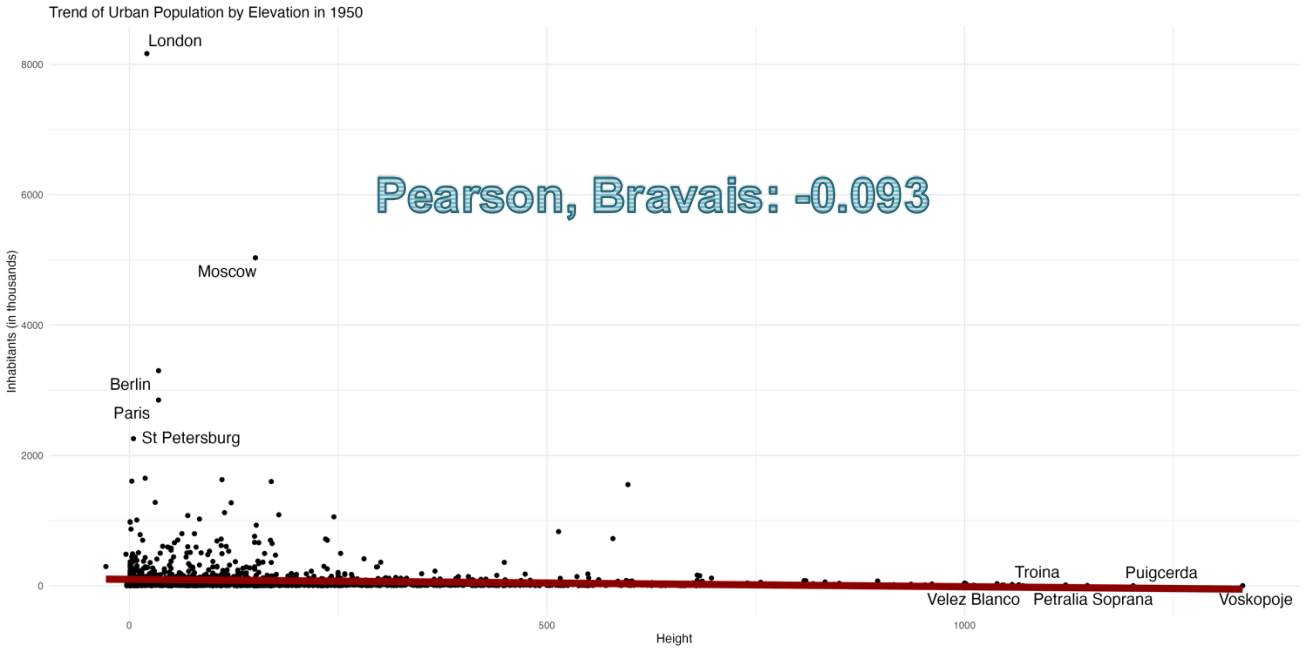
SPEARMAN FOR NON-LINEAR RELATIONSHIPS



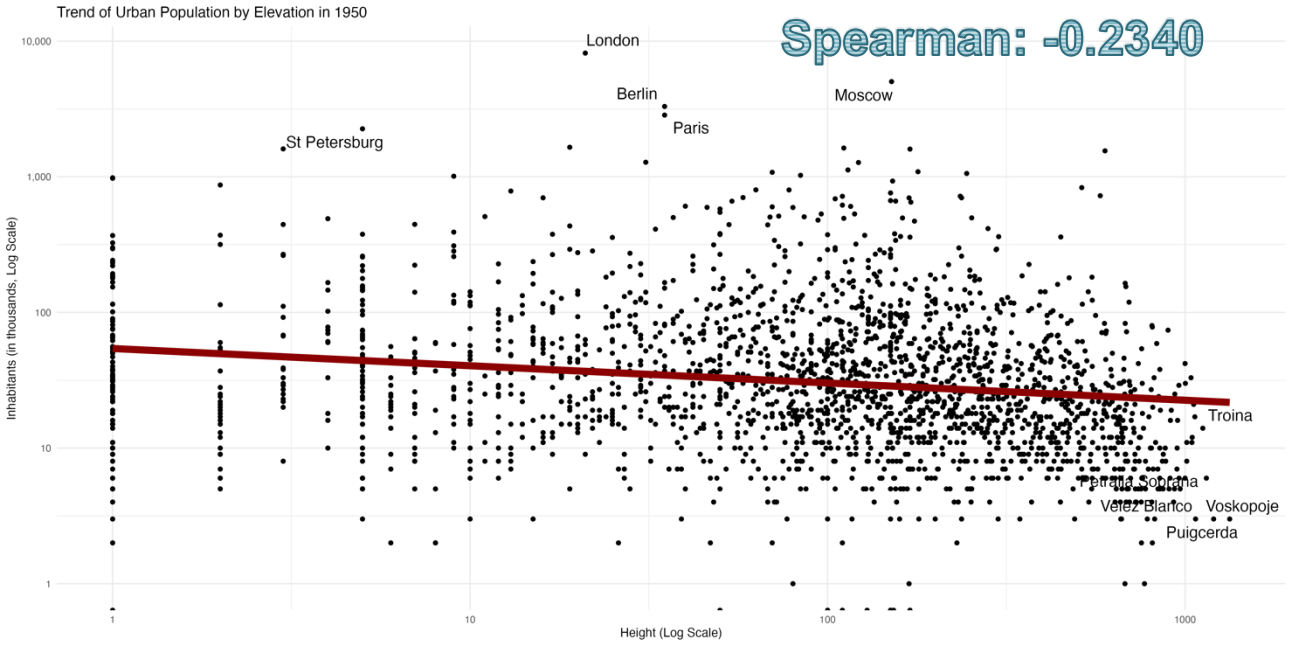
SPEARMAN FOR NON-LINEAR RELATIONSHIPS



SPEARMAN FOR NON-LINEAR RELATIONSHIPS



SPEARMAN FOR NON-LINEAR RELATIONSHIPS



SPEARMAN FOR NON-LINEAR RELATIONSHIPS

- **Summary:** Clearly, overall there is a very strong monotonic relationship.
- The rank transformation, at least in this case, even has a **somewhat linearizing effect**.
- The rank correlation coefficient is thus very **well suited for the numerical description of a monotonic relationship in metric data**.

UNIVERSITÄT
LEIPZIG

SEE YA'LL NEXT WEEK!

Dr. Ing. Andreas Niekler
Computational Humanities

Paulinum, Augustusplatz 10, Raum P 616, 04109 Leipzig
T +49 341 97-32239
andreas.niekler@uni-leipzig.de
https://www.uni-leipzig.de/personenprofil/mitarbeiter/dr-andreas-niekler