UNIVERSITÄT
LEIPZIG

10-207-0003: Introduction to Stochastics

Variance, Transformation, CDF

14.04.2025, Leipzig

Dr. Ing. Andreas Niekler

Computational
Humanities
UNIVERSITÄT LEIPZIG

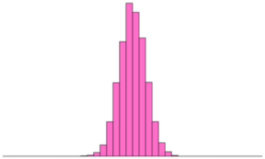
UNIVERSITÄT
LEIPZIG

SYLLABUS

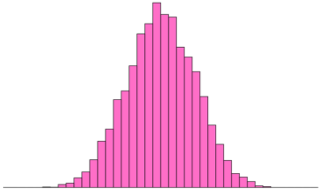
1. Empirical research and scale levels
2. Univariate description and exploration of data
3. Graphical representation of characteristics / Explorative data analysis
4. ***Measures of data distribution***
5. (Combinatorics, permutation) -- > Multivariate Problems, Correlation
6. Probability theory
7. Probability 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

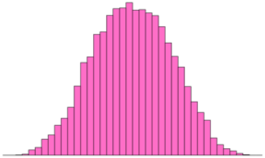
MEASURES OD DISPERSION



Low variability



Medium variability



High variability

- Measures of dispersion answer questions such as:
  - How much do the observations vary, or how large is the variability?
  - What is the average deviation from the mean?
  - Over what range do the observations extend?
- We are already familiar with:
  - Range R or extreme values (xmin, xmax)
  - Interquartile range (IQR) or quartiles (Q1, Q3)

👉 The standard deviation s is a kind of **average distance** of the individual raw values x_i from the mean $\bar{x}$.

VARIANCE AND STANDARD DEVIATION

- The standard deviation s is formally defined via the so-called variance s^2, which itself is a measure of dispersion.

👉 Variance: Let $x_1, \dots, x_n$ be the raw data [or: sample data] of a continuous variable X. Then

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

is called the (empirical) vaiance. The (empirical) standard deviation s of X is defined as the positive square root of the variance:

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

VARIANCE AND STANDARD DEVIATION

i	1	2	3	4	5	…	97	98	99	100	$\sum$
x_i	51	54	76	4	12	…	4	7	10	131	851
$x_i - \bar{x}$	42,49	45,49	67,49	-4,51	3,49	…	-4,51	-1,51	1,49	122,49	-7,105E-1
$(x_i - \bar{x})^2$	1805,4001	2069,3401	4554,9001	20,3401	12,1801	…	20,3401	2,2801	2,2201	15003,8001	26056,99

- S^2 = 263.20
- S = = 16.22
- The data thus **spread** around the mean value 8.51 with a standard deviation of 16.22.

VARIANCE AND STANDARD DEVIATION

- The standard deviation s measures the **dispersion around the mean value** $\bar{x}$ and should only be used when the mean is used as a measure of central tendency.
- The standard deviation vanishes, i.e., $s = 0$, **if and only if there is no dispersion**. Otherwise, it holds that $s > 0$. The further the values are spread, the larger s generally becomes.
- The standard deviation, like the mean $\bar{x}$, is not a robust measure, meaning that a **few outliers can extremely distort the value of** s, which is why the raw data should always be **examined for possible outliers first and potentially cleaned**.
- The standard deviation represents a **natural unit of scale** → standardization of raw values.

VARIANCE SHIFT THEOREM OR COMPUTATIONAL FORMULA FOR VARIANCE

- The following theorem is sometimes used for the more convenient calculation of the variance s^2 or, consequently, the standard deviation s.

- **Variance Shift Theorem:**

$$s^2 := \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \overline{x^2} - \bar{x}^2$$

- **Proof:**

$$s^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)$$

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

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

HOW TO COMPARE EMPIRICAL DISTIBUTIONS

- A quantitative characteristic X can be measured in various units of measurement [X], so that, for example, for a ratio-scaled characteristic X, we can determine the number of units [X] for a specific observation or measurement.
  - We could measure temperature in °C (Celsius) or °F (Fahrenheit)
  - Even in the metric system we can measure distance in km, m, or cm

- But here too, **what we measure is the extent of a property**, which itself is independent of the unit of measurement and is only put in relation to it.

- Metric characteristics are now precisely characterized by the fact that the **choice of the unit of measurement is mostly just expedient, but one can transform a characteristic** X with unit $[X]$ into an equivalent characteristic Y with unit $[Y]$ through a so-called linear transformation.

LINEAR TRANSFORMATION

- The linear transformation

$$Y = a + b \cdot X$$

  transforms a variable $x \in X$ into a new variable $y \in Y$, where each value $x \in \mathrm{X}$ gets transformed with

$$y = a + b \cdot x$$

  Here, $a \in R$ describes a shift, while $b \in R$ causes a scaling.

- Linear transformation does not change the shape of the distribution of the data, but only its location and scaling.

- Inverse Transformation:

$$Y = a + b \cdot X \Leftrightarrow X = \frac{Y - a}{b}$$

LINEAR TRANSFORMATION

- Let $(x_1, \dots, x_n)$ be the raw data list of a quantitative characteristic X.

- Question: How do $\bar{x}$ and s_X change under a transformation from X to Y using the linear transformation $Y = a + b \cdot X$?

- Let $(y_1, \dots, y_n)$ be the corresponding raw data list of the linearly transformed characteristic Y. It holds that $\bar{y} = a + b\bar{x}$, $s_Y^2 = b^2 \cdot s_X^2$ and $s_Y = |b| \cdot s_X$

- The standard deviation s_X and the mean $\bar{x}$ are thus scaled by the factor $|b|$ and b respectively, whereby the scaled mean is additionally shifted by a.

EXAMPLE CELSIUS - FAHRENHEIT

- The conversion between Celsius and Fahrenheit is a linear transformation described by the formula:

$$F = (C \times \frac{9}{5}) + 32$$

- Let's use this to find the boiling point of water in Fahrenheit, given that the boiling point is 100°C.

- **Step-by-Step Conversion:**
  - Start with the temperature in Celsius: C=100°C.
  - Substitute the Celsius value into the formula: $F = (100 \times \frac{9}{5}) + 32$
- Perform the multiplication: $F = \left(\frac{900}{5}\right) + 32 = 180 + 32$

- **Result:**
  - The boiling point of water, which is 100°C, is equal to 212°F.

Z TRANSFORMATION

- A special transformation of a characteristic X to a characteristic Z is obtained by means of

  $$Z = \frac{X - \bar{x}}{s_X} \left(= -\frac{\bar{x}}{s_X} + \frac{1}{s_X} X \right)$$

  that is, we center the data around their mean and choose the standard deviation as the new unit of measurement, i.e., $[Z] = s_X$.

- This transformation is characterized by the fact that for such a transformed raw data, it holds that:
  $\bar{z} = 0$ and $s_Z = 1$
-
  This special transformation is also known as standardization or z-transformation.

- Inverse transformation:

  $$Z = \frac{X - \bar{x}}{s_X} \Leftrightarrow X = \bar{x} + \mathrm{Z} \cdot s_X$$

STANDARDIZATION AND Z-VALUES

- Let $x_1, \ldots, x_n$ be the raw data of a quantitative characteristic X with mean $\bar{x}$ and standard deviation s. Then

  $$z_i = \frac{x_i - \bar{x}}{s}$$

  is the so-called z-score or standardized value of x_i.

- The z-score of a raw value x_i thus indicates how many standard deviations s the value x_i is away from the mean $\bar{x}$.

- This becomes particularly clear when we consider the inverse transformation:

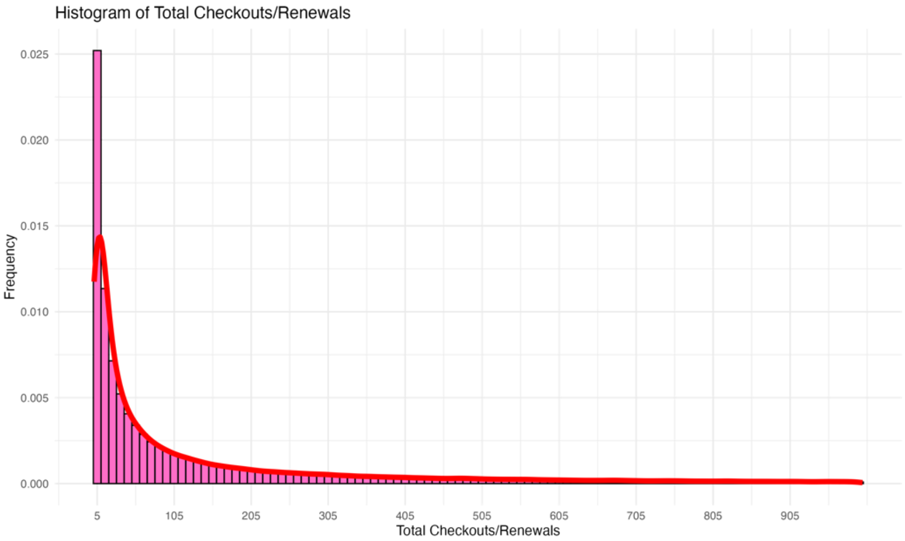
  $$z_i = \frac{x_i - \bar{x}}{s} \Leftrightarrow x_i = \bar{x} + z_i \cdot s$$

STANDARDIZATION AND Z-VALUES

i	1	2	3	4	5	…	97	98	99	100	$\sum$
x_i	51	54	76	4	12	…	4	7	10	131	851
$x_i - \bar{x}$	42,49	45,49	67,49	-4,51	3,49	…	-4,51	-1,51	1,49	122,49	-7,105E-14
z_i	2,61904045	2,8039574	4,16001505	-0,2779918	0,21512006	…	-0,2779918	-0,0930749	0,09184209	7,55015919	-

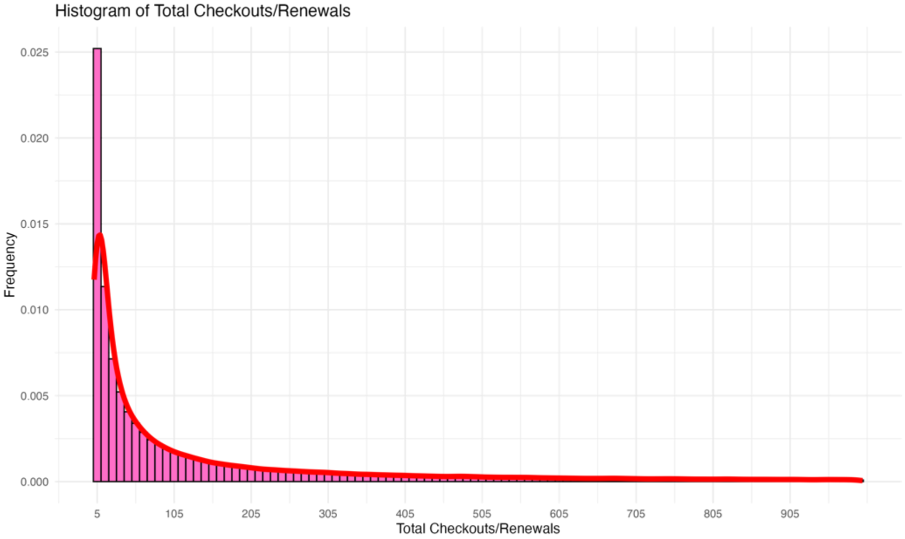
- Some datapoints vary strong from the expectation (mean)
- This has consequences for the usage of those points; outliers etc.
  - Also, we will see later that this variation **can be interpreted as confidence or match (t-test, Hypothesis testing)**

FROM EMPIRICAL DISTRIBUTIONS TO THEORETICAL DISTRIBUTIONS



- Many empirical distributions can be approximated by theoretical distributions: The theoretical distribution is described by a so-called density function, here $f(x) \approx a \cdot e^{-a \cdot x}$, which approximates the histogram with a density curve.

FROM EMPIRICAL DISTRIBUTIONS TO THEORETICAL DISTRIBUTIONS



- Interpretation of density values: Given the scaling, **density indicates approximately how high the proportion of cases per scale unit is**.
- Example. A density of approximately 0.025 at 5 items corresponds to an approximate (local) accumulation of about 2,5% of all cases.

THEORETICAL DENSITY FUNCTION

👉 **Density Function:** A (theoretical) density function f of a quantitative characteristic X, which describes a density curve, is a function $f: X \to \mathbb{R}$ such that

$$f(x) \geq 0, x \in X$$

and

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

– Given a density function that describes a theoretical distribution of the data, the **relative proportion of observations that lie between two values** $a < b$ is given by the corresponding area under the density curve, i.e.,

$$\int_{a}^{b} f(x)dx = 1.$$

CDF FOR THEORETICAL DISTRIBUTION

- Analogous to the empirical cumulative distribution function F(x), which describes how large the relative proportion of raw values is that are less than or equal to x, one can now also define the cumulative distribution function for a theoretical distribution

$$\mathrm{F(x)} := \int_{-\infty}^{\infty} f(x)dx.$$

- It then holds that

$$\int_{a}^{b} f(x)dx = F(b) - F(a).$$

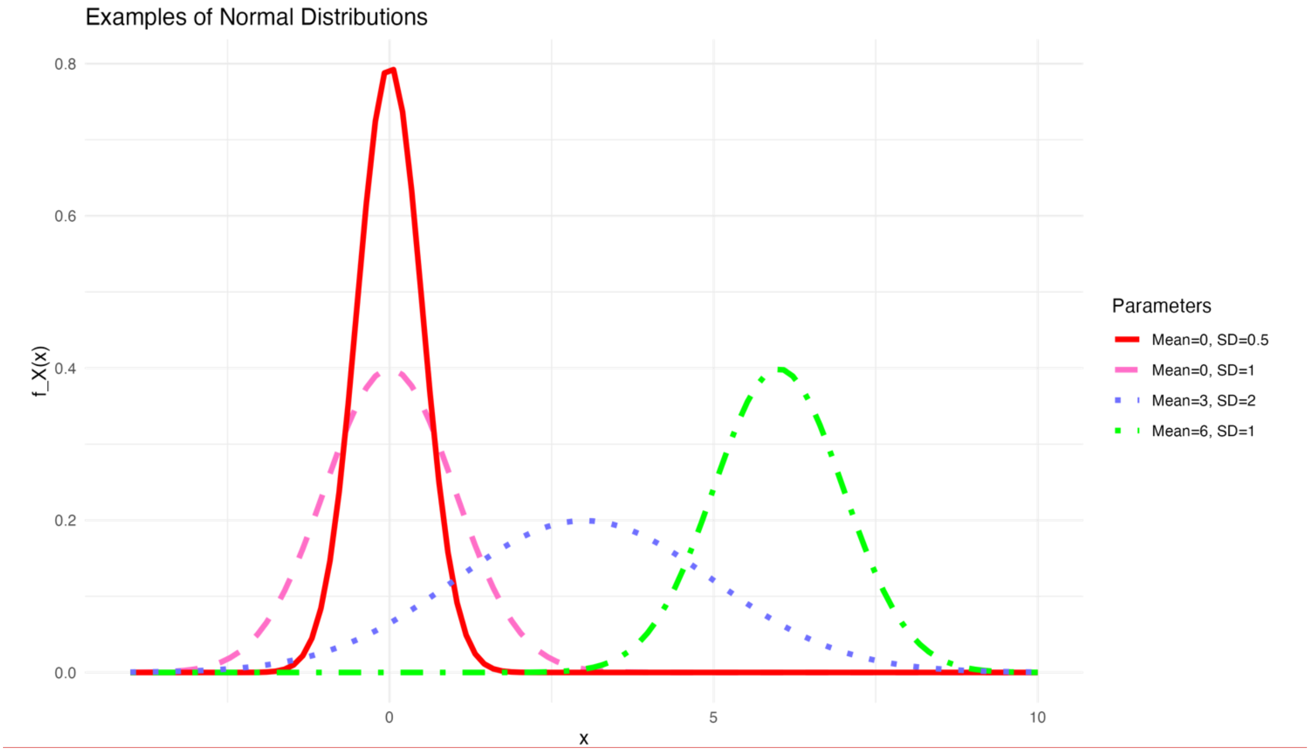
NORMAL DISTRIBUTION

- Normal Distribution: A characteristic X is called normally distributed with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma \in (0, \infty)$, if the density function is given by

$$F_X(x) = \frac{1}{\sigma\sqrt{\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, x \in \mathbb{R}.$$

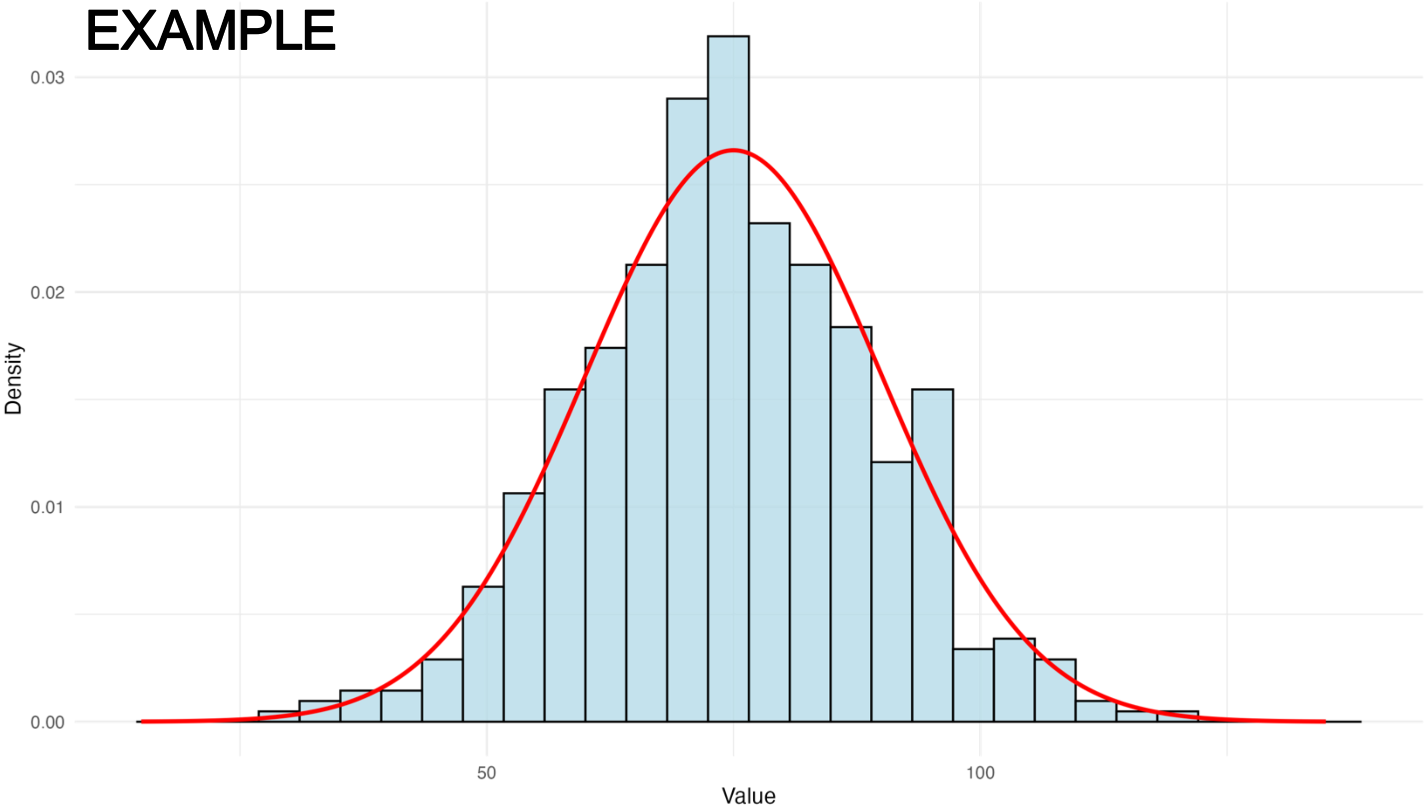
- The distribution is briefly denoted by $N(\mu, \sigma)$.
- μ, mean
- σ, standard deviation (or s as defined before)

NORMAL DISTRIBUTION

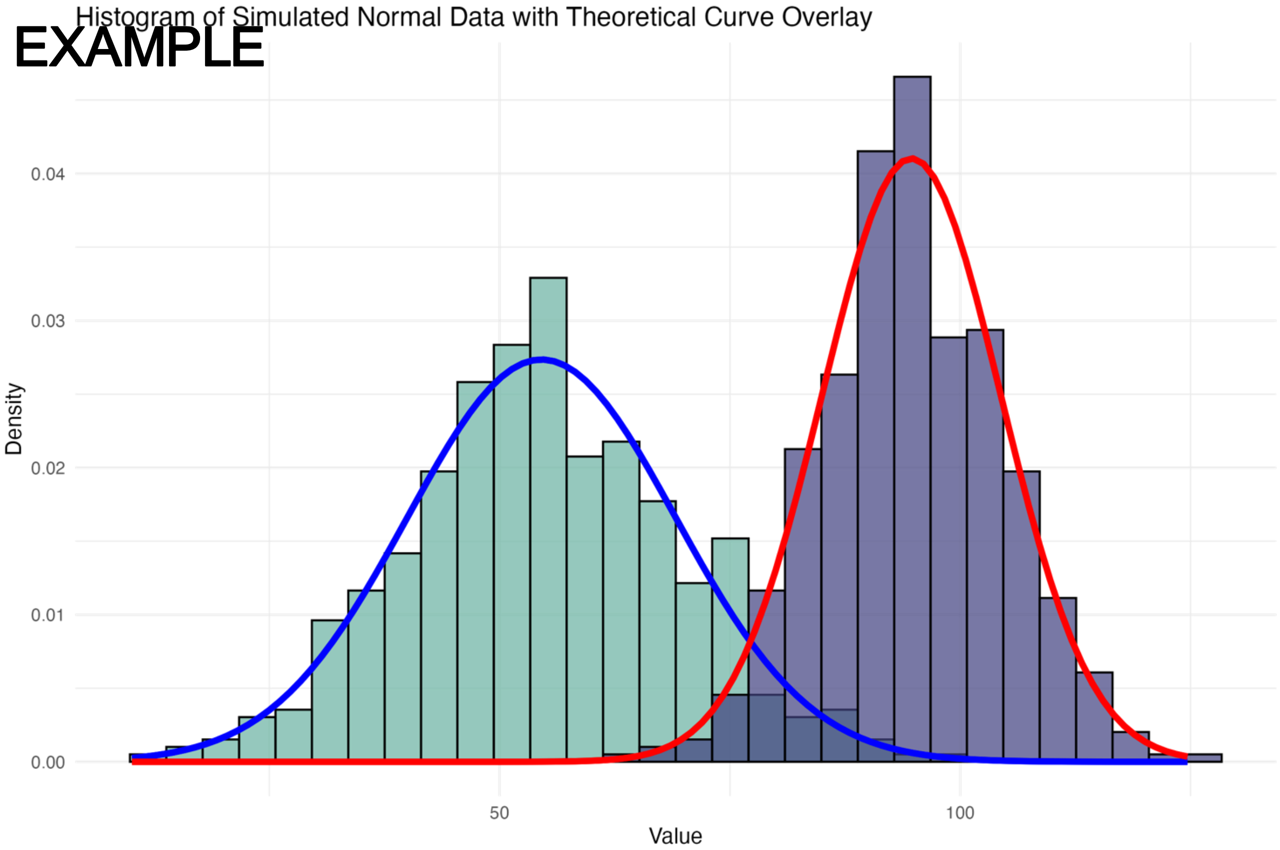


Histogram of Simulated Normal Data with Theoretical Curve Overlay

EXAMPLE



EXAMPLE


Histogram of Simulated Normal Data with Theoretical Curve Overlay
Density
0.04
0.03
0.02
0.01
0.00
50
100
Value

STANDARD NORMAL DISTRIBUTION

- A characteristic Z is called standard normally distributed if Z is normally distributed with $\mu = 0$ and $\sigma = 1$, i.e., the distribution $N(0,1)$ underlies it.
- The standard normal distribution is the prototype of a normal distribution. The corresponding cumulative distribution function F is denoted by Φ and is given by

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{y^2}{2}} dz$$

- For this integral, the so-called Gaussian error integral, no analytical solution exists, **which is why it is either determined numerically or tabulated**.

Z-TRANSFORMATION (STANDARDIZATION)

- Let X be normally distributed with μ and σ. Then it follows that the standardized variable

$$Z(X) := \frac{x-\mu}{\sigma}$$

  is normally distributed.

- Interpretation: If X is normally distributed, then Z describes the corresponding z-transformation of the characteristic or random variable X.

- Important Calculation Rule:

$$F(x) = \Phi\big(z(x)\big)$$

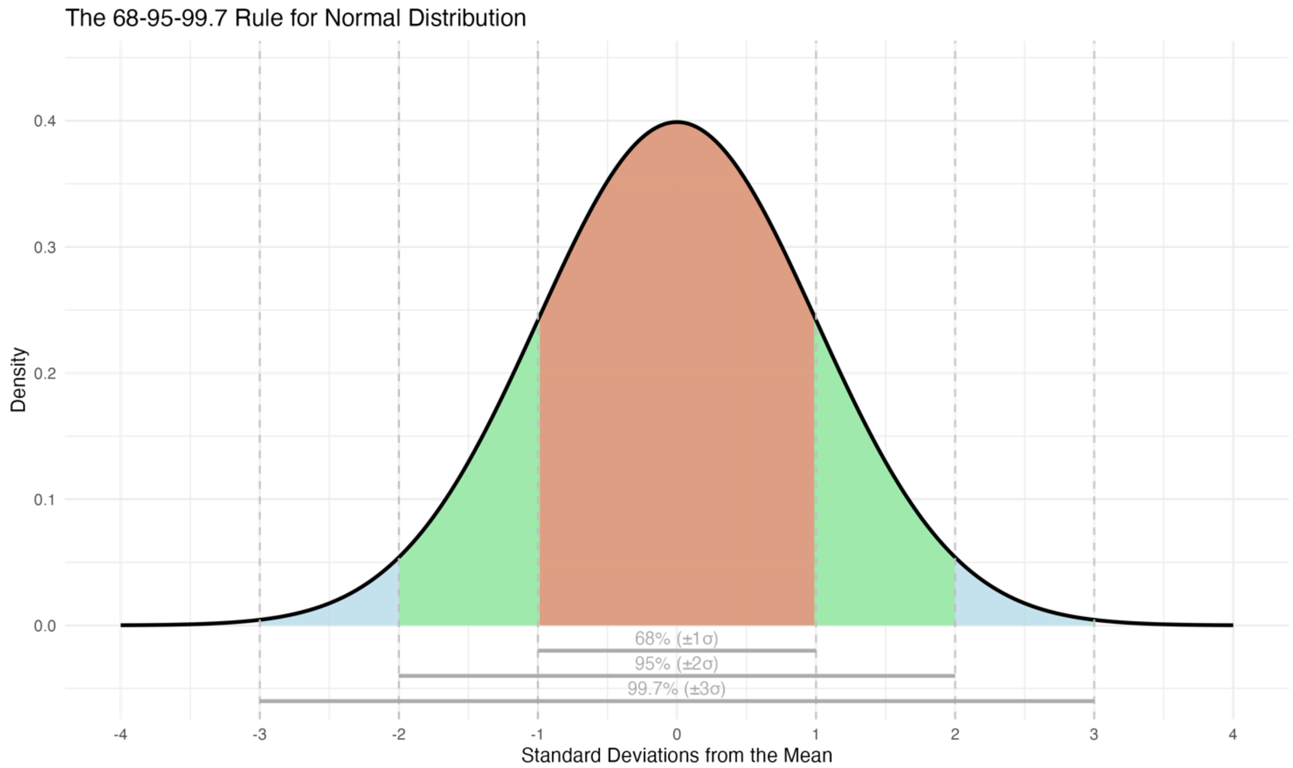
- where the z-value of x is given as

$$Z(X) := \frac{x-\mu}{\sigma}$$

68-95-99.7-RULE FOR NORMAL DISTRIBUTION

- The 68-95-99.7 Rule for normally distributed data states:
  - Approximately 68% of the data is located within the interval $[\mu - \sigma, \mu + \sigma]$
  - Approximately 95% of the data is located within the interval $[\mu - 2\sigma, \mu + 2\sigma]$
  - Approximately 99.7% of the data is located within the interval $[\mu - 3\sigma, \mu + 3\sigma]$

68-95-99.7-RULE FOR NORMAL DISTRIBUTION



EXAMPLE BURROWS DELTA

- **Burrow's Delta:** A technique for authorship attribution, introduced by John F. Burrows, a key figure in modern stylometry.
- **Goal:** Attributes an anonymous text to one of several candidate authors using example texts.
- **Method:** Can be viewed as a simple machine learning method for text classification.
- **Fitting (Training):** Stores example texts from candidate authors.
- **Testing (Prediction):** Attributes a new text to the author of the *most stylistically similar* training document (the "nearest neighbor").
- **Effectiveness:** Simple yet often produces strong results, particularly with longer texts (e.g., novels).
- **Core Metric:** Calculates the stylistic distance based on the mean of absolute differences between **z-scores of word frequencies** in the test text and training texts.

John Burrows. 'Delta': A measure of stylistic difference and a guide to likely authorship.*Literary and Linguistic Computing*, 17(3):267–287, 2002.

DEFINITION BURROWS DELTA

- $z(\vec{x}_i) := \frac{\vec{x}_i - \mu_i}{\sigma_i}$, μ_i and σ_i respectively stand for the sample mean and sample standard deviation of the word's frequencies in the reference corpus.

- Delta or stylistic difference between $\vec{x}_i$ and $\vec{y}_i$ following, Burrows's definition, $\Delta(\vec{x}_i, \vec{y}_i)$ is then the mean of the absolute differences for the z-scores of the words in them:

$$\Delta(x, y) = \frac{1}{n} \sum_{i=1}^{n} |z(x_i) - z(y_i)|$$

John Burrows. 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287, 2002.

PROPERTIES OF BURROWS DELTA

- **Focus on Function Words:** Burrows's Delta, like much stylometric research, primarily analyzes function words (e.g., prepositions, articles).
- **Definition:** These are a small set of typically short, grammaticalized words that carry little meaning in isolation.
- Advantages for Author Identification:
  - **Frequency & Distribution:** They are frequent and evenly distributed across texts, unlike content words which vary by topic.
  - **Topic/Genre Independence:** Not tied to specific subjects or genres.
  - **Unconscious Use:** Psycholinguistic research suggests they are used less consciously, making them harder to imitate or forge stylistically.

John Burrows. 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287, 2002.

NEIREST NEIGHBOR CLASSIFIER

- $\operatorname*{argmin}_{i \in \{1,\ldots,n\}} \Delta(x, y_i)$
- Classification: Select the author where the distance in training document in minimized.
- Training: Determine μ_i and σ_i from exiting data
- Filter corpus by functional words
  - Top n frequency words – functional words
  - Word List for functional words

John Burrows. 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287, 2002.

APPLICATION OF BURROWS DELTA

- **Burrows's Delta:** Originally a **supervised classification** method using **nearest neighbor** logic to attribute authorship based on labeled training data ("ground truth").
  - **Supervised Learning:** Relies on labeled example data for training (like Burrows's Delta).
  - **Unsupervised Learning / Exploratory Data Analysis:** Analyzes data without ground truth labels to find inherent structure or groups (e.g., in anonymous texts).
- **Clustering:** A common unsupervised technique in stylometry (and other text analysis) to group stylistically similar texts.
  - **Hierarchical (Agglomerative) Clustering:** A bottom-up clustering method that merges texts into increasingly larger clusters based on similarity, often visualized with a dendrogram.

John Burrows. 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287, 2002.

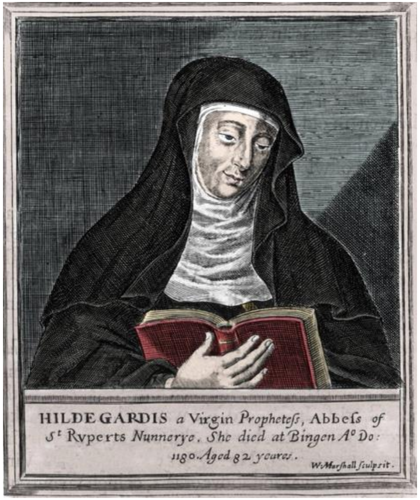
PROBLEM EXAMPLE

- **Subject:** A stylometric study examining the authenticity of letters attributed to Hildegard of Bingen, a rare female Latin prose author from the 12th century. Hildegard von Bingen (1098-1179) was a German Benedictine abbess, polymath, composer, and mystic of the Middle Ages. Known for her writings on religion, medicine, and science, she was an influential advisor and preacher. She is venerated as a saint and recognized as a Doctor of the Catholic Church, with commemorations in other Christian denominations as well.

- **Comparison Corpus:** Hildegard's letters were compared to those of two contemporaries: Bernard of Clairvaux and Guibert of Gembloux (her last secretary).

- **Method/Finding:** Stylometric analysis showed remarkably clear clusters for each author.

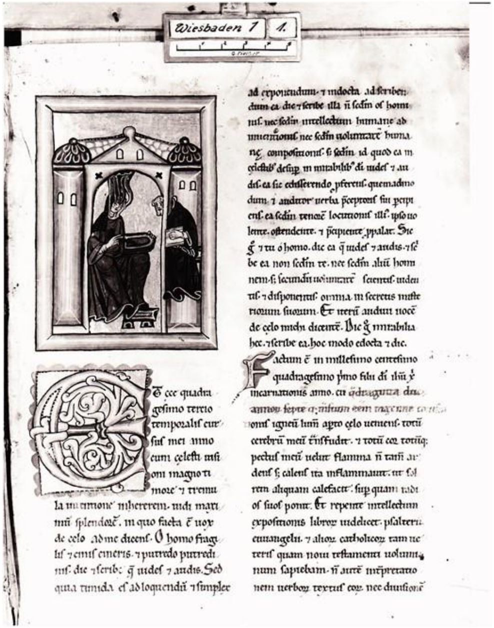
- **Conclusion:** The distinct clusters indicate that the three authors employed markedly different writing styles.


HILDEGARDIS a Virgin Prophetess, Abbess of
St Rvperts Nunnerye. She died at Bingen Ao Do:
1180. Aged 82 yeares.
W.Marshall sculpsit.

John Burrows. 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287, 2002.

PROBLEM EXAMPLE

- Two letters are extant which are commonly attributed to Hildegard herself, although philologists have noted that these are much closer to Guibert’s writings in style and tone. These texts titles have been abbreviated: D_Mart.txt (*Visio de Sancto Martino*)
and D_Missa.txt (*Visio ad Guibertum missa*). Note that the prefix D_ reflects their **D**ubious authorship.

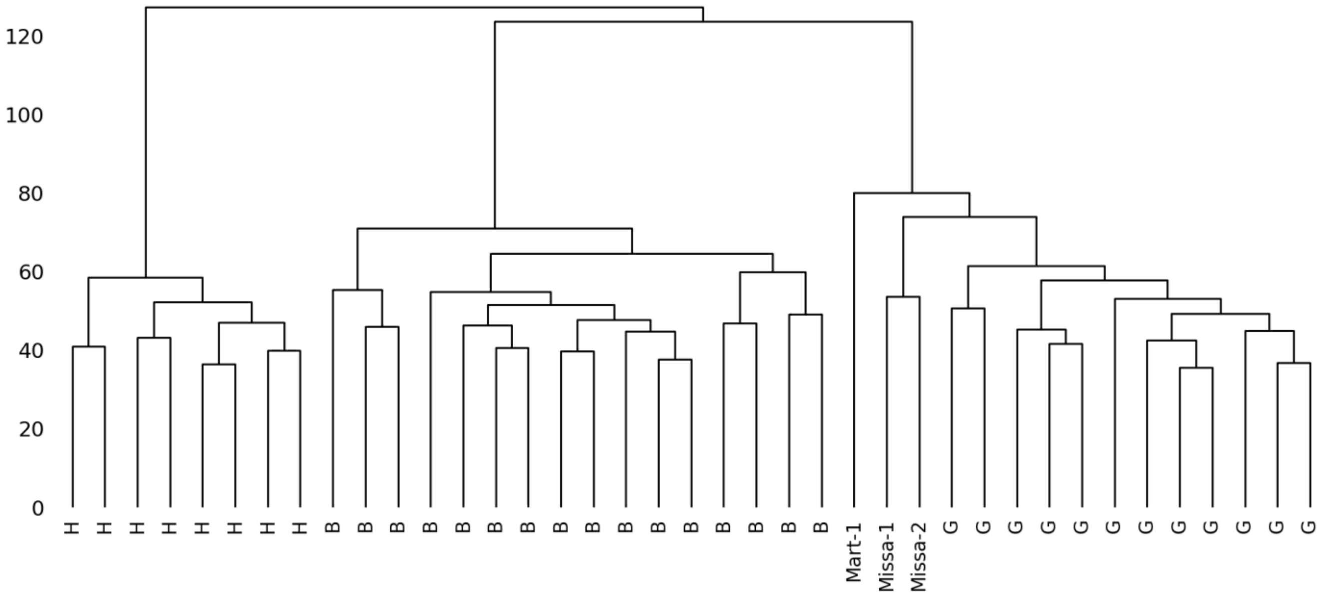


John Burrows. 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287, 2002.

https://www.humanitiesdataanalysis.org/stylometry/notebook.html

Mike Kestemont, Sara Moens, Jeroen Deploige, Collaborative authorship in the twelfth century: A stylometric study of Hildegard of Bingen and Guibert of Gembloux, *Digital Scholarship in the Humanities*, Volume 30, Issue 2, June 2015, Pages 199–224, https://doi.org/10.1093/llc/fqt063

DEFINITION BURROWS DELTA



John Burrows. 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287, 2002.

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