



UNIVERSITÄT
LEIPZIG

10-207-0003: Introduction to Stochastics

Univariate description and exploration of data

14.04.2025, Leipzig

Dr. Ing. Andreas Niekler



SYLLABUS

1. Empirical research and scale levels
2. Univariate description and exploration of data
3. *Graphical representation of characteristics / Explorative data analysis*
4. The random experiment
5. Combinatorics, permutation
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

DATA

- In the context of descriptive statistics, we will not make a relevant distinction between a sample and the population for the time being
- The data with n statistical (u)nits u_1, \dots, u_n is defined as a set and is designated with a large letter, for example

$$G = \{u_1, \dots, u_n\}.$$

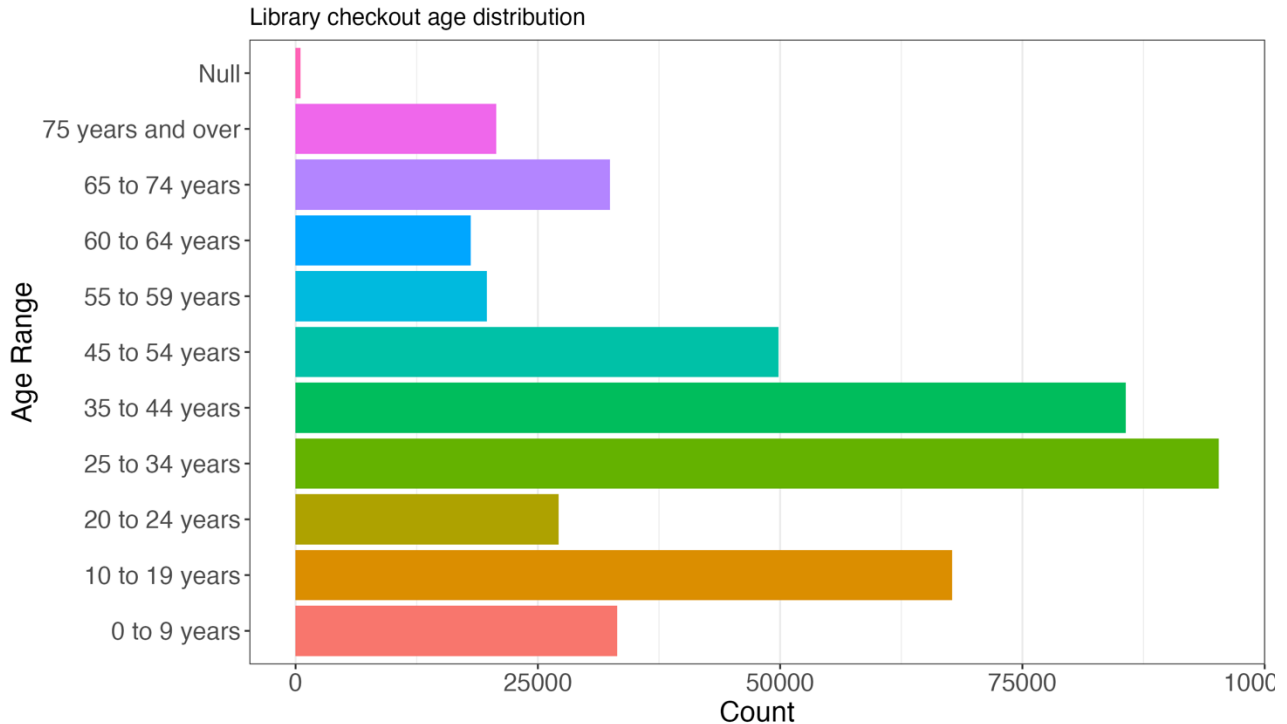
- The numer $n \in \mathbb{N}$ of static units in G is called the *size* or *scope*

$$n := |G|$$

FEATURES

- We do not measure the statistical units themselves, but rather the **features** they possess. Characteristics are usually denoted by capital Latin letters such as X, Y, Z .
 - Example: X — Gender, Y — Age
- All possible values for a feature form a set
 - $X = \{1, 2, 3\}$, where 1 represents *male*, 2 represents *female*, and 3 represents *diverse*.
 - $Y = \{y \in \mathbb{N} | y \geq 0\}$
- Values of features are denoted by lowercase Latin letters such as x, y, z .
 - $y = 2$ is the concrete observation of the characteristic value *female*

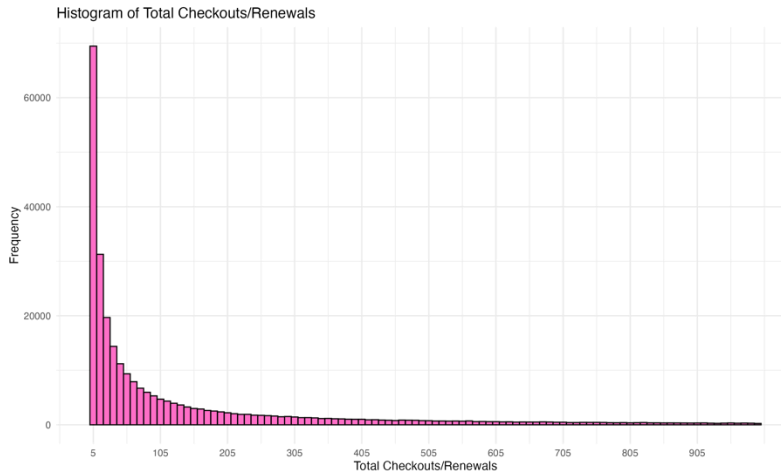
PLOTTING OF FREQUENCY DISTRIBUTIONS



LA library services

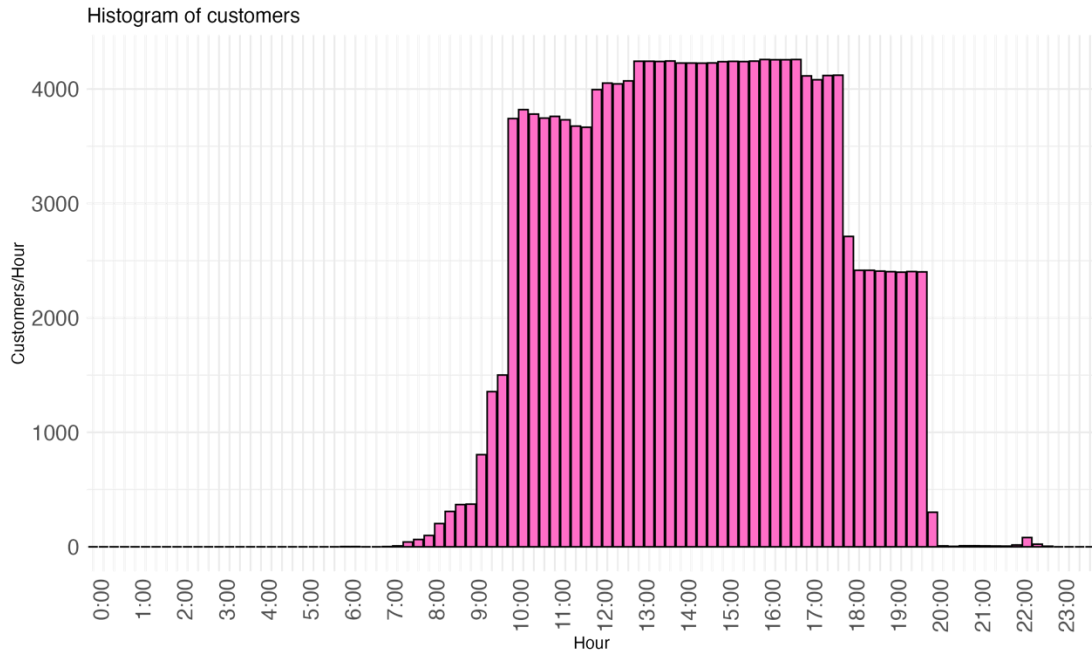
HISTOGRAM (EXAMPLE CONTINUED)

- We choose $d = 10$ and $m_1 = 5$ etc.
- The interval is $[0, 1000]$ which fall in $\frac{1000}{10} = 100$ classes $[0,10), [10,20), \dots, [990, 1000]$



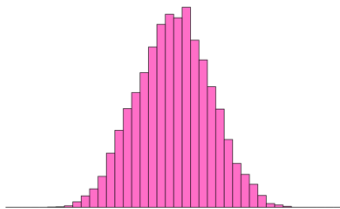
FROM HISTOGRAMS TO DISTRIBUTIONS

- Consider a Histogram of library checkouts during the day

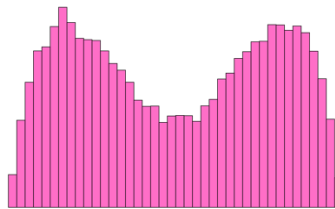


FORMS OF DISTRIBUTIONS - MODES

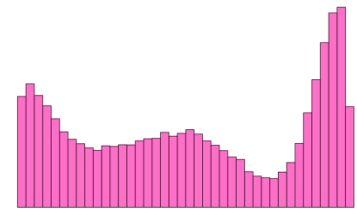
- The shape of a distribution can be assessed, for example, by the **number of modes (peaks)**:



Unimodal = single-peaked



Bimodal = double-peaked



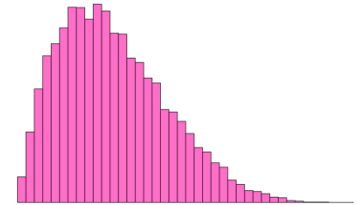
Multimodal = multi-peaked

- If several **clearly separable modes** occur, this often indicates a **mixture of several subpopulations**.

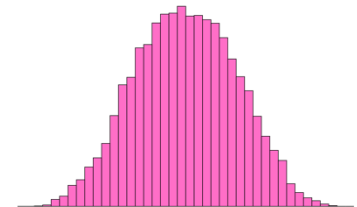
FORMS OF DISTRIBUTIONS - SYMMETRY

- The shape of a distribution can be assessed by its **symmetry or skewness**:

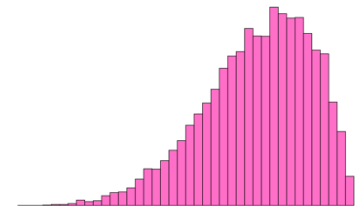
- **left-skewed or right-tailed**: distribution falls much more steeply to the left than to the right



- **symmetric**: right and left halves are approximately mirror images

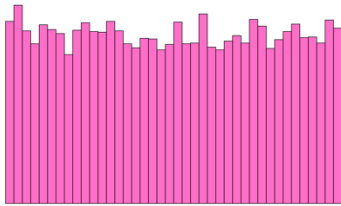


- **right-skewed or left-tailed**: distribution falls much more steeply to the right than to the left

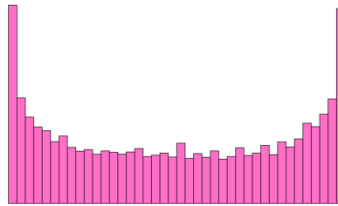


FORMS OF DISTRIBUTIONS - OTHERS

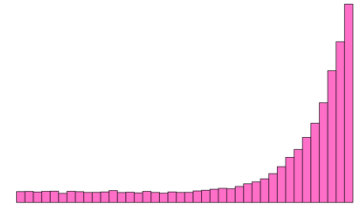
- Other typical forms of distributions are uniform, u-shaped and j-shaped distributions:



Uniform distribution



U-shaped distribution



J-shaped distribution

CUMULATIVE ABSOLUTE FREQUENCIES (CAF)

- The frequency distributions c_1, \dots, c_k or f_1, \dots, f_k of a characteristic in a population G describe the respective number or relative proportion of the values a_1, \dots, a_k in the feature list x_1, \dots, x_n .
 - For many problems, it's more interesting to know what the **relative proportion of statistical units is that fall below or exceed a certain value**.
- Example: Library visitors
 - What is the proportion of those who visit the library before 12 o'clock?"

CAF DEFINITION

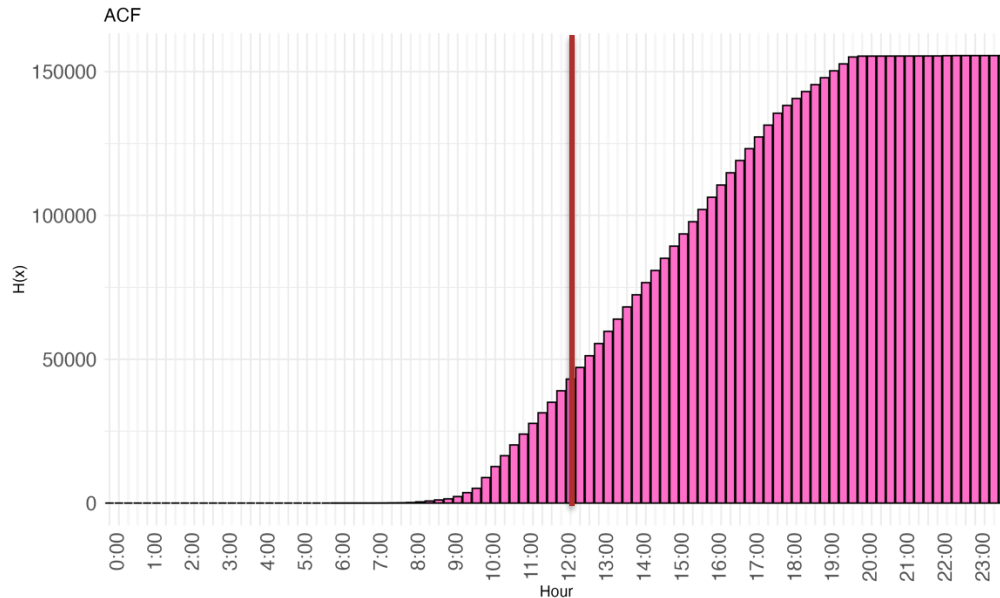
- Given the data list x_1, \dots, x_n of a characteristic $X \in \mathcal{X}$ with frequency distribution c_1, \dots, c_k or f_1, \dots, f_k .
- 👉 For each $x \in \mathcal{X}$, the *cumulative absolute frequency* $H(x)$ is defined as the **number** of (raw) values x_i such that $x_i \leq x$.
The function $H : \mathcal{X} \rightarrow N_0$ is called the *cumulative absolute frequency distribution*.
- Note: *Cumulative* means to accumulate or gather.
- The formal definition is:

$$H(x) := \sum_{j: a_j \leq x} c_j$$

- Note: The notation $\sum_{j: a_j \leq x}$ means that the sum is taken over all indices j for which $a_j \leq x$.

CAF LIBRARY EXAMPLE

- What is the **count** of those who visit the library before 12 o'clock?"



CUMULATIVE DENSITY FUNCTION (CDF)

- Analogous to cumulative absolute frequencies, relative empirical frequencies can also be cumulated.

👉 For each $x \in \mathcal{X}$, the *cumulative density function* $F(x)$ is defined as the **proportion** of (raw) values x_i such that $x_i \leq x$.
The function $F: \mathcal{X} \rightarrow [0,1]$ is called the *cumulative density function*.

- The formal definition is:

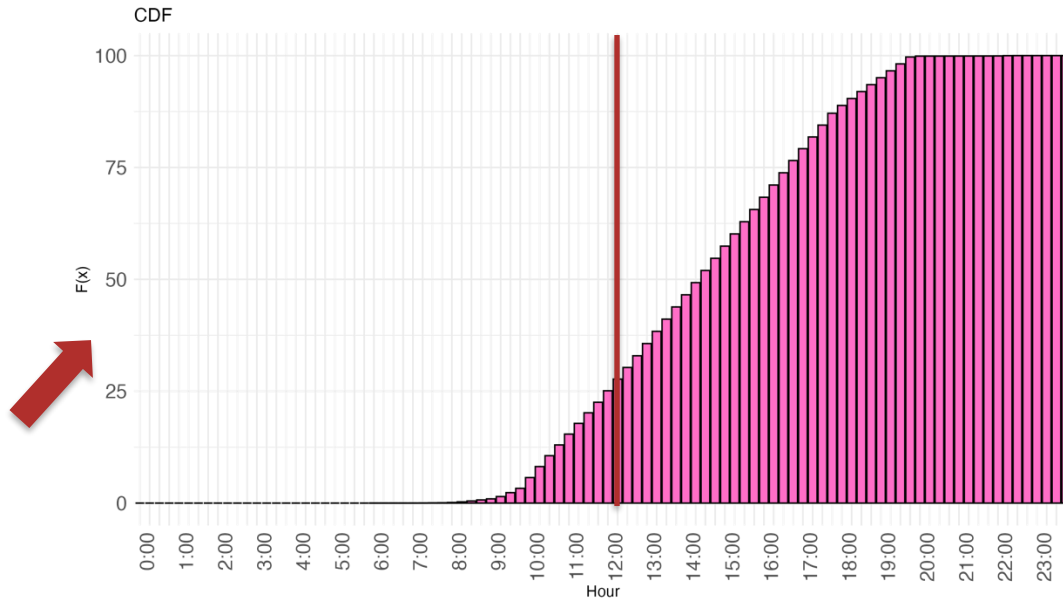
$$F(x) := \sum_{j: a_j \leq x} f_j$$

- Relationship with $H(x)$:

$$F(x) = \frac{1}{n} H(x) \Leftrightarrow H(x) = n \cdot F(x)$$

CDF LIBRARY EXAMPLE

- What is the **proportion** of those who visit the library before 12 o'clock?"



SUMMARY DISTRIBUTION

- Graphical representations of frequency distributions give us a general impression of the distribution of data of an empirical characteristic:
 - Location and center of the data
 - Dispersion of the data around this center
 - Skewness/symmetry/unimodality/multimodality of the data

CENTRAL TENDENCY

- A statistical measure or statistical characteristic t is formally defined as a mapping or calculation rule such that each possible raw data list x_1, \dots, x_n is assigned a statistical characteristic $t(x_1, \dots, x_n) \in \mathbb{R}$.
 - The most trivial characteristic is, in a way, n , i.e., the size of the population under consideration.
 - Another simple characteristic would be, for example, the **number of actually observed values**.
- Statistical characteristics generally aim to **reduce complex information to meaningfully interpretable numbers** (actual information gain through complexity reduction).

EXAMPLE: WAITING TIME

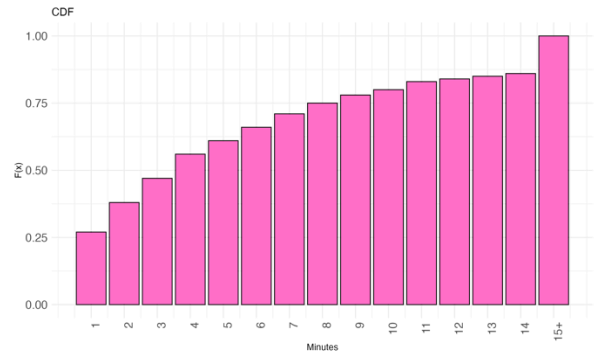
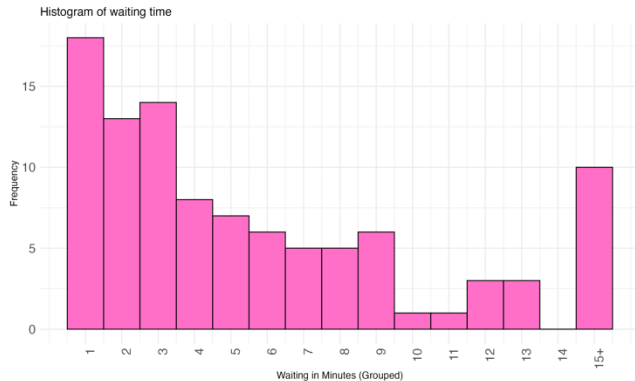
- The librarians Olivia and Carl both record the time X (in minutes) they must wait for a customer on a given day:
- Here is the raw data:

```
# A tibble: 100 × 3
  CheckoutDateTime waiting_time
<dtm>             <dbl>
1 2017-01-02 08:13:00         13
2 2017-01-02 09:40:00        87
3 2017-01-02 09:45:00         5
4 2017-01-02 09:58:00        13
5 2017-01-02 11:02:00        64
6 2017-01-02 11:05:00         3
7 2017-01-02 11:13:00         8
8 2017-01-02 11:15:00         2
9 2017-01-02 11:24:00         9
10 2017-01-02 11:27:00         3
# i 90 more rows
```


- First informative indicators are $n = 100$ and $S = \sum x_i = 700$,
- Together, they waited a total of $700 \text{ minutes} * \frac{1}{60} \approx 14h$.

EXAMPLE: WAITING TIME

— Histogramm and Cummulated Density Function



MEASURES OF CENTRAL TENDENCY

- Using the example, we will particularly discuss various measures of central tendency, which are intended to describe the **central tendency of the data using numerical values**.
 - A measure of central tendency should accomplish approximately the following:
 - Where are **most of the observations** located?
 - Where is the **center of gravity** of the distribution?
 - Where is the **middle of the distribution**?
 - What is a **typical observed value**?
 - Of course, there is not one measure of central tendency, but rather **various appropriate measures**.
-  How appropriate a measure of central tendency **depends on the distribution shape of the data and the scale level**.

ARITHMETIC MEAN

👉 Let x_1, \dots, x_n be the raw data list of a characteristic X . Then

$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} (x_1 + \dots + x_n)$$

is called the arithmetic mean of the data x_1, \dots, x_n

- The arithmetic mean is commonly referred to as the mean, average, or simply mean.
- Example: Library ($S = 701$, $n = 100$):
 - Arithmetic mean: $\tilde{x} = \frac{1}{100} (5 + 10 + \dots + 6 + 16) = 7,01$
 - The librarians waited on average 7 minutes and 1 seconds for a customer.
 - **Note:** $0.01min * \frac{60s}{1min} \approx 1s$

ARITHMETIC MEAN

👉 Let x_1, \dots, x_n be the raw data list of a characteristic X . Then

$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} (x_1 + \dots + x_n)$$

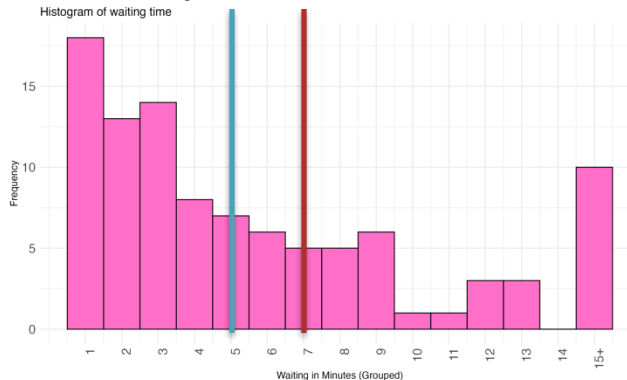
is called the arithmetic mean of the data x_1, \dots, x_n

- The arithmetic mean is commonly referred to as the mean, average, or simply mean.
- Example: Library ($S = 701$, $n = 100$):
 - Arithmetic mean: $\tilde{x} = \frac{1}{100} (5 + 10 + \dots + 6 + 16) = 7,01$
 - The librarians waited on average 7 minutes and 1 seconds for a customer.
 - **Note:** $0.01min * \frac{60s}{1min} \approx 1s$

ARITHMETIC MEAN

👉 An interesting characterization of the arithmetic mean as a **measure of the location of a frequency distribution** results from its center-of-mass property

- There are values larger than 15 in this example
- **Note:** The blue line would be the mean if we take only values lower or equal to 15 into the mean calculation



ARITHMETIC MEAN

- Arithmetic means are particularly **suitable for projections, extrapolations, or estimations**:
- The the library is opened on average $y = 12$ hours per day and $z = 25$ days per month. How many people k can he expect in a month?
 - $k = \frac{12 \cdot 25 \cdot 60}{7,01} \approx 2568$
 - Note: 1h = 60min
- We expect 2568 people to visit the library from the observation of one day!

ARITHMETIC MEAN REMARKS

- Alternatively, we could also calculate:

$$\bar{x} := \frac{1}{n} \sum_{j=1}^k a_j c_j = \sum_{j=1}^k a_j f_j$$

- While you can mathematically calculate the average of any set of numbers, the **result is only truly meaningful when dealing with data that has a clear, numerical scale** (metric data).



Arithmetic means are **extremely sensitive to outliers in the data or changes in a few values** in the raw data.

ARITHMETIC MEAN REMARKS

- The **Will Rogers phenomenon**, also known as "stage migration", is a statistical phenomenon where the average values of two groups increase (or decrease) when members of one group are moved to the other. This can lead to misleading conclusions.
 - A library, initially categorizing customers as "Frequent" (5+ books/month) or "Occasional" (<5 books/month), **changes its criteria to 7+ and <7 books**. This reclassification shifts customers who borrowed 5-6 books from "Frequent" to "Occasional." Consequently, **the average books borrowed by both groups appears to increase**, not because customers are reading more, but solely due to the change in categorization, creating a **misleading impression of improved reading habits**.

MEDIAN

- The (empirical) median \tilde{x} is the midpoint of a distribution, which can be determined as follows:

1. Create the ordered raw data:

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

2. If n is odd, then:

$$\tilde{x} := x_{\left(\frac{n+1}{2}\right)}$$

i.e., the observed middle value of the data.

3. If n is even, then:

$$\tilde{x} = \frac{1}{2} \left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right)$$

i.e., the arithmetic mean of the lower median $x_{\left(\frac{n}{2}\right)}$ and the upper median $x_{\left(\frac{n}{2}+1\right)}$.

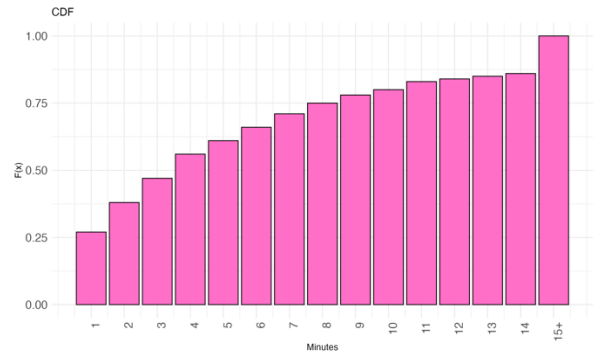
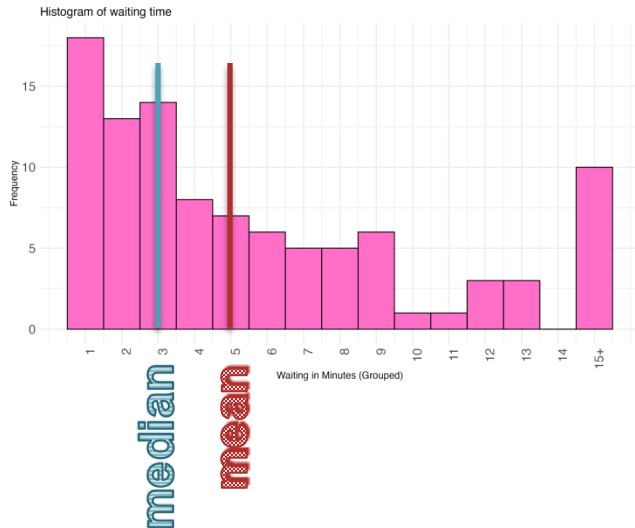
There isn't just **one median, but rather something like a median interval**. The median defined here, however, is the **midpoint of this interval**.

MEDIAN

- An alternative measure of central tendency can be defined by a different meaningful center-of-mass property:
 - **50% of the data should lie to the right and left of a central point.**
 - If we **track the median number of books borrowed** in each group, the change in categorization will have a smaller impact.
 - While the mean might show a significant jump, the **median will likely show a more subtle change**, reflecting the actual change in the "typical" customer's behavior.
- Example:
 - Frequent > 5 Books: $Occ=[1,1,2,\mathbf{2},3,4,4]$, $Freq=[5,6,7,\mathbf{7},8,9,100]$
 - $Mean_{Occ} = 2,4$, $Mean_{Freq} = 20,2$
 - Frequent > 7 Books: $Occ=[1,1,2,2,\mathbf{3},4,4,5,6]$, $Freq=[7,7,\mathbf{8},9,100]$
 - $Mean_{Occ} = 3,4$, $Mean_{Freq} = 26,2$

MEDIAN

- Example Library: If the waiting time is classified as the *typical* middle value using the arithmetic mean instead of the median, they systematically overestimate it by 2 minutes more than it is?



CRITIQUE

- However, stating a measure of central tendency alone can be quite misleading.
- Example:
 - In one library, all users borrow an average of 4 items.
 - In another library, 25% borrow 2 items, 50% borrow 4 items, and 25% borrow 6 items.
 - In both libraries, both the average and the median are 4.
 - Obviously, the dispersion is greater in one library, while the borrowing in the other library is constant.
- The difference in spread is crucial for understanding the data fully.

MINIMUM AND MAXIMUM

- We denote the smallest value of an observation in a raw data list x_1, \dots, x_n as x_{min} , and the corresponding largest value as x_{max} :

$$x_{min} = \min\{x_1, \dots, x_n\} \text{ and } x_{max} = \max\{x_1, \dots, x_n\}$$

- With the minimum and maximum, we can roughly estimate the range in which we effectively observe characteristic values of a feature.

MINIMUM, MAXIMUM, RANGE

👉 We denote the smallest value of an observation in a raw data list x_1, \dots, x_n as x_{min} , and the corresponding largest value as x_{max} :

$$x_{min} = \min\{x_1, \dots, x_n\} \text{ and } x_{max} = \max\{x_1, \dots, x_n\}$$

- With the minimum and maximum, we can roughly estimate the range in which we effectively observe characteristic values of a feature.

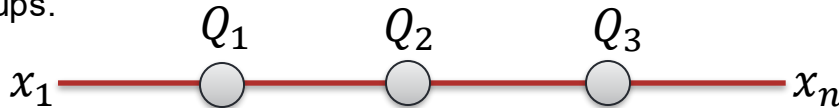
👉 The range R is defined as the difference between the minimum and maximum of a raw data list x_1, \dots, x_n :

$$R = x_{max} - x_{min}$$

- The range R is a measure of dispersion, but it reflects less what one might intuitively understand as the typical dispersion because of its sensitivity to outliers.

QUARTILE

👉 The quartiles Q_1, Q_2, Q_3 of a raw data list x_1, \dots, x_n are those midpoints that divide the ordered raw data list $x_{(1)}, \dots, x_{(n)}$ into four approximately equal-sized groups.



- This definition is also not sufficiently precise – see median. There are approximately 10 different ways in the literature to determine the quartiles specifically."
- However, we will content ourselves here with the following pragmatic definition by [1]:
 - We choose the median defined above as the second quartile,

$$Q_2 := \tilde{x}$$
 - The first quartile Q_1 is the median of all observations $x_{(1)}, \dots, x_{(n)}$ that lie to the left of \tilde{x} in the ordered data.
 - The third quartile Q_3 is the median of all observations $x_{(1)}, \dots, x_{(n)}$ that lie to the right of \tilde{x}

[1] Moore, D. S. and McCabe, G. P. Introduction to the Practice of Statistics, 4th ed. New York: W. H. Freeman, 2002.

QUARTILE EXAMPLE

- The following four examples list all possible *complications* (forming arithmetic mean intermediate values) when determining the three quartiles.
 - Sorted data $(1, 2, 3, 4) \rightarrow (1, 2) (3, 4)$
 - $(Q_1, \tilde{x}, Q_3) = (1.5, 2.5, 3.5)$
 - Sorted data $(1, 2, 3, 4, 5) \rightarrow (1, 2) (4, 5)$
 - $(Q_1, \tilde{x}, Q_3) = (1.5, 3, 4.5)$
 - Sorted data $(1, 2, 3, 4, 5, 6) \rightarrow (1, 2, 3) (4, 5, 6)$
 - $(Q_1, \tilde{x}, Q_3) = (2, 3.5, 5)$
 - Sorted data $(1, 2, 3, 4, 5, 6, 7) \rightarrow (1, 2, 3) (5, 6, 7)$
 - $(Q_1, \tilde{x}, Q_3) = (2, 4, 6)$

INTER QUARTILE RANGE (IQR)

- The lower quartile Q_1 and the upper quartile Q_3 are themselves measures of central tendency and describe the center of the data - **the central 50% of the data**.
 - With the help of these two values, a meaningful **measure of dispersion for metric characteristics** can now be defined.
- Interquartile Range: IQR is defined as the difference between the first and third quartiles:

$$IQR := Q_3 - Q_1$$

- The interquartile range is **robust against outliers**, since only the central 50% of the data are included.

IQR EXAMPLE

- Library Example:
 - $Q_1 = 2,0$
 - $Q_3 = 7.25$
 - $IQR = 7.25 - 2.0 = 5.25$
- The **three quartiles together with the minimum and maximum** of a raw data list give us a suitable numerical summary of the distribution of the data.

$$(x_{min}, Q_1, \tilde{x}, Q_3, x_{max}) = (1.0, 2.0, 3.0, 7.25, 189)$$

FIVE-NUMBER SUMMARY

👉 The five-number summary of a raw data list x_1, \dots, x_n consists of the minimum, first quartile, median, third quartile, and maximum, in order of size $\rightarrow (x_{min}, Q_1, \tilde{x}, Q_3, x_{max})$

- Example Library:
 - $(x_{min}, Q_1, \tilde{x}, Q_3, x_{max}) = (1.0, 2.0, 3.0, 7.25, 189)$
 - From these five values, one can quite well recognize the general characteristics of the distribution of waiting times.
 - For example, the distance from Q_1 to \tilde{x} is smaller than that from Q_3 to \tilde{x} , which is typical for a left-steep or right-skewed distribution.
 - Also, with $x_{max} = 189$, an unusually high value occurs, since $IQR = 5.25$ and thus x_{max} deviates from the upper quartile Q_3 by approximately $(189 - 7.25) / 5.25 = 34.6$ IQRs.

OUTLIERS

- With the help of the IQR, we can specify a criterion for what a possible outlier might be.

👉 **Possible Outlier Value:** We suspect an observation value x_i from the raw data list as a possible outlier if it is more than $1.5 \times IQR$ above Q_3 or below Q_1 , i.e., outside the interval

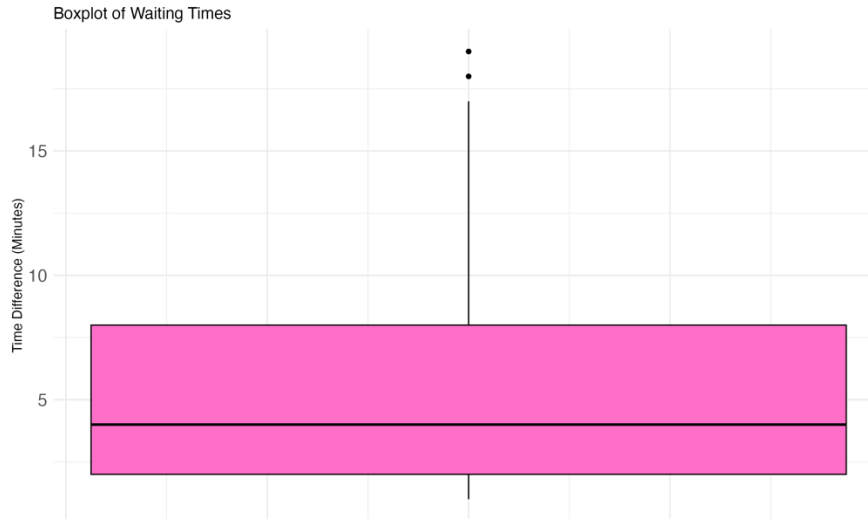
$$[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$$

👉 **Example:** $(x_{min}, Q_1, \tilde{x}, Q_3, x_{max}) = (1.0, 2.0, 3.0, 7.25, 189)$

👉 IQR: $7.25 - 2 = 5.25$

👉 Outlier: $189, 97, 85, 20, 131 \notin [2 - 1.5 \times 5.25, 7.25 + 1.5 \times 5.25]$

BOX PLOT



BOX PLOT DEFINITION

- 👉 Box-Whisker-Plot: A box plot is a graphical representation of the five-number summary:
- A box extends between Q_1 and Q_3 , describing the location of the central 50% of the data.
 - A line within the box marks the median \tilde{x} , representing the center of the distribution.
 - Whiskers extend from the box to the largest and smallest observed values, which are not suspected outliers.
 - Any suspected outliers are shown separately.
-
- The numerical five-number summary or its graphical representation as a box plot is a very suitable way to quickly assess larger datasets of metric characteristics.
 - The first representation of a box plot can be found under the name *range-bar* in **Mary Eleanor Spear's** book *Charting Statistics* [2] from 1952. There, the whiskers extend to the extreme values.
 - The term *box plot* goes back to **John W. Tukey**, who speaks of *box-and-whisker plots* in his book *Exploratory Data Analysis* [3] from 1977. In it, he proposes limiting the length of the whiskers to 1.5 times the interquartile range.



[2] Mary Eleanor Spear: *Charting Statistics*. McGraw-Hill, 1952, S. 164–166.

[3] John W. Tukey: *Exploratory data analysis*. Addison-Wesley, 1977,



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>