



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

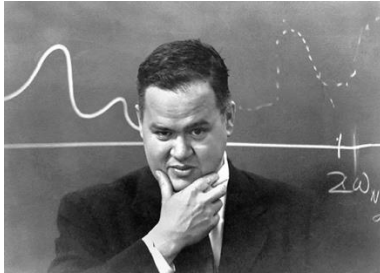
# DESCRIPTIVE STATISTICS

- 👉 Scientific work aims to condense individual pieces of information and observations into **generally valid statements**. Descriptive statistics leads to a **clear and descriptive presentation of information** and descriptive statistics.
- 👉 The aim is to create frequency distributions and graphical representations. The material collected is prepared in such a way that a **quick overview of the distributions of characteristics** found in the sample examined can be obtained. However, **generalized interpretations** of descriptive statistical analyses that go beyond the collected material **are speculative**.
- 👉 **Features or variables** that have certain **values** are **described or displayed**.

# EXPLORATIVE DATA ANALYSIS (EDA)

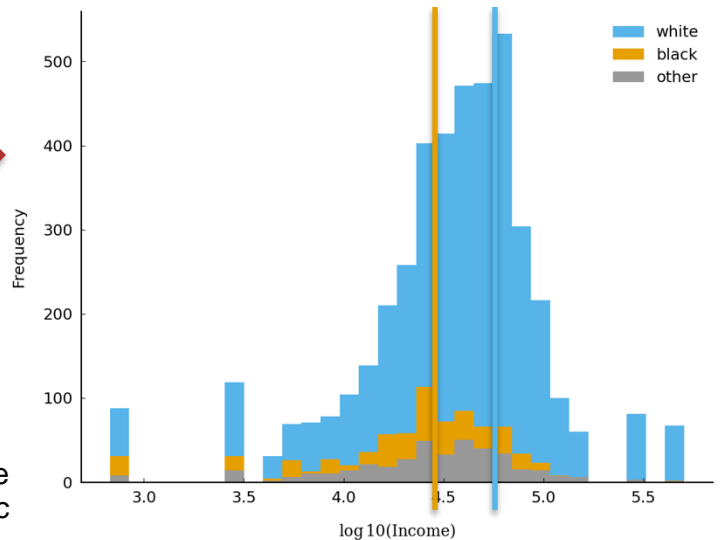
👉 EDA is an approach to analyzing datasets that emphasizes **understanding their main characteristics**, often through **visual methods and summary statistics**. Pioneered by statistician John Tukey, EDA diverges from traditional statistics by **prioritizing the exploration of data to uncover patterns**, identify anomalies, and generate new hypotheses. It's a crucial first step in any data analysis project, helping to **gain insights into the data's structure**, identify potential outliers, and assess the suitability of statistical techniques, **ultimately guiding further analysis and modeling** efforts.

# EXAMPLE FOR EDA



<https://www.amphilsoc.org/item-detail/john-tukey>

This analysis uses data from the 1998 and 2002 General Social Survey (GSS), a **nationally representative US survey**, to explore novel reading habits. In those years the GSS included questions related to culture and the arts. The GSS provides demographic data (**age, sex, race, education, income, region**) alongside **responses about reading fiction** and attending classical music performances. This allows to analyze factors influencing cultural engagement in the US.



<https://www.humanitiesdataanalysis.org/statistics-essentials/notebook.html#summarizing-location-and-dispersion>

# 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*

$$n := |G|$$

## EXAMPLE

- The members of a band build the set:

$$G = \{John, Paul, George, Ringo\}$$

- Size:

$$n = |G| = 4$$

- We can order the set arbitrarily

$$(u_1, u_2, u_3, u_4) = (John, Paul, George, Ringo) \text{ or } (u_1, u_2, u_3, u_4) = (George, Paul, John, Ringo)$$

- The only important thing about the index  $i$  of  $u_i$  is that we clearly know which **unit** is meant.

# 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*



## MEASUREMENT OF FEATURES

- The measurement of a feature  $X$  on a unit  $u \in G$  results in the **observation of a specific value**  $x_u \in X$ , the observed feature value.
- A feature represents a function or **mapping between the population  $G$  and the feature space  $X$** , i.e.,  $X: G \rightarrow X$ , such that each unit  $u \in G$  is assigned a specific value  $X(u) = x_u$ .
- Less abstractly, one could also say that a **feature is a well-defined measurement rule** that assigns a measurement value to a unit.

## BAND EXAMPLE

- Dataset

$$G = \{John, Paul, George, Ringo\}$$

- Gender  $X \in \{1 = male, 2 = female, 3 = diverse\}$

$$X(John) = x_{John} = 1$$

- Ordered dataset

$$(u_1, u_2, u_3, u_4) = (John, Paul, George, Ringo)$$

- Measurement

$$(u_1, u_2, u_3, u_4) \xrightarrow{X} (x_1, x_2, x_3, x_4) = (1, 1, 1, 1)$$

# RAW DATA

- The **measurement of an arbitrarily scaled feature  $X$**  on  $n$  statistical units  $u_1, \dots, u_n$  of a population results in  $n$  observed feature values or measurements  $x_1, \dots, x_n$ , which **are called a univariate raw data list** or simply a raw data list.
- The **simultaneous measurement of other features** in the population then leads to **further raw data lists**, which can best be represented in a so-called **data matrix or data table**.
  - These could also be referred to as a multivariate raw data.
  - Rows represent the respective statistical units.
  - Columns represent the respective features.
  - A cell contains the concretely observed value of the respective feature on the respective statistical unit.

## RAW DATA EXAMPLE

- Let be  $X$  Gender,  $Y$  Age and  $Z$  Height for the (ordered) dataset  $(u_1, u_2, u_3, u_4) = (John, Paul, George, Ringo)$

Unit $i$	Name	$X$	$Y$	$Z$
1	John	1	23	179
2	Paul	1	21	180
3	George	1	21	177
4	Ringo	1	23	168
...	...	...	...	...
$n$	?	$x_n$	$y_n$	$z_n$

# ABSOLUTE (EMPIRICAL) FREQUENCIES

- For each  $a_j$ ,  $f_j$  denotes the **absolute frequency or count** of the value  $a_j$ , i.e., the number of statistical units in  $G$  for which the value  $a_j$  was **observed or measured**. The absolute frequencies  $f_j$ ,  $j = 1, \dots, k$  as a whole are **called the absolute frequency distribution of the feature  $X$**  in the raw data.

$$c_j := |\{u \in G | X(u) = a_j\}|$$

- It holds:

$$c_j \in \mathbb{N}_0, j = 1, \dots, k$$

$$\sum_{j=1}^k c_j = c_1 + c_2 + \dots + c_k = n$$

# FREQUENCY DISTRIBUTION

- An absolute frequency distribution serves to **describe how often each value of a feature  $X$**  was observed in a population  $G$ .
  - Example for gender ( $a_1$  – male,  $a_2$  – female,  $a_3$  – diverse)

$$c_1 = 3, c_2 = 2 \text{ and } c_3 = 1$$

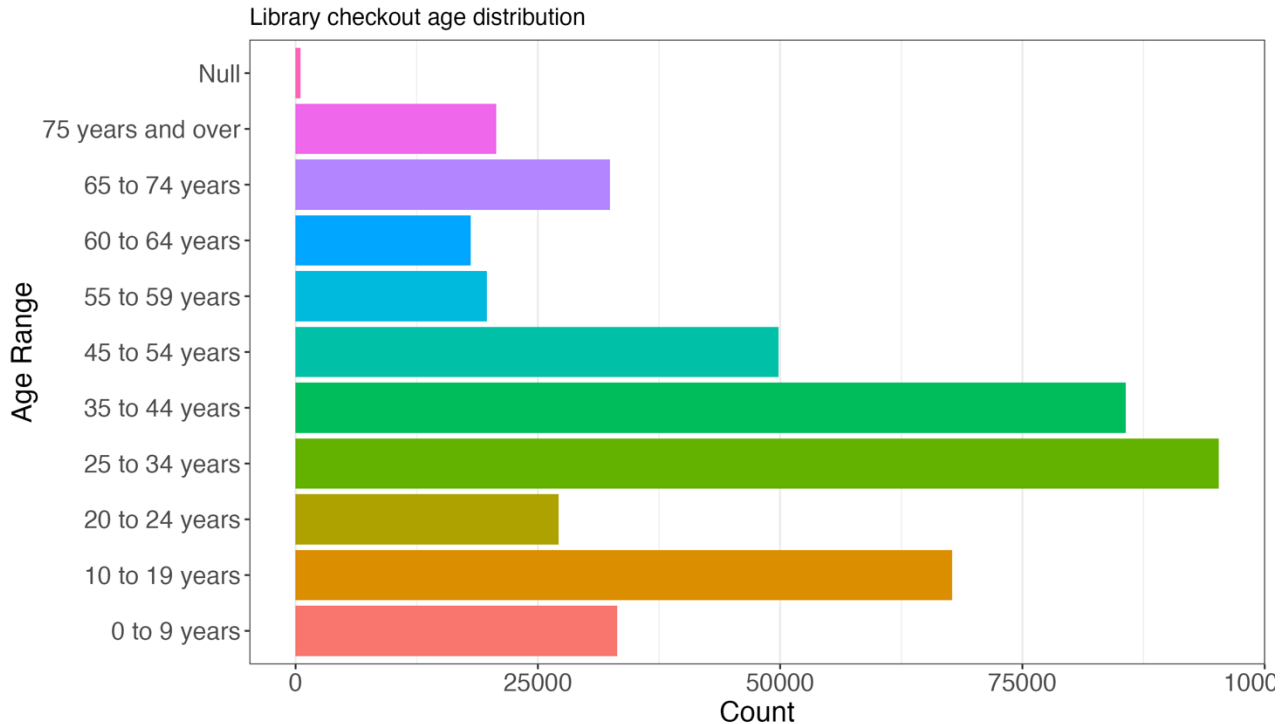
- With few different feature values, the frequencies can be described quite well using a frequency table:

$a_j$	$h_j$
1 (male)	3
2 (female)	2
3 (diverse)	1
$\Sigma$	6

# PLOTTING OF FREQUENCY DISTRIBUTIONS

- An absolute frequency distribution can be suitably represented with a bar chart, stick diagram, or pie chart.
- A bar chart visually represents a frequency distribution using columns:
  - Each column corresponds to a feature value ( $a_j$ ), with its height proportional to the absolute frequency ( $c_j$ ).
  - The columns do not touch, clearly separating categories.
  - For nominal features, the order of feature values is arbitrary.

# PLOTTING OF FREQUENCY DISTRIBUTIONS

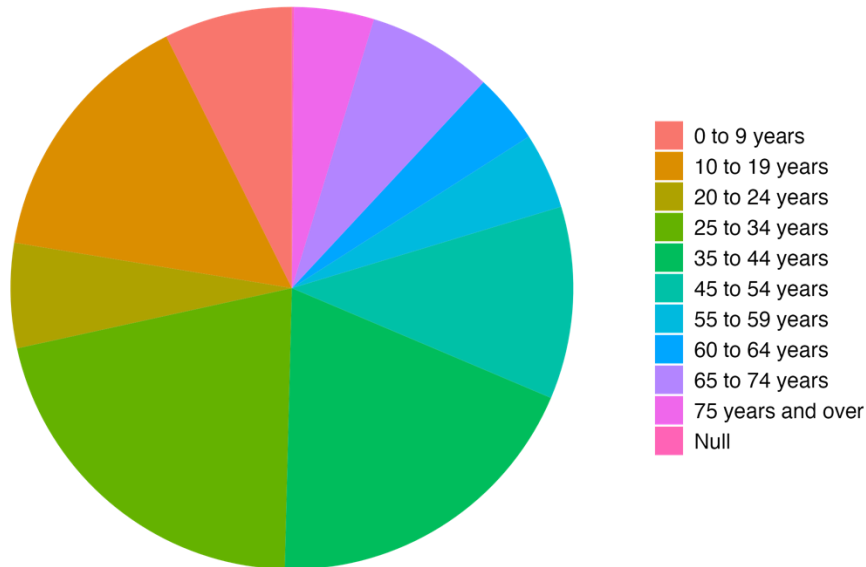


LA library services



# PLOTTING OF FREQUENCY DISTRIBUTIONS


Library checkout age distribution



LA library services

# MODE

- Another important ordering of feature values is the ordering by the magnitude of their frequencies:
  - Choose an arbitrary ordering  $a_1 < \dots < a_k$
  - Determine the corresponding frequencies  $f_1$  from the raw data.
  - Determine the most frequently occurring value  $a_{(1)}$ , the second most frequent value  $a_{(2)}$ , etc.
  - Recode such that:  $a_{(1)} < \dots < a_{(k)}$

 **Mode:** The value  $a_{(1)}$  for which the largest frequency was observed or measured is called **the mode of the frequency distribution**.

## RELATIVE (EMPIRICAL) FREQUENCIES

- Absolute frequency distributions **obscure relative proportions** within a dataset and **hinder comparisons between populations**, particularly in multivariate analyses.
- Example:
  - In a population of 273 kids in the age 10 to 19 years 63 like Romantic Vampire Stories (🧛❤️)
  - In a population of 1517 adults in the age 55 to 59 years there are 75 people who like 🧛❤️
  - In absolute numbers more adults like 🧛❤️
  - $$\frac{63}{273} \approx 0.231 = 23.1\% \text{ and } \frac{75}{1517} \approx 0.049 = 4.9\%$$
  - For every employee who likes 🧛❤️ there are  $23/5 = 4.60$  kids who do the same.

# RELATIVE FREQUENCY DISTRIBUTIONS

- For each  $a_j$ ,  $f_j$  denotes the relative frequency of the value  $a_j$ , i.e., the relative proportion of statistical units in the population for which  $a_j$  is observed. The relative frequencies  $f_j, j = 1, \dots, k$  as a whole are called the relative frequency distribution of the feature  $X$  in the population.
- Formal definition:

$$f_j := \frac{|\{u \in G | X(u) = a_j\}|}{|G|} = \frac{c_j}{n}$$

- It holds:

$$c_j \geq 0, j = 1, \dots, k$$

$$\sum_{j=1}^k c_j = 1$$

## REMARKS

- **Small relative frequencies:** Small relative frequencies like 0,368% of people want to live at the South Pole are not intuitive  $\rightarrow 1/0,00368 = 271,74$ 
  - **One of 272** people want to live at the South Pole
- For a summary of nominal, ordinal or interval scaled data, several categories can be collapsed:
  - *Former GDR = {Mecklenburg – Vorpommern, Brandenburg, Sachsen – Anhalt, Sachsen, Thüringen, Berlin}.*
  - *0 to 9 years = {0,1,2,3,4,5,6,7,8,9}*
    - For ordinal scaled data we must preserve order when grouping.

# HISTOGRAM (PROBLEM DEFINITION)

- Consider a **metric characteristic**  $X$ , i.e. an at least interval-scaled characteristic (e.g. age, income, home distance to library etc.).
- Metric characteristics are always **quantitative and usually continuous**.
- Problem:
  - Either an **uncountably large number of possible characteristic expressions** can occur (continuous) or an **uncountably infinite number** (quasi-continuous or discrete).
  - Example age in years:  $X = \{0, 1, 2, \dots\} = \mathbb{N}_0$
  - Even if we concentrate on the observed expressions, many different expressions can occur.
  - Example age in years  $X_{effective} = \{0, 1, \dots, 969(Methuselah)\}$
  - In the worst case, we, **observe as many different characteristic expressions as there are data points**
- 👉 We observe for **many possible characteristic expressions** either a frequency of 0 or 1 and a **suitable representation in frequency tables is then usually not possible**.

# HISTOGRAM (EXAMPLE)

- Consider the dataset from the San Francisco library
  - It contains 450,359 transactions each showing different amounts of checkouts.
  - A sample from the dataset

[1]	432	49	11	6	0	2	118	15	20	5	0	2	0	1	21
[16]	85	0	0	0	12	320	166	34	0	1	145	46	1	0	0
[31]	<b>1098</b>	11	47	0	<b>2110</b>	6	0	276	15	0	493	9	725	114	240
[46]	0	7	3	272	17	10	0	1851	5	0	0	1	195	0	0
[61]	0	0	0	0	100	96	0	67	<b>1284</b>	608	49	39	12	81	57
[76]	6	447	0	0	0	1	1	5	0	0	222	0	0	503	0
[91]	85	0	51	20	17	47	9	0	74	0					

- Transactions with more than 1000 items are considered outliers since they represent some kind of digital items (17781)

## HISTOGRAM (EXAMPLE CONTINUED)

- Consider the dataset from the San Francisco library
  - In the dataset we have indeed **1000 different values**
  - A direct representation using a frequency table, **bar chart or pie chart is not possible** (1000 categories!), but how can the frequency distribution for metric characteristics still be meaningfully represented or described?
  - In principle, as with nominally or ordinally scaled characteristics, by **condensing the data into classes**

 We call this a **Histogram**



# HISTOGRAM (FORMAL)

- Feature:  $x_1, \dots, x_n$
- Determine a lower bound  $c_0$  and upper bound so that

$$c_0 \leq \min\{x_1, \dots, x_n\} \text{ and } c_k \geq \max\{x_1, \dots, x_n\}$$

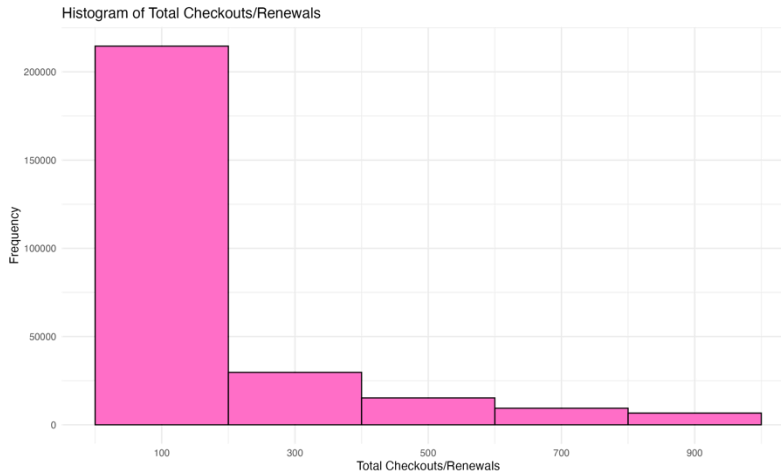
- Determine class boundaries  $c_1, \dots, c_{k-1}$  and form classes (bins):

$$B_1 = [c_0, c_1), B_2 = [c_1, c_2), \dots, B_k = [c_{k-1}, c_k]$$

- For each class  $B_j : d_j = c_j - c_{j-1}$ , the width of the interval and absolute frequency  $h_j$  of the  $j$ -th class.
- Determine the absolute or relative frequency for each class (bin).
- Histogram: Draw a rectangle of width  $d_j$  over each interval  $B_j$  so that the area is proportional to  $h_j$ .
- Note: Sometimes the choice of classes results from what the characteristic  $X$  describes.
- Ideally, all  $d$  are chosen constant and the position of the bar on the x-axis is the center point of the class

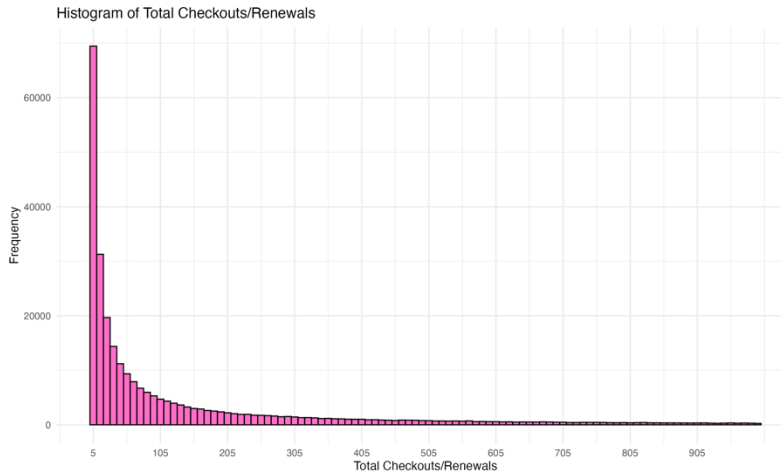
## HISTOGRAM (EXAMPLE CONTINUED)

- We choose  $d = 200$  and  $m_1 = 100$  etc.
- The interval is  $[0, 1000]$  which fall in  $\frac{1000}{200} = 5$  classes  
 $[0, 200)$ ,  $[200, 400)$ ,  $[400, 600)$ ,  $[600, 800)$ ,  $[800, 1000]$



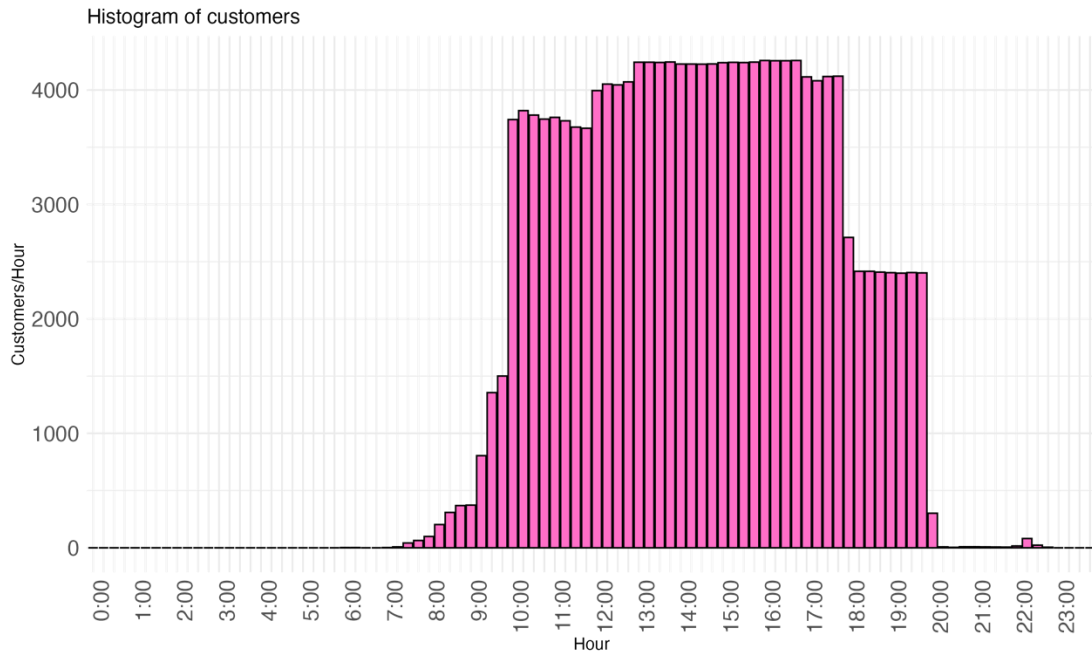
## HISTOGRAM (EXAMPLE CONTINUED)

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



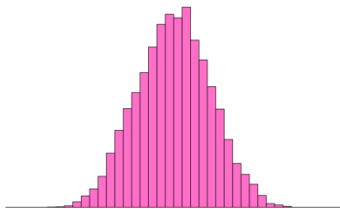
# FROM HISTOGRAMS TO DISTRIBUTIONS

- Consider a Histogram of library checkouts during the day

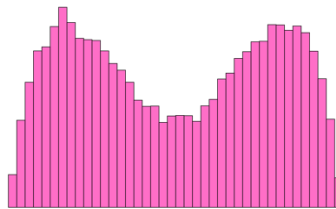


# FORMS OF DISTRIBUTIONS - MODES

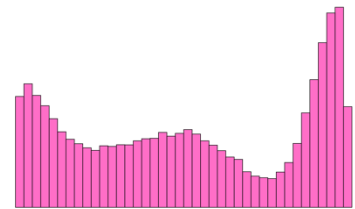
- The shape of a distribution can be assessed, for example, by the **number of modes (peaks)**:



Unimodal = single-peaked



Bimodal = double-peaked



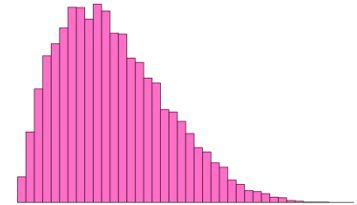
Multimodal = multi-peaked

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

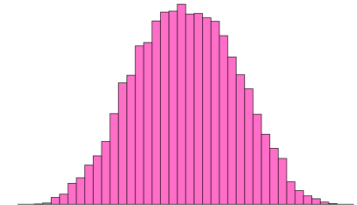
# FORMS OF DISTRIBUTIONS - SYMMETRY

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

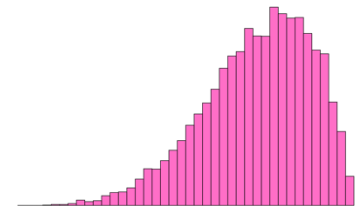
- **left-skewed or right-tailed**: distribution falls much more steeply to the left than to the right



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

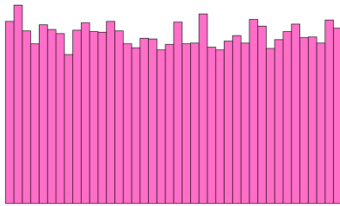


- **right-skewed or left-tailed**: distribution falls much more steeply to the right than to the left

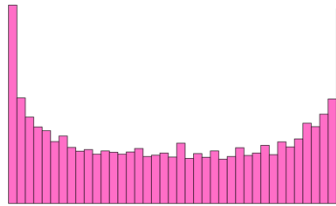


## FORMS OF DISTRIBUTIONS - OTHERS

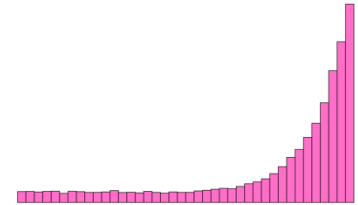
- Other typical forms of distributions are uniform, u-shaped and j-shaped distributions:



Uniform distribution



U-shaped distribution



J-shaped distribution



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](mailto:andreas.niekler@uni-leipzig.de)

<https://www.uni-leipzig.de/personenprofil/mitarbeiter/dr-andreas-niekler>