Frank Klinker

Statistische Grundlagen

Teil 1: Datenreihen und ihre Darstellung, Häufigkeiten

1 Datenreihen und Diagramme

Unter einer **Datenreihe** versteht man eine Menge von Werten. Diese notieren wir typischerweise in der Form

$$x_1, x_2, x_3, \ldots, x_{n-1}, x_n$$

und die Anzahl n der Werte wird auch **Umfang** der Datenreihe genannt.

Solche Datenreihen ergeben sich oft aus einer Messung, z. B. die Temperaturwerte im Verlauf des Tages, die Anzahl der Autos pro Stunde an einer Kreuzung, die Bevölkerungszahlen in verschiedenen Staaten, die Ergebnisse einer Landtagswahl oder die Anzahl des 6er-Wurfs bei einer vorgegebenen Zahl von Würfen.

Eine gängige Art Datenreihen zu notieren ist die **Tabelle**.

Beispiel 1.

1. Präsidentschaftswahl in Frankreich 2017

Kandidat/Kandidatin	Stimmen in %
Marin LePen	21,7
Emmanuel Macron	23,9
Fran cois Fillon	20,0
Benoit Hamon	6,3
Jean-Luc Mélanchon	19,2

2. Entwicklung der Weltbevölkerung

Jahr	1950	1960	1970	1980	1990	2000	2010
Bev.-Zahl in Mrd	2,5	3,1	3,8	4,5	5,4	6,1	6,9

3. Sitzverteilung im 19. Bundestag

Partei	CDU/CSU	SPD	AfD	FDP	Linke	B90/Grüne
Sitze	246	152	89	80	69	67

Adresse: Eduard-Spranger-Berufskolleg, 59067 Hamm
E-Mail: mail@frank-klinker.de
Version: 3. Oktober 2024

4. Monatliche Temperaturen (gemessen in der Monatsmitte)

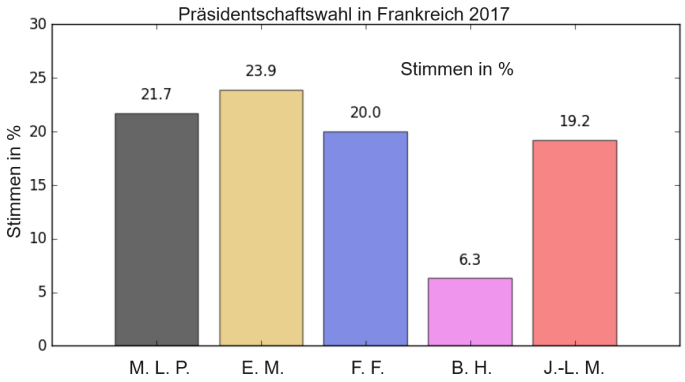
Monat	1	2	3	4	5	6	7	8	9	10	11	12
Temp. Ort 1 in $°C$	7	6,9	9,5	14,5	18,4	21,5	25,2	26,5	23,3	18,3	13,9	9,6
Temp. Ort 1 in $°C$	3,9	4,2	5,7	8,5	11,9	15,2	17	16,6	14,2	10,3	6,6	4,8

Weil Tabellen unübersichtlich sein können, kann man Datenreihen auch mit Hilfe von **Diagrammen** darstellen. Damit lassen sich spezielle Eigenschaften der Datenreihe auf einen ersten Blick erkennen.

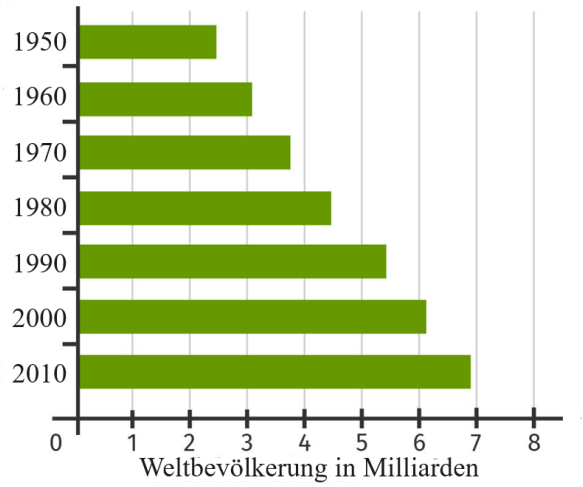
Im folgenden Beispiel sind verschiedenen Arten von Diagrammen passend zu den Tabellen aus Beispiel 1 aufgeführt:

Beispiel 2.

1. **Säulendiagramm**[1]

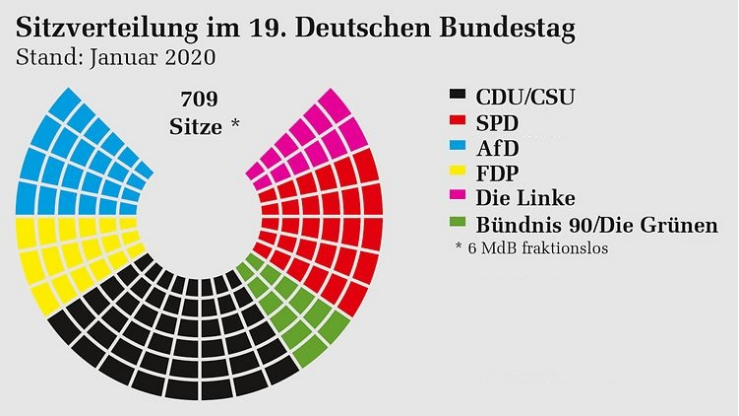


2. **Balkendiagramm**

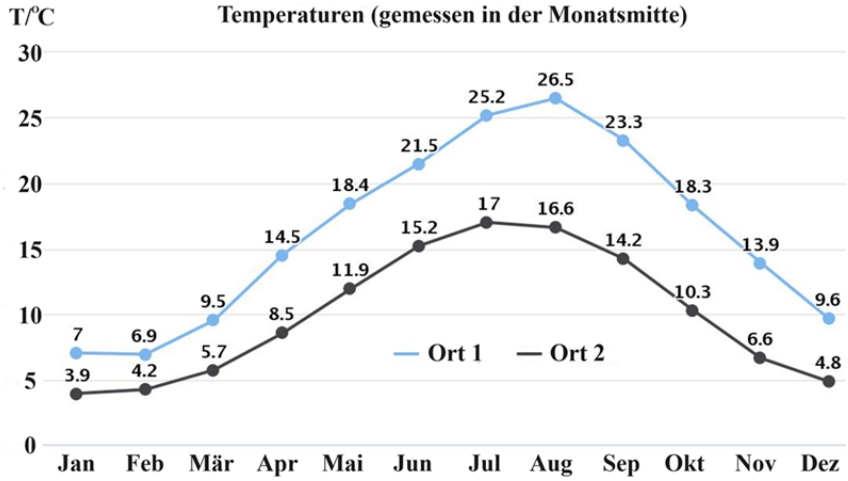


[1] Quelle des Originaldiagramms: https://bodo-schoenfeld.de/jupyter-notebook-balkendiagramm-erstellen/. Leichte Anpassung der Beschriftung.

3. **Tortendiagramm**[2]



4. **Liniendiagramm**



2 Absolute und relative Häufigkeit

Gibt es in einer Datenreihe mit n Werten viele gleiche Werte, dann ist es übersichtlicher nur die unterschiedlichen Werte zu notieren.
Um aber die gesamte Reihe rekonstruieren zu können, ist es wichtig anzugeben, wie oft jeder Wert x_i in der Reihe vorkommt. Diese Zahl nennt man dann **absolute Häufigkeit** H_i:

(Original-)Datenreihe

x_i	10	1	3	1	9	8	9	7	7	9	2	2	3	1	9

Datenreihe und absolute Häufigkeiten

x_i	10	1	3	9	8	7	2
H_i	1	3	2	4	1	2	2

[2]Quelle des Originaldiagramms: https://www.bundestag.de/278118-278118. Leichte Anpassung der Beschriftung.

Wie bereits gesagt, ist die Angabe der absoluten Häufigkeiten unbedingt notwendig: man sieht der zweiten Tabelle nämlich ohne diese Information den Umfang n der (Original-)Datenreihe nicht mehr an.

Ist $\{x_1; x_2; \ldots; x_k\}$ eine Datenreihe mit den absoluten Häufigkeiten $\{H_1; H_2; \ldots; H_k\}$, dann ist $k \leq n$ und es gilt

$$H_1 + H_2 + \ldots + H_k = n\,.$$

Insbesondere ist $0 \leq H_i \leq n$. Hierbei bedeutet $H_i = 0$, dass der entsprechende Datenwert in der Datenreihe gar nicht vorkommt (in bestimmten Situationen kann es sinnvoll sein, solche Werte mit aufzuführen).

Neben der absoluten Häufigkeit gibt die **relative Häufigkeit** h_i an, wie groß der Anteil des Datenwertes x_i an der gesamten Datenreihe ist:

$$h_i = \frac{H_i}{n}\,.$$

Damit ist $0 \leq h_i \leq 1$ und $h_1 + h_2 + \ldots + h_k = 1$:

Datenreihe:	x_i	10	1	3	9	8	7	2	
absolute Häufigkeiten:	H_i	1	3	2	4	1	2	2	Summe = 15
relative Häufigkeiten:	h_i	$\frac{1}{15}$	$\frac{3}{15}$	$\frac{2}{15}$	$\frac{4}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	Summe = 1

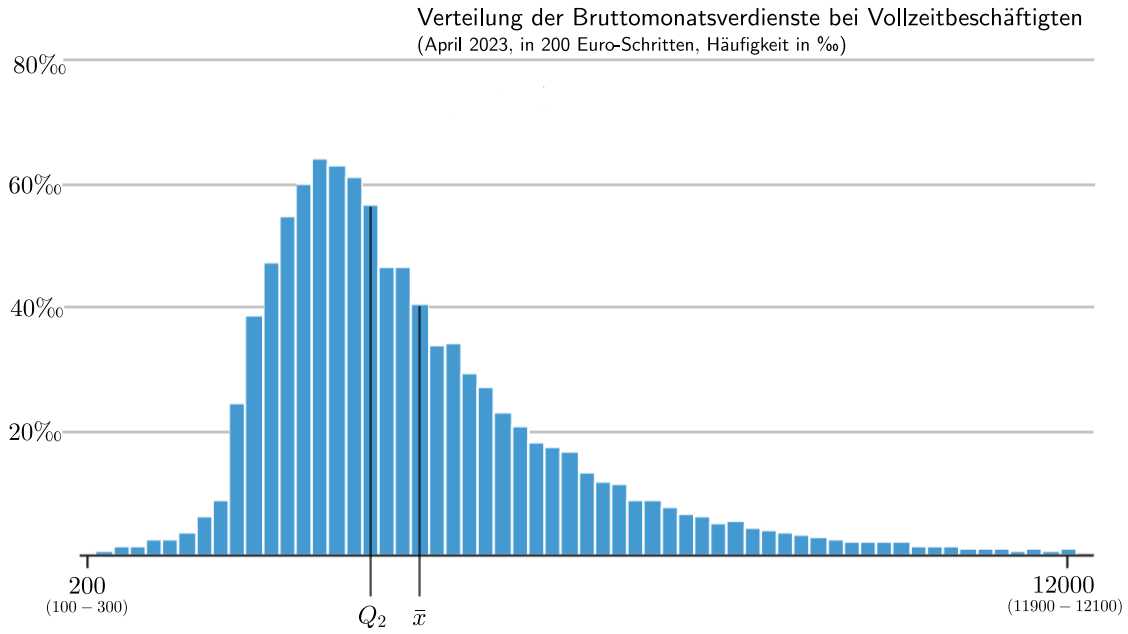
Verabredung:

- Wenn wir $\{x_1; x_2; \ldots; x_{\boldsymbol{n}}\}$ schreiben, dann meinen wir stets eine Datenreihe mit Umfang n.
- Wenn wir $\{x_1; x_2; \ldots; x_{\boldsymbol{k}}\}$ schreiben, dann meinen wir stets eine Datenreihe mit zusammengefassten gleichen Datenwerten.
  Es ist also $k \leq n$ und die Angabe der Häufigkeiten (relativ oder absolut) ist zur vollständigen Beschreibung der Datenreihe notwendig.

Bemerkung 3 (zu relativen Häufigkeiten).

- Wenn man zu den Datenwerten lediglich die relativen Häufigkeiten angibt, dann kann man den echten Umfang der Datenreihe nicht mehr rekonstruieren.
- Einen sinnvollen Wert für den Umfang erhält man dann, indem man einen Faktor sucht, für den alle relativen Häufigkeiten ganzzahlig werden.
  Da die relativen Häufigkeiten oft mit einer festen Zahl an Nachkommastellen angegeben sind, ist so ein Faktor in der Regel eine Zehnerpotenz.
  Dann interpretiert man die Produkte als Häufigkeiten und den Faktor als Umfang.
- Sind die relativen Häufigkeiten in % oder ‰ angegeben, dann ist der Faktor 100 oder 1000 falls keine Nachkommastellen angegeben sind.

3 Beispiel: Einkommensverteilung in Deutschland, April 2023



Die Daten[3] sind in der folgenden Tabelle sortiert:

x_i	200	400	600	800	1000	1200	1400	1600	1800	2000	2200	2400	2600	2800	3000
h_i	0	1	2	1	3	3	4	6	9	25	39	48	55	60	64

x_i	3200	3400	3600	3800	4000	4200	4400	4600	4800	5000	5200	5400	5600	5800	6000
h_i	63	61	57	47	47	41	34	35	31	27	23	21	19	18	17

x_i	6200	6400	6600	6800	7000	7200	7400	7600	7800	8000	8200	8400	8600	8800	9000
h_i	14	13	13	9	9	8	7	7	5	6	5	5	4	4	3

x_i	9200	9400	9600	9800	10000	10200	10400	10600	10800	11000	11200	11400	11600	11800	12000
h_i	3	2	2	2	2	2	2	2	1	1	1	1	1	1	1

Die relative Häufigkeit ist hier in ‰ ohne Nachkommastellen angegeben. Rundungsbedingt ist der Umfang $n = 997$ (statt 1000).

[3] Originalgrafik und Originaldaten: Statistisches Bundesamt (Destatis). Beides wurde leicht an die hier verwendete Notation angepasst.