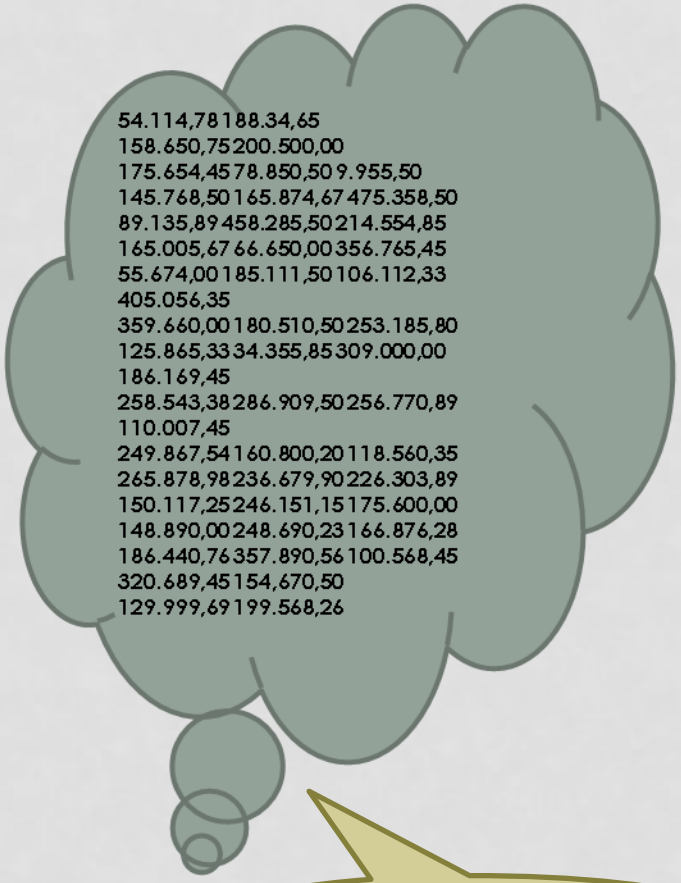
WIRTSCHAFTSSTATISTIK
MODUL 3: HÄUFIGKEITEN UND HÄUFIGKEITSVERTEILUNGEN

WS 2020/21 DR. E. MERINS

DATEN

INPUT



Datenerhebung

OUTPUT

Umsätze der Meyer AG über die Großhändler in NRW im Jahr 2008

Umsatzklasse in Tsd. €	Anzahl Großhändler (absolute Häufigkeit)	Anteil Großhändler von Gesamt in % (relative Häufigkeit)
0 bis unter 100	7	14%
100 bis unter 200	23	46%
200 bis unter 300	12	24%
300 bis unter 400	5	10%
400 bis unter 500	3	6%
Summe	50	100%

(Quelle: Umsatzstatistiken der Vertriebsabteilung, 2008)
Tabelle 1

Datenaufbereitung

DATENDOKUMENTATION

Formen als Dokumentation der Daten:

Einzelwerte (Einzelbeobachtungen) → ungeordnete Reihe (**Urliste, Rohdaten, Primärdaten**) → INPUT-Blase auf der Folie 2

→ Die Urliste ist im Bereich der Statistik das direkte Ergebnis einer Datenerhebung

Vorteile:

Die Urliste enthält alle Beobachtungswerte und damit: keine Auslassungen, keine Übertragungsfehler und keine verlorene Information

Nachteile:

Urlisten können in der Praxis tausende oder Millionen von Datensätze enthalten, die für sich genommen unübersichtlich und nicht auswertbar sind; außerdem können bei einer unkorrigierten Urliste noch offensichtliche Fehler, wie Zahlendreher oder unplausible Daten enthalten sein

HÄUFIGKEITSVERTEILUNGEN

Die Daten einer Urliste müssen in der Praxis also aufbereitet werden, um ihren Zweck zu erfüllen. Das geschieht meist durch das Bilden von Häufigkeitsverteilungen:

Schritt 1a: Sortieren der Daten → **geordnete Reihe**: Reihung nach irgendeiner Ordnung, z. B. alphabetische Ordnung der Merkmalsträger oder die Ordnung nach der Größe der Merkmalsausprägung

Schritt 1b: Verdichten der sortierten Daten auf Merkmalsausprägungen und zählen wie oft diese vorkommen → geordnete Menge von Wertepaaren (Merkmalsausprägung und zugehörige Häufigkeit) heißt **Häufigkeitsverteilung**

Schritt 1c: Darstellen tabellarisch von nach Merkmalsausprägungen sortierten Häufigkeitsverteilungen → die **Häufigkeitstabelle**

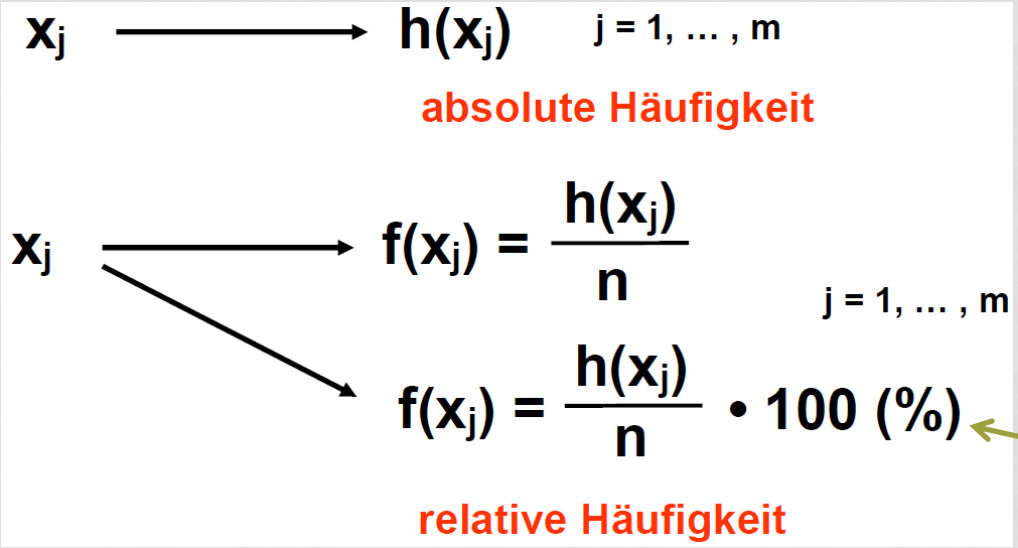
Schritt 2a: Einteilung der Werte in Klassen → **klassierte Daten**

Schritt 2b: Verdichten der klassierten Daten → **Häufigkeitsverteilung** für klassierte Daten (klassierte Verteilung)

Schritt 2c: Darstellen der klassierten Daten → **Häufigkeitstabelle** für klassierte Daten

ABSOLUTE UND RELATIVE HÄUFIGKEITEN

- **Merkmalsausprägung und zugehörige Häufigkeit**



Bezug zur Grundgesamtheit

- **absolute Häufigkeit** → die Anzahl des Auftretens einer bestimmten Merkmalsausprägung
- **relative Häufigkeit** → das Verhältnis der absoluten Häufigkeit und der Summe der Einzelhäufigkeiten

EINDIMENSIONALE HÄUFIGKEITSVERTEILUNG

- **Beispiel 1:**

  n = 20 Personen wurden gefragt nach dem

  Merkmal X: Familienstand

  mit den **j Merkmalsausprägungen**:

  x1 = ledig, x2 = verheiratet, x3 = geschieden, x4 = verwitwet

Primärdaten:

ledig, verheiratet, geschieden, ledig, verheiratet, verwitwet, verheiratet, ledig, verheiratet, verwitwet, verheiratet, ledig, verheiratet, geschieden, ledig, verheiratet, verwitwet, verheiratet, ledig, verheiratet

Was können Sie über die Daten sagen? Charakterisieren Sie die Daten

EINDIMENSIONALE HÄUFIGKEITSVERTEILUNG

- **Beispiel 1:**

  Schritt a: Sortieren

  Schritt b: Verdichten in eine **Häufigkeitsverteilung**

  Schritt c: Darstellen als eine **Häufigkeitstabelle**

j	x_j	Anzahl $h(x_j)$	Anteil $f(x_j)$	Anteil in % $f(x_j)$ (%)
1	ledig	6	0,30	30
2	verheiratet	9	0,45	45
3	geschieden	2	0,10	10
4	verwitwet	3	0,15	15
	Summe	20	1,00	100

EINDIMENSIONALE HÄUFIGKEITSVERTEILUNG

- **Beispiel 2:**

  Frage: Wo wohnen Sie?

  Antworten: B C A B C B B B A A D k.A. A B B A k.A. A B B (*k.A. = keine Antwort*)

→ **Verdichten** in eine **Häufigkeitsverteilung und Darstellen als Häufigkeitstabelle**

i	Wohnort x_i	Anzahl $h(x_i)$	Anteil $f(x_i)$ (%) (bezogen auf alle Antworten)	Anteil $f(x_i)$ (%) (bezogen auf die gültigen Antworten)
1	A	6	30,0%	33,3%
2	B	9	45,0%	50,0%
3	C	2	10,0%	11,1%
4	D	1	5,0%	5,6%
5	k. A.	2	10,0%	-
Summe		20	100,0%	100,0%

SUMMENHÄUFIGKEITEN

→ sinnvoll nur für Rangmerkmale und metrische Merkmale

absolute Summenhäufigkeiten

(absolute kumulierte Häufigkeit)

$H(x_1) = h(x_1)$

$H(x_2) = h(x_1)+h(x_2)$

$H(x_3) = h(x_1)+h(x_2)+h(x_3)$

…

$H(x_j) = h(x_1)+h(x_2)+ \dots +h(x_j)$

…

$H(x_i) = h(x_1)+h(x_2)+ \dots +h(x_i) = n$

relative Summenhäufigkeiten

(relative kumulierte Häufigkeit)

$F(x_1) = f(x_1)$

$F(x_2) = f(x_1)+f(x_2)$

$F(x_3) = f(x_1)+f(x_2)+f(x_3)$

…

$F(x_j) = f(x_1)+f(x_2)+ \dots +f(x_j)$

…

$F(x_i) = f(x_1)+f(x_2)+ \dots +f(x_i) = 1\ (100\%)$

EINDIMENSIONALE HÄUFIGKEITSVERTEILUNG MIT SUMMENHÄUFIGKEITEN

→ sinnvoll nur für Rangmerkmale und metrische Merkmale

Index	Merkmals-ausprägungen	absolute Häufigkeit	relative Häufigkeit	relative Häufigkeit in %	absolute Summenhäufigkeit	relative Summenhäufigkeit	relative Summenhäufigkeit
i	x_i	$h(x_i)$	$f(x_i)$	$f(x_i)$ (%)	$H(x_i)$	$F(x_i)$	$F(x_i)$ (%)
1	a	28	0,11	11,0%	**28**	**0,11**	11,0%
2	b	102	0,402	40,2%	28 + 102 = **130**	0,110 + 0,402 = **0,512**	51,2%
3	c	61	0,24	24,0%	130 + 61 = **191**	0,512 + 0,240 = **0,752**	75,2%
4	d	39	0,154	15,4%	191 + 39 = **230**	0,752 + 0,154 = **0,906**	90,6%
5	e	11	0,043	4,3%	230 + 11 = **241**	0,906 + 0,043 = **0,949**	94,9%
6	f	9	0,035	3,5%	241 + 9 = **250**	0,949 + 0,035 = **0,984**	98,4%
7	h	3	0,012	1,2%	250 + 3 = **253**	0,984 + 0,012 = **0,996**	99,6%
8	g	1	0,004	0,4%	253 + 1 = **254**	0,996 + 0,004 = **1,000**	100,0%
Summe		**254**	**1**	**100%**	**-**	**-**	**-**

EINDIMENSIONALE KLASSIERTE HÄUFIGKEITSVERTEILUNG MIT SUMMENHÄUFIGKEITEN

Klasse Nr.	Klasse	absolute Häufigkeit	relative Häufigkeit in %	absolute Summen-häufigkeit	relative Summen-häufigkeit	Klassenbreite	Klassenmitte
i		h_i	f_i (%)	H_i	F_i (%)	b_i	m_i
1	0 b.u. 20	30	15%	30	15%	20-0=20	(20+0)/2=10
2	20 b.u. 50	60	30%	30+60=90	15+30=45%	50-20=30	(50+20)/2=35
3	50 b.u. 100	80	40%	90+80=170	45+40=85%	100-50=50	(100+50)/2=75
4	100 b.u. 200	30	15%	170+30=**200**	85+15=**100%**	200-100=100	(200+100)/2=150
Summe		**200**	**100%**	-	-	-	-

Klassenbreite b_i:

Die Differenz aus der oberen Klassengrenze und unteren Grenze heißt Klassenbreite **b** der Klasse **i** → $b_i = x_{k-1} - x_k$

Klassenmitte m_i :

Das arithmetische Mittel aus der unteren Klassengrenze und der oberen Klassengrenze heißt Klassenmitte **m** der Klasse **i** → $m_i = 1/2\ (x_{k-1} + x_k\ )$

ZWEIDIMENSIONALE HÄUFIGKEITSVERTEILUNG

Zweidimensionale Häufigkeitsverteilung: G → M Kreuztabelle

Randverteilung: eindim. Häufigkeits-verteilung von M

Randverteilung: eindim. Häufigkeits-verteilung von G

M/G →	w	m	∑
A	400 40,0% 33,3% 20,0%	800 80,0% 66,7% 40,0%	1.200 60%
B	600 60,0% 75,0% 30,0%	200 20,0% 25,0% 10,0%	800 40%
∑	1.000 50%	1.000 50%	2.000 100%

Absolute Häufigkeiten der Merkmals-ausprägungskombinationen

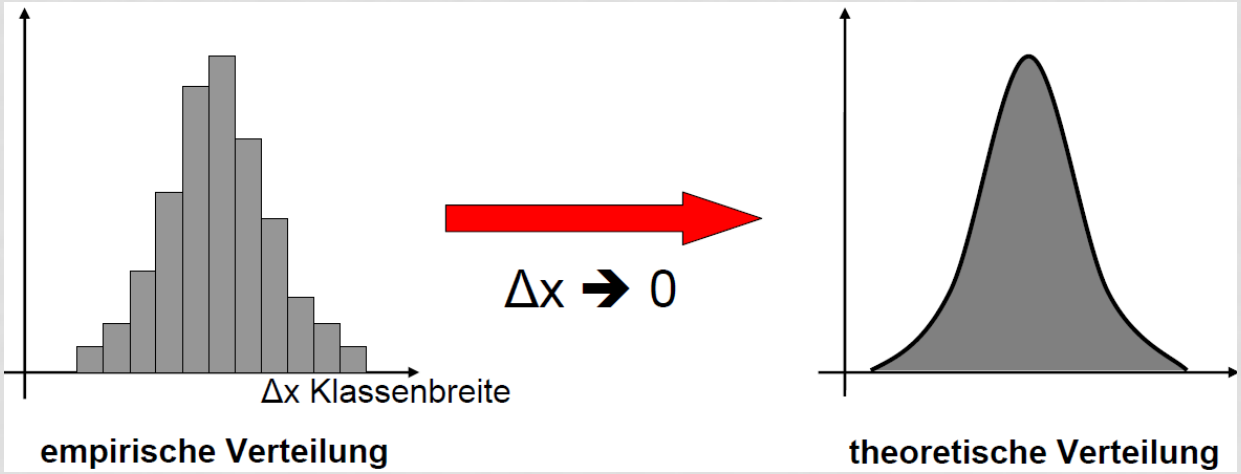
Relative Zeilenhäufigkeiten (bedingte relative Häufigkeiten)

Relative Spaltenhäufigkeiten (bedingte relative Häufigkeiten)

Relative Häufigkeiten der Merkmals-ausprägungskombinationen

EMPIRISCHE VERTEILUNGSFUNKTION

Die Verteilungsfunktion enthält die gesamte Information, die in den Daten steckt, nur die ursprüngliche Reihenfolge geht verloren



theoretische Verteilung ist eine „idealisierte" Verteilung

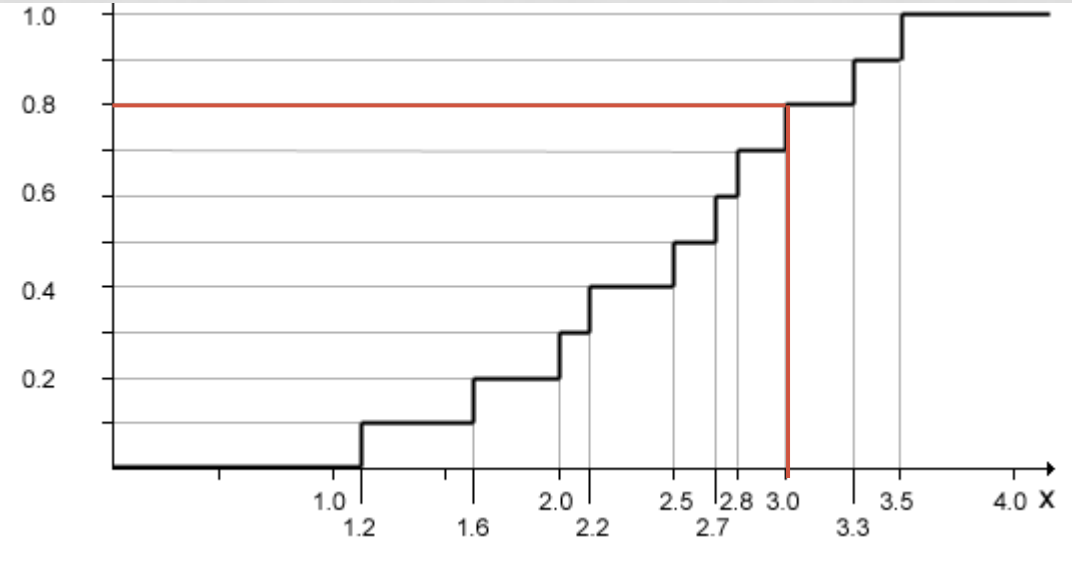
In der mathematischen Statistik klingt das dann in etwa wie folgt:

„Das Maximum der Abweichungen der empirischen Verteilungsfunktion von der theoretisch zugrunde liegenden konvergiert mit Wahrscheinlichkeit Eins gegen Null."

EMPIRISCHE VERTEILUNGSFUNKTION

- **Die empirische Verteilungsfunktion F(x) ist (relative) Summenhäufigkeitskurve, relative Summenfunktion**
- **Die empirische Verteilungsfunktion F(x) gibt für jede beliebige reelle Zahl x den Anteil der Merkmalsträger an, für die das Merkmal X einen Wert x_i annimmt, der kleiner oder gleich x ist**
- **Wertebereich: $0 \leq F(x) \leq 1$**
- **F(x) ist monoton nichtfallend (steigt oder ist konstant)**
- **F(x) ist eine Treppenfunktion mit Sprungstellen bei x_1, x_2, ..., x_i**
- **Die Größe der Sprünge beträgt $f_i = F(x_i) - F(x_{i-1})$**

EMPIRISCHE VERTEILUNGSFUNKTION



Treppenfunktion

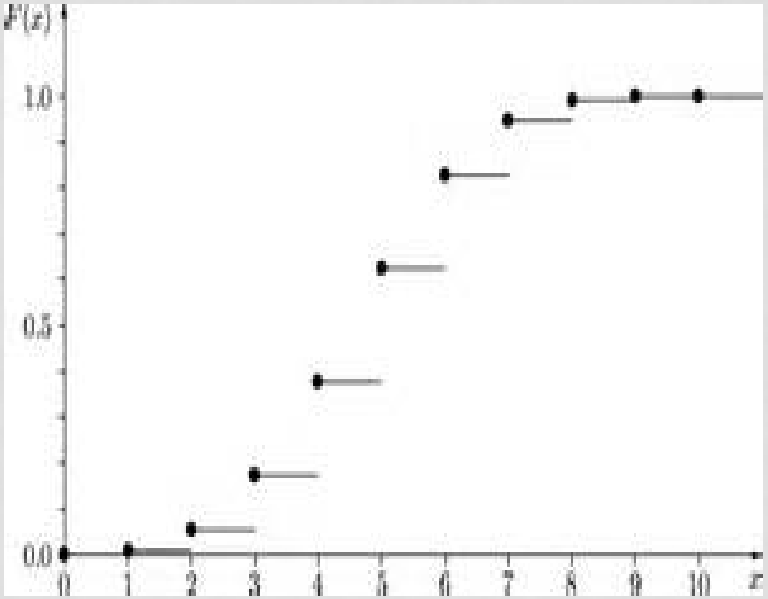
Die Abbildung zeigt die empirische Verteilungsfunktion für das Merkmal Abiturnoten. Greift man auf der x-Achse den Wert 3 heraus, so lässt sich der dazugehörige y-Wert 0.8 wie folgt interpretieren: 80 % der Abiturienten haben im schlechtesten Fall den Notendurchschnitt 3 bekommen.

<u>Anders formuliert:</u>

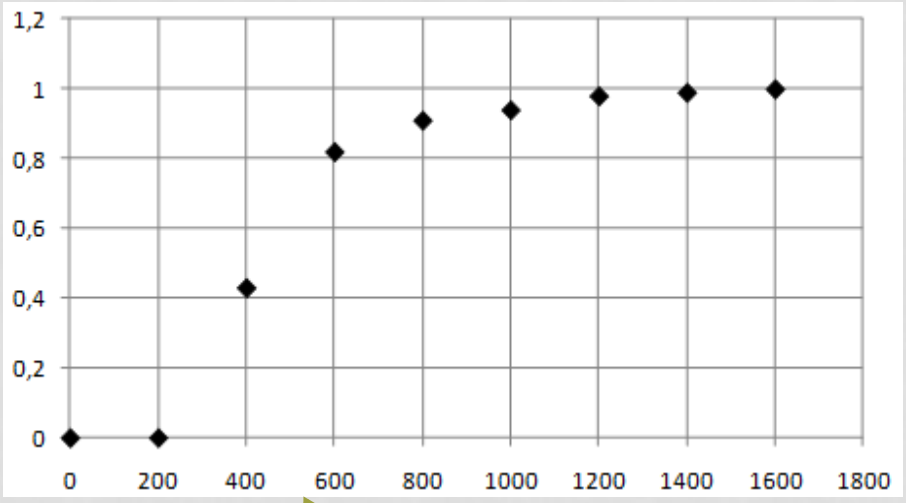
Der Notendurchschnitt ist bei 80 % der Schüler kleiner oder gleich 3.

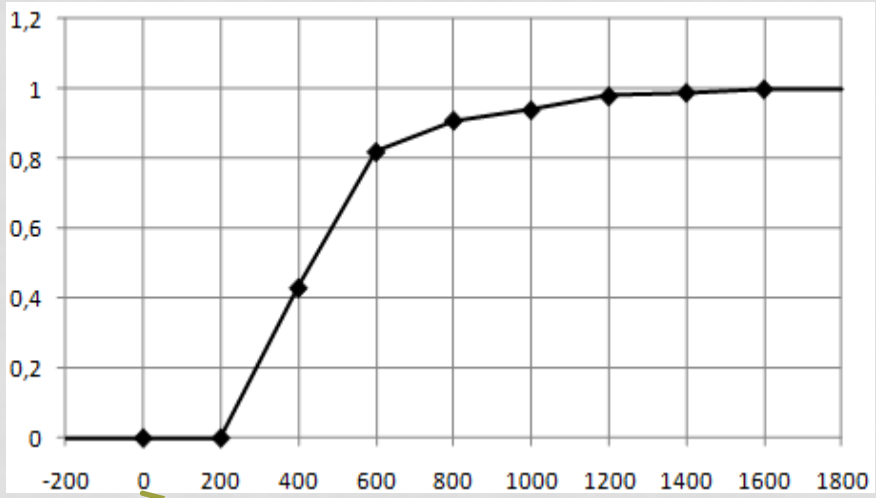
EMPIRISCHE VERTEILUNGSFUNKTION

Beispiel für eine andere Darstellung der Treppenfunktion:



EMPIRISCHE VERTEILUNGSFUNKTION BEI KLASSIERTEN DATEN





Obere Klassengrenze

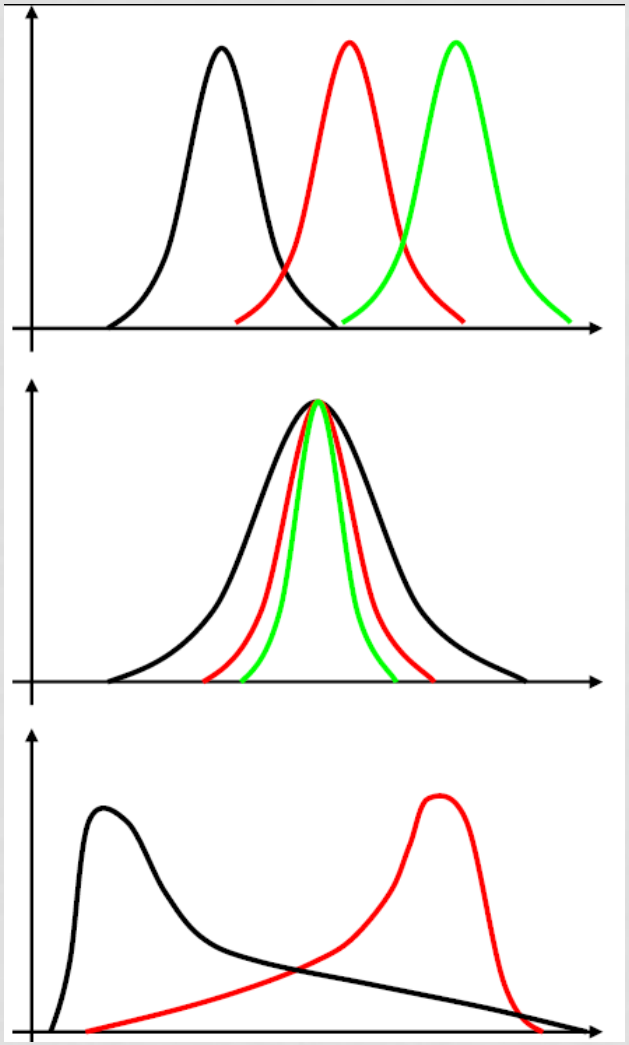
EIGENSCHAFTEN DER HÄUFIGKEITSVERTEILUNGEN

Lage

Streuung

= Wölbung, Form

Schiefe



GRAFISCHE DARSTELLUNG DER HÄUFIGKEITSVERTEILUNG

- **Pro: ein anschauliches Bild der Daten**

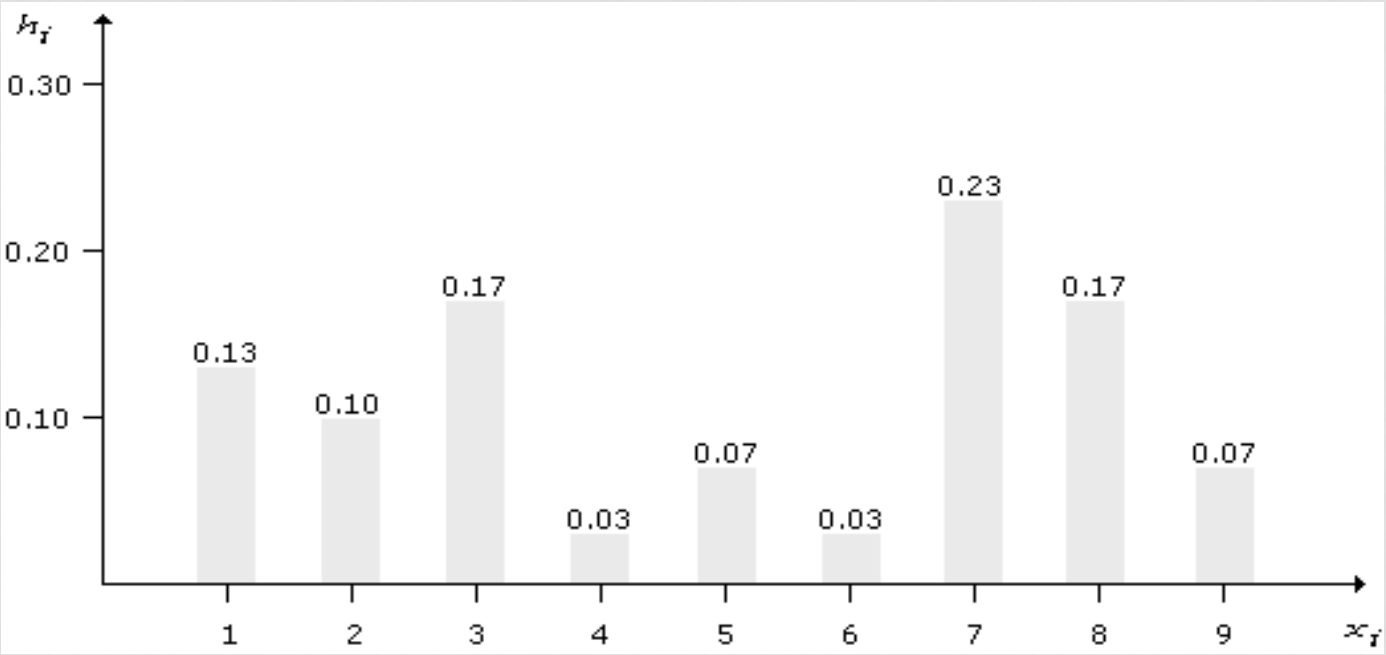
  Ziel: das Wesentliche der Verteilung aufzuzeigen
- **Wahlentscheidung:**
  - **Form der grafischen Darstellung**
  - **Achsenmaßstab**
  - **Evtl. Ausschnitt**

    → *Manipulationen sind denkbar (optische Täuschung!)*
- **Die am weitesten verbreiteten grafischen Darstellungsformen:**
  - **Säulendiagramm**
  - **Stabdiagramm**
  - **Balkendiagramm**
  - **Kreisdiagramm**
  - **Histogramm (bei klassierten Daten)**

SÄULENDIAGRAMM

Säulendiagramm

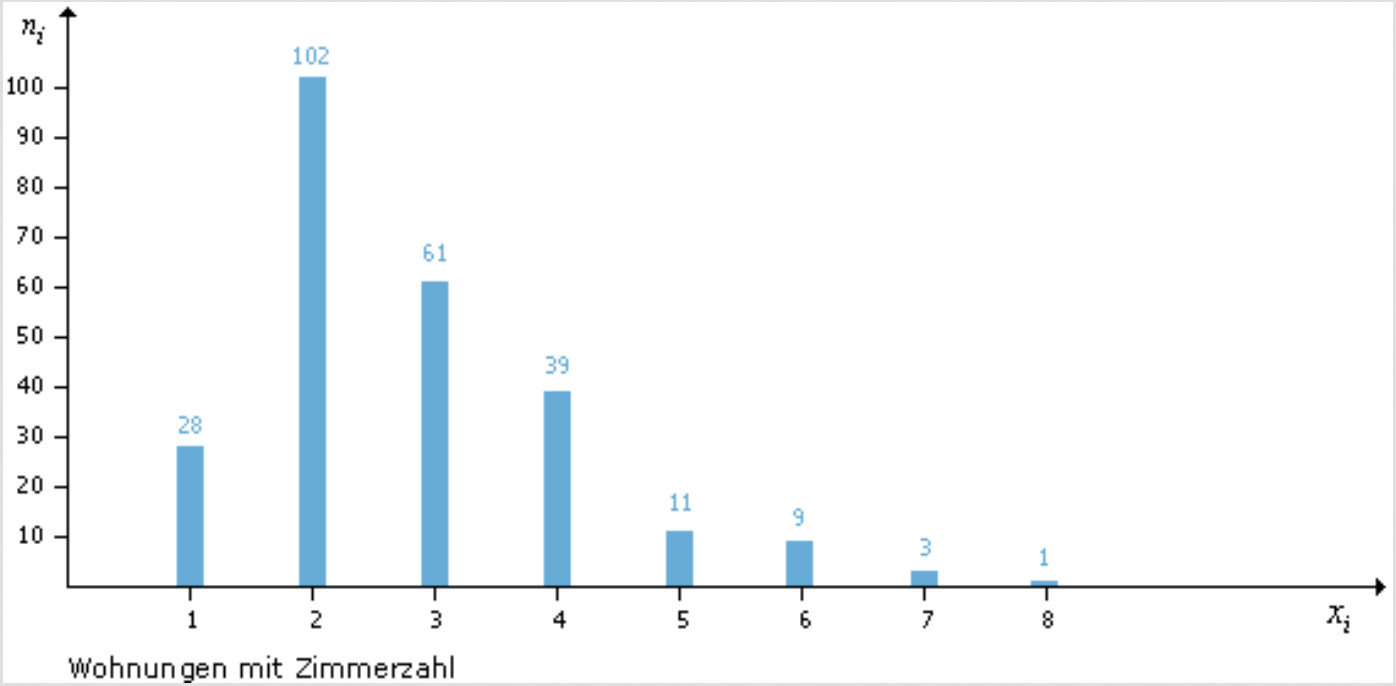
- höhenproportionale Darstellungsform einer Häufigkeitsverteilung durch auf der **x-Achse senkrecht stehende, nicht aneinandergrenzende Säulen** (Rechtecke mit bedeutungsloser Breite)
- eignet sich besonders, um wenige Ausprägungen (bis ca. 15) zu veranschaulichen. Bei mehr Kategorien leidet die Anschaulichkeit und es sind Liniendiagramme zu bevorzugen.



STABDIAGRAMM

Stabdiagramm

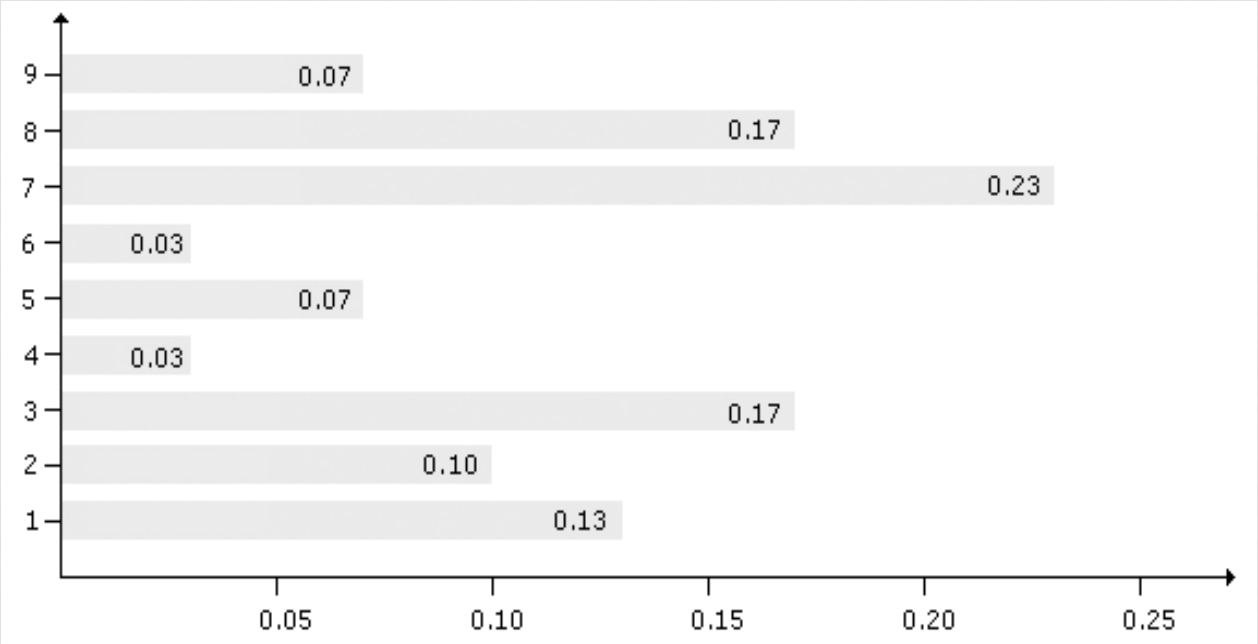
→ Säulendiagramm mit sehr schmalen Säulen



BALKENDIAGRAMM

Balkendiagramm

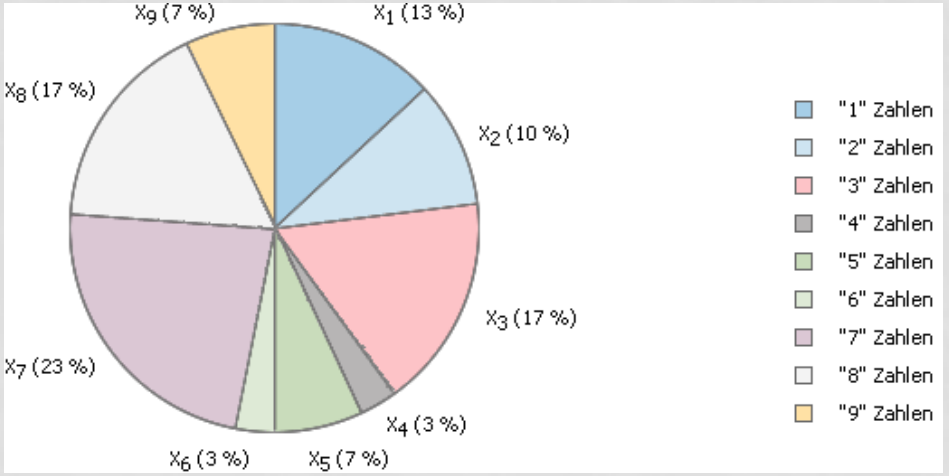
- einer der häufigsten Diagrammtypen
- ist dem Säulendiagramm sehr ähnlich. Unterschied besteht in der Art der Visualisierung: anstatt der vertikalen Säulen sind **horizontale Balken** zu sehen
- eignet sich sehr gut zur Darstellung von Rangfolgen (= Reihenfolge mehrerer vergleichbarer Objekte, deren Sortierung eine Bewertung festlegt, z.B. hier Weltrangliste in Musik)



KREISDIAGRAMM

Kreisdiagramm (Kuchen- oder Tortendiagramm)

- **Kreisförmig, in mehrere Sektoren eingeteilt**, wobei jeder Kreissektor einen Teilwert und der Kreis somit die Summe der Teilwerte (das Ganze) darstellt
- Faustregel: max. 7 Teilwerte, sonst unübersichtlich. Zur besseren Übersichtlichkeit die Teilwerte im Uhrzeigersinn der Größe nach sortieren
- eignet sich zur Darstellung von diskreten Daten, besonders für das Nominal- und das Ordinalskalenniveau zu empfehlen.
- Verwenden wenn:
  - nur eine Datenreihe wird dargestellt
  - keine negativen Werte auftreten
  - keine Nullwerte vorhanden sind
  - die Kategorien Teile des gesamten Kreisdiagramms repräsentieren



HISTOGRAMM

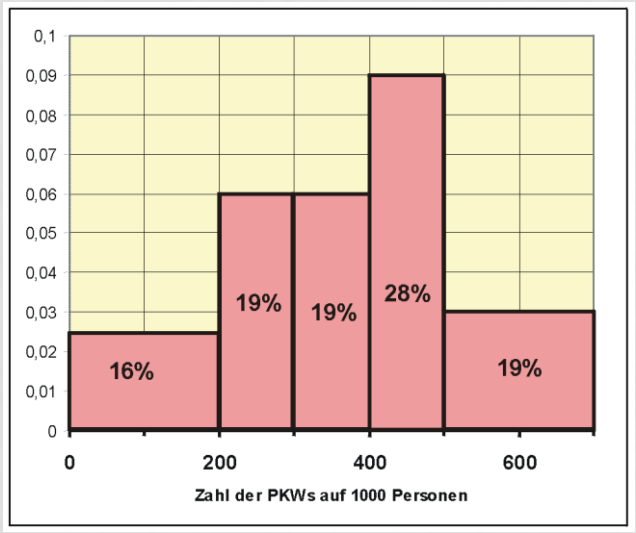
Histogramm

- grafische **flächenproportionale Darstellung der Häufigkeiten von klassierten Daten**
- Im Unterschied zum Säulendiagramm muss bei einem Histogramm die x-Achse immer eine Skala sein, deren Werte geordnet sind und gleiche Abstände haben
- direkt **nebeneinanderliegende Rechtecke** (keine Abstände dazwischen) **von der Breite der jeweiligen Klasse** gezeichnet **(Breite der Rechtecke = Klassenbreite)**
- Absolute oder relative Häufigkeiten der Klassen werden durch die Flächen der Rechtecke dargestellt: **Fläche = Breite x Höhe**
  - Die Breite der Rechtecke entspricht der Breite der Klasse
  - Die Höhe der Rechtecke entspricht den Klassenhäufigkeiten
  - Die Fläche eines Rechtecks = $c \cdot f(x_j)$, wobei $f(x_j)$ die relative Klassenhäufigkeit der Klasse j und c ein Proportionalitätsfaktor ist. Ist c gleich dem Stichprobenumfang (c = n), so ist die Fläche eines jeden Rechtecks gleich der absoluten Klassenhäufigkeit. Das Histogramm wird absolut genannt wenn Summe der Flächeninhalte aller Rechtecke = n. Verwendet das Histogramm die relativen Klassenhäufigkeiten (c = 1), wird das Histogramm relativ oder normiert genannt (Summe der Flächeninhalte aller Rechtecke ist 1).

HISTOGRAMM

Histogramm

Klasse Nr.	Zahl der PKW pro 1.000 Personen	absolute Häufigkeit	Klassenbreite	Rechteckhöhe
i		Anzahl der Länder (h_i)	b_i	$r_i=h_i/b_i$
1	0 b. 200	5	200-0=200	5/200=0,025
2	ü. 200 b. 300	6	300-200=100	6/100=0,06
3	ü. 300 b. 400	6	400-300=100	6/100=0,06
4	ü. 400 b. 500	9	500-400=100	9/100=0,09
5	ü. 500 b. 700	6	700-500=200	6/200=0,03
Summe		**32**	**-**	**-**



HISTOGRAMM

Beispiel:

Vier Histogramme für den gleichen Datensatz: die Klassenbreiten sind in jedem Histogramm gleich 2.0, aber der Beginn der ersten Klasse verschiebt sich von -6,0 über -5,5 und -5,0 auf -4,5.

Fazit: Neben dem Problem der Klassenanzahl bzw. Klassenbreite spielt also auch die Wahl der (linken) Klassengrenzen eine Rolle