Deskriptive Statistik

„Data is merely the raw material of knowledge.“

Ziel: Beschreibung von Daten mit möglichst geringem Informationsverlust

- Eigenschaften und Strukturen sichtbar machen
- Graphisch und durch Kennwerte
- Eindimensional und mehrdimensional
- Zunächst keine Schlüsse auf die Grundgesamtheit oder allgemeine Phänomene

Rohdaten und Datenmatrix

Die Daten liegen in der Regel als Datenmatrix vor:

- Zeilen entsprechen Untersuchungseinheiten
- Spalten entsprechen Merkmalen
- Elemente der Matrix sind die Merkmalsausprägungen
- Fragen mit Mehrfachnennungen als einzelne binäre Merkmale definieren

Hinweise zur Eingabe unter `http://www.stablab.stat.uni-muenchen.de/Datensaetze_mit_Excel`

Beispiel: Befragung von Redakteuren

Bitte füllen Sie diesen Fragebogen nur aus, wenn Sie Chefredakteur bzw. Redaktionsleiter einer Print-Zeitung sind

Sehr geehrter Teilnehmer,
zunächst haben wir einige allgemeine Fragen zur Organisation Ihrer Redaktion:

1. *Die Redaktionen von Print-Zeitungen in Deutschland sind unterschiedlich groß. Wie viele Journalisten (festangestellte und freie) arbeiten in der Stammredaktion Ihrer Print-Tageszeitung?*

_____ festangestellte Redakteure und _____ freie Mitarbeiter.

2. *In jeder Redaktion gibt es verschiedene Positionen zu besetzen. Bitte geben Sie an, welche der folgenden Positionen es in Ihrer Print-Redaktion gibt und wie oft sie gegebenenfalls besetzt sind (also z.B. „2" wenn es zwei Chefs vom Dienst gibt).*

Es gibt.....

___	(Anzahl)	Chefredakteur(e)
	(Anzahl)	Stellvertretende(n) Chefredakteur(e)
___	(Anzahl)	Chef(s) vom Dienst
	(Anzahl)	Ressortleiter
___	(Anzahl)	Leitende(n) Redakteur(e)
___	(Anzahl)	weitere Position und zwar _____

3. *Der Alltag von Journalisten wird durch verschiedene Tätigkeiten bestimmt. Bitte geben Sie an, wie intensiv die Print-Redakteure die folgenden Tätigkeiten im Redaktionsalltag ausüben.*

	täglich	mehrmals pro Woche	einmal pro Woche	mehrmals pro Monat	einmal pro Monat	seltener als einmal pro Monat	nie
Verfassen eigener Artikel	☐	☐	☐	☐	☐	☐	☐
Redigieren von Agenturmeldungen/ Pressemitteilungen	☐	☐	☐	☐	☐	☐	☐
Redigieren von Beiträgen anderer Autoren	☐	☐	☐	☐	☐	☐	☐
Recherche vor Ort	☐	☐	☐	☐	☐	☐	☐
Recherche vom Schreibtisch aus	☐	☐	☐	☐	☐	☐	☐
Bearbeiten von Fotos	☐	☐	☐	☐	☐	☐	☐
Technische Produktion/ Layout der Beiträge	☐	☐	☐	☐	☐	☐	☐

Eindimensionale Häufigkeitsverteilung

- Ordnen der Daten nach einem Merkmal
- Auszählen der Häufigkeiten der einzelnen Merkmalsausprägungen
- Relative Häufigkeiten = Häufigkeit/Anzahl der Untersuchungseinheiten
- Kumulative Häufigkeiten bei ordinal oder metrisch skalierten Merkmalen sinnvoll:
  $F(x)$:=(Summe der relativen Häufigkeiten $\leq$ x)
  empirische Verteilungsfunktion

Häufigkeitsverteilung

Im Weiteren:

$X, Y, \ldots$ Bezeichnung für Merkmal

n Untersuchungseinheiten

$x_1, \ldots, x_i, \ldots, x_n, \quad i = 1, \ldots, n$ beobachtete Werte bzw. Merkmalsausprägungen von X

$\{ x_1, \ldots, x_i, \ldots, x_n; \quad i = 1, \ldots, n\}$ Rohdaten, Urliste

Häufigkeiten I

$a_1 < a_2 < \ldots < a_k$, $k \leq n$ der Größe nach geordnete, *verschiedene* Werte der Urliste $x_1, \ldots, x_n$

Beispiel: Absolventenstudie
Für die Variable D "Ausrichtung der Diplomarbeit" ist die Urliste durch die folgende Tabelle gegeben.

Person i	1	2	3	4	5	6	7	8	9	10	11	12
Variable D	3	4	4	3	4	1	3	4	3	4	4	3

Person i	13	14	15	16	17	18	19	20	21	22	23	24
Variable D	2	3	4	3	4	4	2	3	4	3	4	2

Person i	25	26	27	28	29	30	31	32	33	34	35	36
Variable D	4	4	3	4	3	3	4	2	1	4	4	4

Häufigkeiten II

Ausprägung	absolute Häufigkeit h	relative Häufigkeit f
1	2	$2/36 = 0.056$
2	4	$4/36 = 0.111$
3	12	$12/36 = 0.333$
4	18	$18/36 = 0.500$

Häufigkeitstabelle für die Variable D „Ausrichtung der Diplomarbeit“

Bemerkungen:

- Für Nominalskalen hat die Anordnung „$<$“ keine inhaltliche Bedeutung.
- Bei kategorialen Merkmalen $\Rightarrow$ k = Anzahl der Kategorien
  Bei stetigen Merkmalen $\Rightarrow$ k oft nicht oder kaum kleiner als n.

Absolute und relative Häufigkeiten

$h(a_j) = h_j$ *absolute Häufigkeit* der Ausprägung a_j,

d.h. Anzahl der x_i aus $x_1, \ldots x_n$ mit $x_i = a_j$

$f(a_j) = f_j = h_j/n$ *relative Häufigkeit* von a_j

$h_1, \ldots, h_k$ *absolute Häufigkeitsverteilung*

$f_1, \ldots, f_k$ *relative Häufigkeitsverteilung*

Bemerkungen:

- Wenn statt der Urliste bereits die Ausprägungen $a_1, \ldots, a_k$ und die Häufigkeiten $f_1, \ldots, f_k$ bzw. $h_1, \ldots, h_k$ vorliegen, sprechen wir von *Häufigkeitsdaten*.
- Klassenbildung, gruppierte Daten:
  Bei metrischen, stetigen (oder quasi-stetigen) Merkmalen oft Gruppierung der Urliste durch Bildung geeigneter Klassen

Beispiel Nettomieten I

Wir greifen aus dem gesamten Datensatz die Wohnungen ohne zentrale Warmwasserversorgung (zh=1) und mit einer Wohnfläche kleiner als 50 qm (wfl<50) heraus. Die folgende Urliste zeigt, bereits der Größe nach geordnet, die Nettomieten dieser $n = 27$ Wohnungen:

81.28	98.85	109.32	130.35	132.24	151.00	163.41
172.23	181.98	183.09	195.72	203.75	224.61	229.06
244.50	268.36	272.24	275.52	314.09	352.79	353.69
357.05	373.37	388.81	389.23	412.61	463.40	

Alle Werte verschieden
$\Rightarrow k = n$ und $\{x_1, \ldots, x_n\} = \{a_1, \ldots, a_k\}$
$\Rightarrow f_j = \frac{1}{27}, \quad j = 1, \ldots, 27.$

Beispiel Nettomieten II

Selektion dieser Daten in R:

```
daten <- read.table(file="miete03.asc", sep="\t", header=T)
daten <- subset(daten, (zh==1) \& (wfl<50) )
attach(daten)
print(sort(nm))
```

(nm=Nettomiete)

Beispiel Nettomieten III

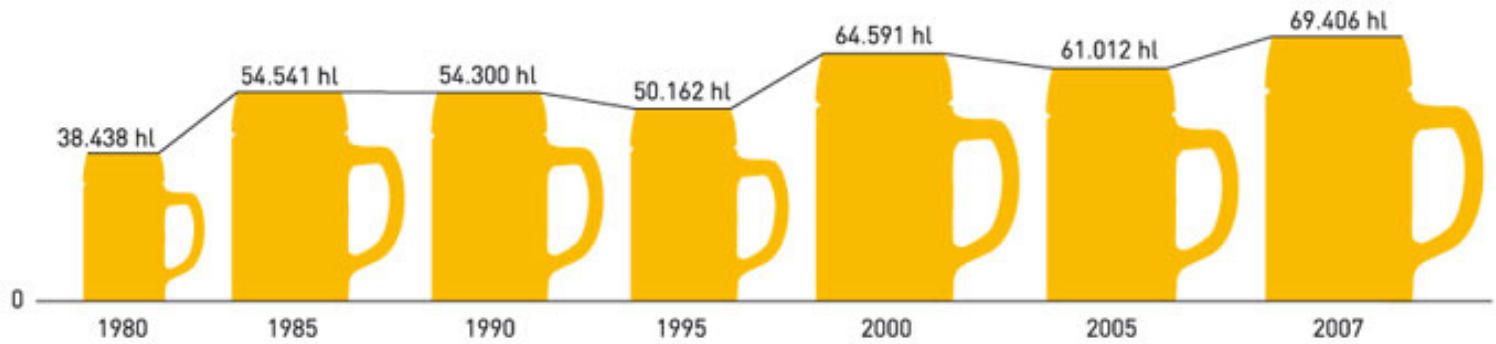
Gruppiert man die Urliste in 5 Klassen mit gleicher Klassenbreite von 100 EURO, so erhält man folgende Häufigkeitstabelle:

Klasse	absolute Häufigkeit	relative Häufigkeit
$50 < \ldots \leq 150$	5	$5/27 = 0.185$
$150 < \ldots \leq 250$	10	$10/27 = 0.370$
$250 < \ldots \leq 350$	4	$4/27 = 0.148$
$350 < \ldots \leq 450$	7	$7/27 = 0.259$
$450 < \ldots \leq 550$	1	$1/27 = 0.037$

Häufigkeiten für gruppierte $n = 27$ Nettomieten

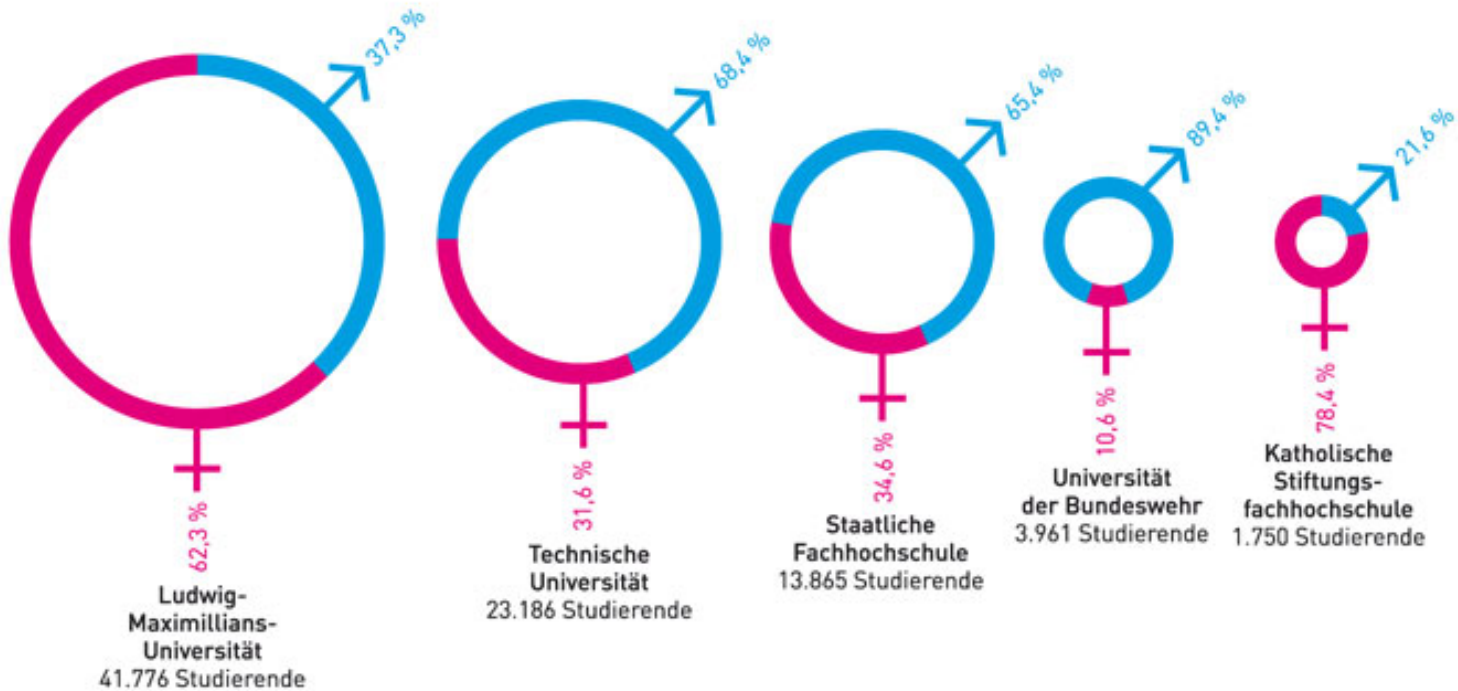
Grafische Darstellungen

„Ein Bild sagt mehr als tausend Worte“



Grafische Darstellungen

„Ein Bild sagt mehr als tausend Worte“

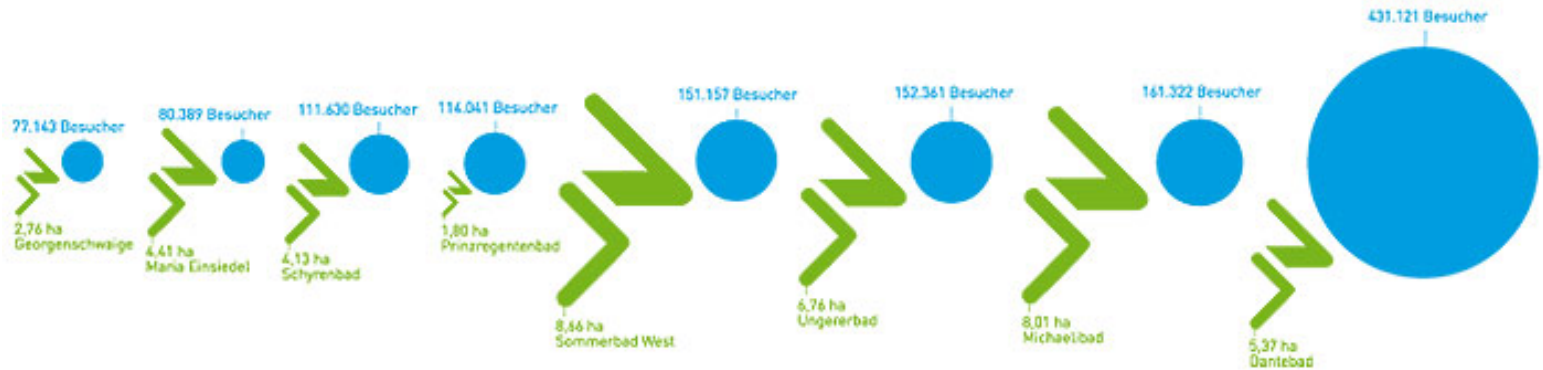


Grafische Darstellungen

„Ein Bild sagt mehr als tausend Worte“

Grafische Darstellungen

„Ein Bild sagt mehr als tausend Worte“



Statistische Grafik

Lit.: Tufte, E. (2001): The visual Display of Information.
Graphic Press 2nd ed.

Principles of Graphical Excellence

- Graphical excellence is the well-designed presentation of interesting data - a matter of *substance*, of *statistics* and of *design*.
- Graphical excellence consists of complex ideas communicated with clarity, precision and efficiency.
- Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.
- Graphical excellence is nearly always multivariate.
- And graphical excellence requires telling the truth about the data.

Allgemeine Kriterien

- Wahl der Skala inkl. Bereich
- Wahl des Prinzips (Längentreue, Flächentreue)
- Einbringen von anderen Visualisierungen (Piktogramme etc.)
- Angemessene Wahl der Variablen
- Angemessene Wahl der Farben

Wahrnehmung von Grafiken

Experimente von Psychologen zeigen Hierarchie der korrekten Interpretation (Cleveland/McGill)

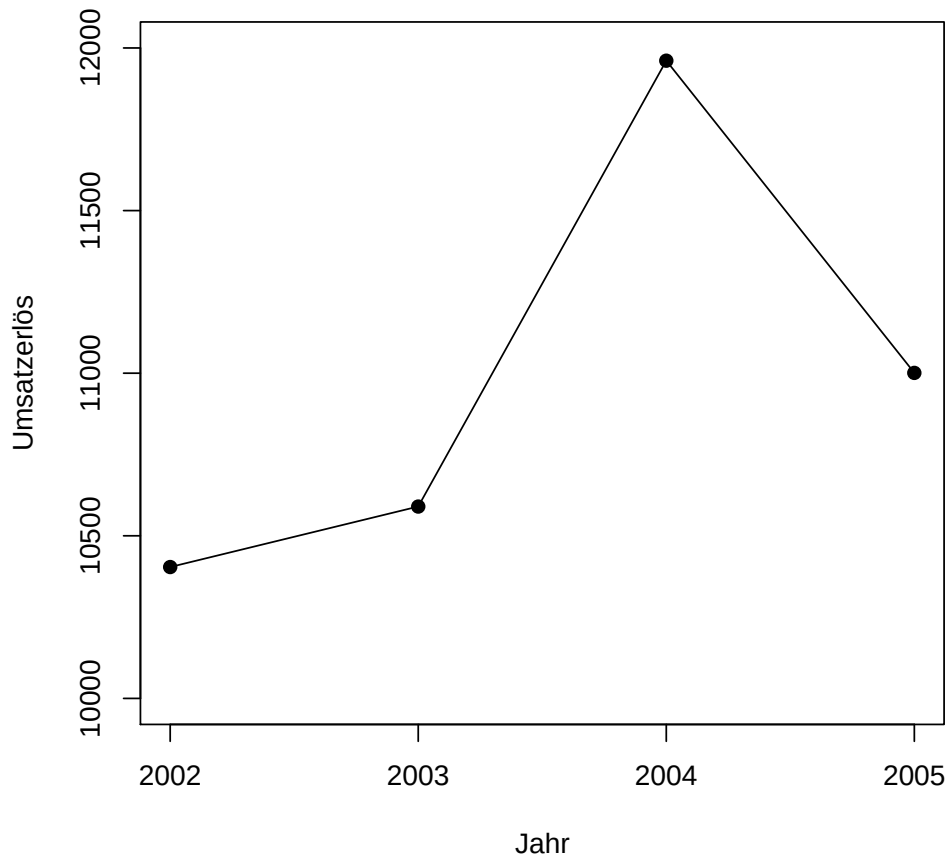
1. Abstände
2. Winkel
3. Flächen
4. Volumen
5. Farbton-Sattheit-Schwärzegrad

Da Abstände am besten wahrgenommen werden, sollten diese bevorzugt verwendet werden.

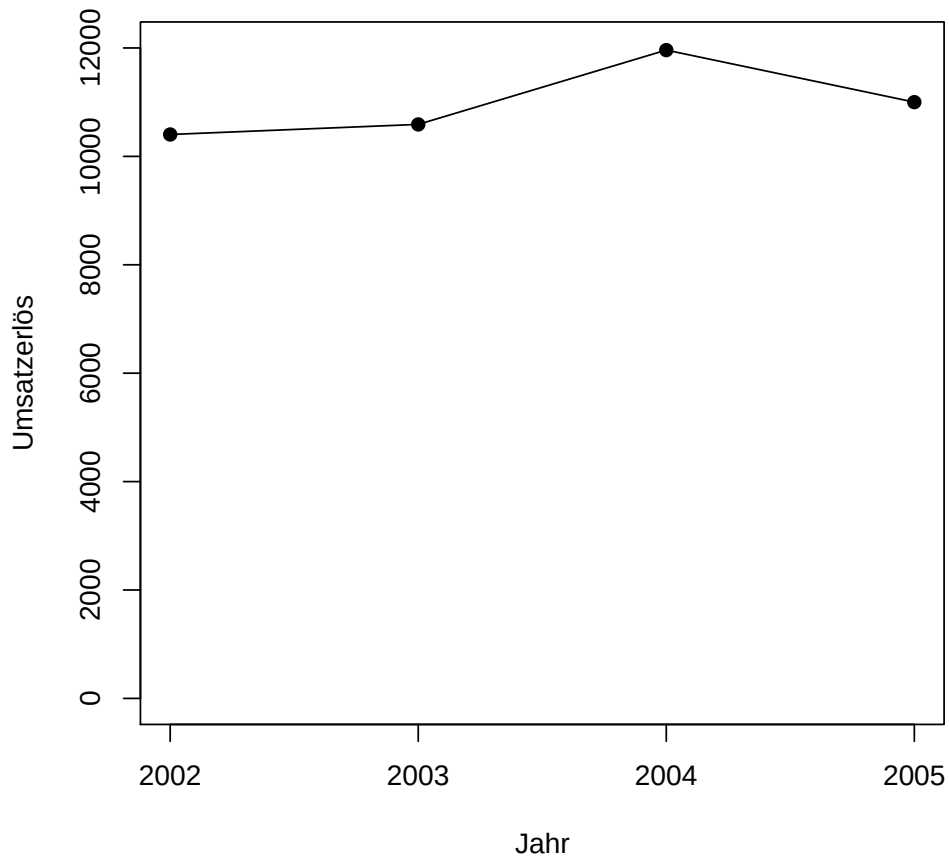
Typen von eindimensionalen Darstellungen

- Stab-, Balken- und Säulendiagramm
- Kreis (Torten)-Diagramm
- Histogramm

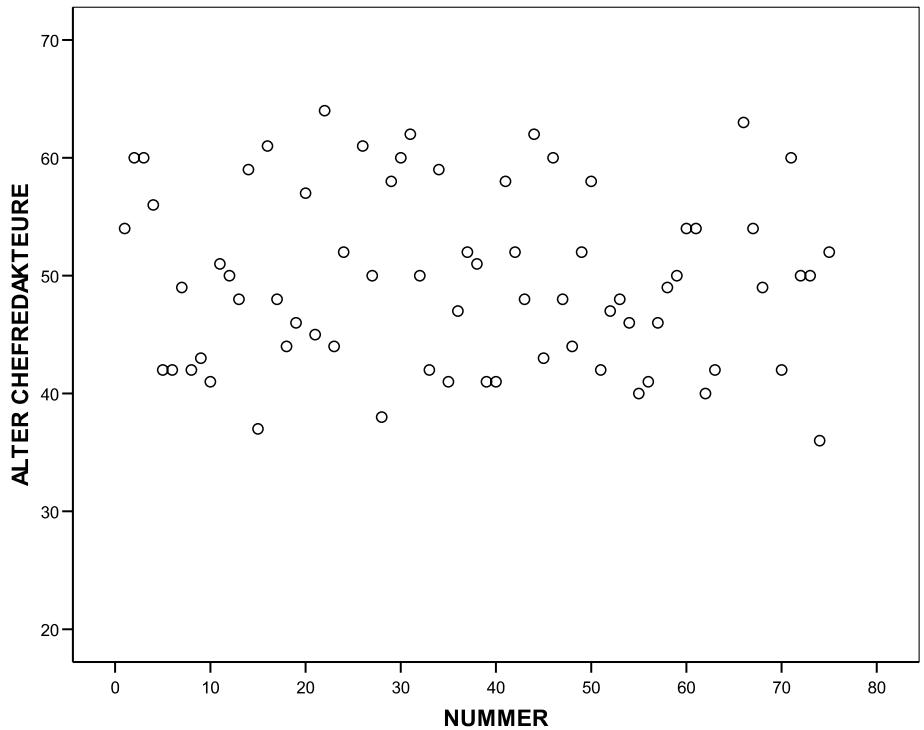
Beispiel: Liniendiagramm (??)



Beispiel: Liniendiagramm (!!)



Beispiel: Streudiagramm



Beispiel: Needleplot