LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
INSTITUT FÜR KOMMUNIKATIONSWISSENSCHAFT UND MEDIENFORSCHUNG

BA KW | Vorlesung

Einführung in die Statistik

Häufigkeiten

Prof. Thomas Hanitzsch

Urliste und Häufigkeitsverteilung

- Urliste:
  - die Gesamtheit aller erhobenen Daten.
- Rangwertreihe:
  - eine „sortierte Urliste“
    alle erhobenen Daten werden nach Größe sortiert.
- Häufigkeitsverteilung:
  - gibt für jeden Merkmalswert die Häufigkeit an, mit der dieser in den erhobenen Daten vorkommt

Die Urliste

Untitled1* × data ×

Filter

	OFBesuche Wie oft haben Sie – dieses Jahr eingenommen – das Oktoberfest b...	Tracht Haben Sie dabei mindestens einmal auch Tracht getragen?	MassBier Wie viele Maß Bier haben Sie bei einem durchschnittlichen Wiesen...	OFgut Ich finde das Oktoberfest gut	OFsauf Für meinen Geschmack wird zu viel getrunken	OFlaut Das Oktoberfest
1	69	2	1	4	2	0
2	50	1	2	3	3	2
3	50	1	3	2	3	1
4	50	1	4	3	1	1
5	40	1	1	4	3	2
6	40	1	3	4	1	0
7	35	1	2	4	4	2
8	35	1	3	4	0	2
9	35	1	1	3	2	2
10	28	1	1	4	4	2
11	25	1	3	4	3	3
12	25	1	3	3	3	1
13	25	1	5	1	1	1
14	20	1	1	3	4	3
15	20	1	2	3	3	1
16	20	1	1	4	2	1
17	20	*NA*	1	4	3	3
18	20	1	4	3	2	1
19	18	1	2	3	3	1
20	17	1	1	1	3	4
21	17	1	1	1	4	4
22	15	1	2	3	3	3
23	15	1	1	2	3	4
24	15	1	2	3	2	0
25	15	1	2	4	3	1
26	15	2	3	3	1	0
27	10	1	1	4	3	1
28	10	1	0	1	4	4
29	10	1	1	4	4	1
30	10	1	3	3	2	0

Showing 1 to 31 of 126 entries, 23 total columns

Die Urliste

Untitled1* data

Filter

	OFBesuche Wie oft haben Sie – dieses Jahr eingenommen – das Oktoberfest b...	Tracht Haben Sie dabei mindestens einmal auch Tracht getragen?	MassBier Wie viele Maß Bier haben Sie bei einem durchschnittlichen Wiesen...	OFgut Ich finde das
1	69	2	1	4
2	50	1	2	3
3	50	1	3	2
4	50	1	4	3
5	40	1	1	4
6	40	1	3	4
7	35	1	2	4
8	35	1	3	4
9	35	1	1	3
10	28	1	1	4
11	25	1	3	4
12	25	1	3	3
13	25	1	5	1
14	20	1	1	3
15	20	1	2	3
16	20	1	1	4
17	20	*NA*	1	4
18	20	1	4	3
19	18	1	2	3

Darstellung von Häufigkeiten

- Urliste und Rangwertreihe sehr unübersichtlich
- Je größer der Datensatz, desto schwieriger wird es, sich in der Urliste zurechtzufinden
- Lösung 1:
  - tabellarische Darstellung
- Lösung 2:
  - grafische Darstellung

Einfache Häufigkeitsverteilungen

- Stellt die absoluten und relativen Häufigkeiten eines Merkmals dar
- Absolute Häufigkeit (f_i):
  - Wie oft tritt eine Ausprägung i des Merkmals X in einer Stichprobe mit dem Umfang N auf?
- Relative Häufigkeit (p_i):
  - Wie oft tritt eine Ausprägung i des Merkmals X im Verhältnis zur Gesamtstichprobe mit dem Umfang N auf?

Einfache Häufigkeitsverteilungen

- Häufigkeitstabelle:
  - Zuordnung der absoluten und relativen Häufigkeiten zu den Ausprägungen des Merkmals X
- Relative Summenhäufigkeit ($F_i = \sum p_i$):
  - Für mindestens ordinalskalierte Daten
  - Prozentualer Anteil derjenigen Merkmalsträger, die kleiner oder höchstens gleich einem bestimmten Wert x_i eines Merkmals X sind

Die empirische Verteilungsfunktion

- Die empirische Verteilungsfunktion $F(X)$ der Datenliste ist der Anteil der Daten, die kleiner oder gleich x_i sind:

$$F(X) = f(\vec{x} \leq x_i)$$

Kumulierte Häufigkeiten

Beispiel: Mass Bier bei Oktoberfestbesuch

Mass Bier (x_i)	Absolute Häufigkeit (f_i)	Relative Häufigkeit (p_i)	Relative Summen-häufigkeit (F_i)
0	15	13,8%	13,8%
1	36	33,0%	46,8%
2	29	26,6%	73,4%
3	18	16,5%	89,9%
4	9	8,3%	98,2%
5	1	0,9%	99,1%
6	1	0,9%	100%

Häufigkeitstabelle: Klassenbildung

- Zur Erinnerung: stetige Daten können unendlich viele Ausprägungen annehmen
- Eine Häufigkeitstabelle wäre dann keine echte Verbesserung zur Rangwertreihe
- Lösung:
  - Klassenbildung bzw. gruppierte Häufigkeitstabelle

Häufigkeitstabelle: Klassenbildung

- Beispiel: tägliche TV-Nutzung (in Minuten)

x_i	f_i	p_i	F_i
0	14	11,2	11,2
2	2	1,6	12,8
10	4	3,2	16,0
15	4	3,2	19,2
20	9	7,2	26,4
30	11	8,8	35,2
40	3	2,4	37,6
45	9	7,2	44,8
50	2	1,6	46,4
60	25	20,0	66,4
70	1	,8	67,2
…	…	…	…
300	2	1,6	100

Klasse	Minuten
1	0 … 60
2	> 60 … 120
3	> 120 … 180
4	> 180

Häufigkeitstabellen: Klassenbildung

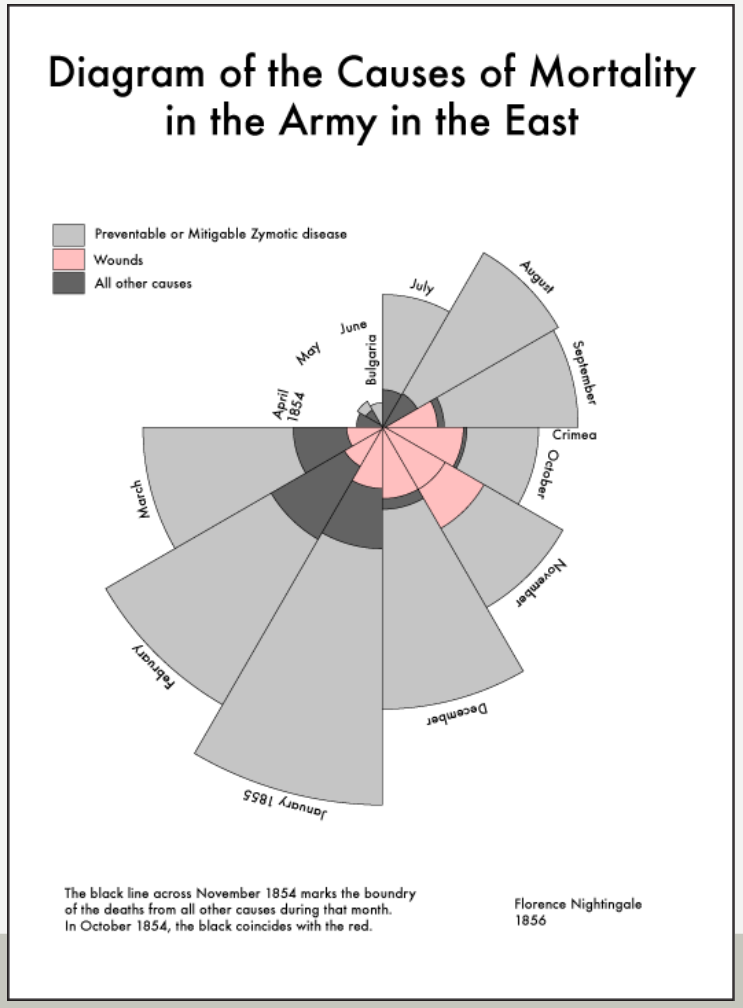
- Prinzip der Überschneidungsfreiheit:
  - Jede Ausprägung muss genau einer Klasse zugeordnet werden können
- Anzahl der Klassen:
  - Je breiter eine Klasse, desto größer der Informationsverlust
  - Je mehr Klassen, desto geringer die Übersicht
  - Anzahl auch abhängig von Untersuchungsziel und Verwendungszweck
- Klassenbreite:
  - Sollte bei allen Klassen möglichst gleich sein (Ausnahme: stark variierende Daten)
- Ausreißer:
  - Bei vielen oder starken Ausreißern „offene Randgruppen“ bilden („weniger als…“ bzw. „mehr als…“)

Warum Grafiken: Florence Nightingale

- 1820-1910
- Sammelte während des Krimkrieges Daten über Todesursachen der Soldaten
- niemand schenkte ihren endlosen Tabellen Beachtung
- Lösung: Darstellung der Ergebnisse mit einer Grafik

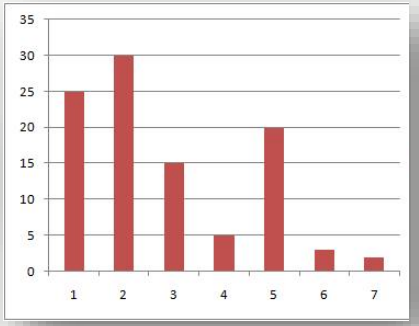
Warum Grafiken: Florence Nightingale

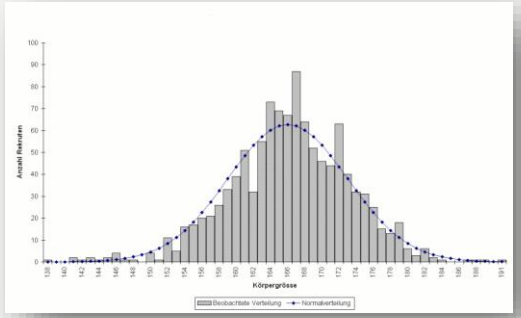
- ihre Grafiken überzeugten die Regierung von einer grundlegenden Reform des Sanitätssystems.

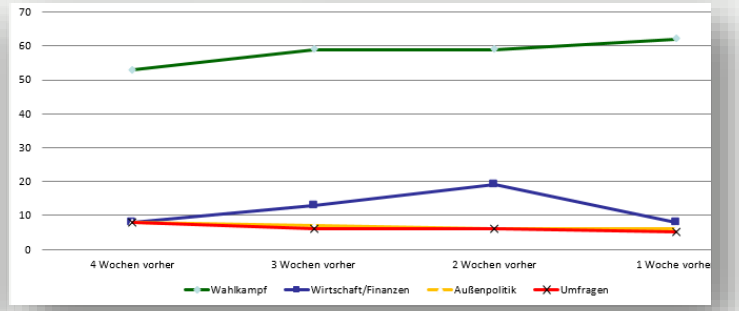


Grafische Darstellungsformen

- Darstellungsform hängt vom Skalenniveau des Merkmals ab
- Nominale und ordinale Merkmale
  - Kreisdiagramm
  - Säulen- / Balkendiagramm
- Metrische Merkmale
  - Histogramm
  - Liniendiagramm

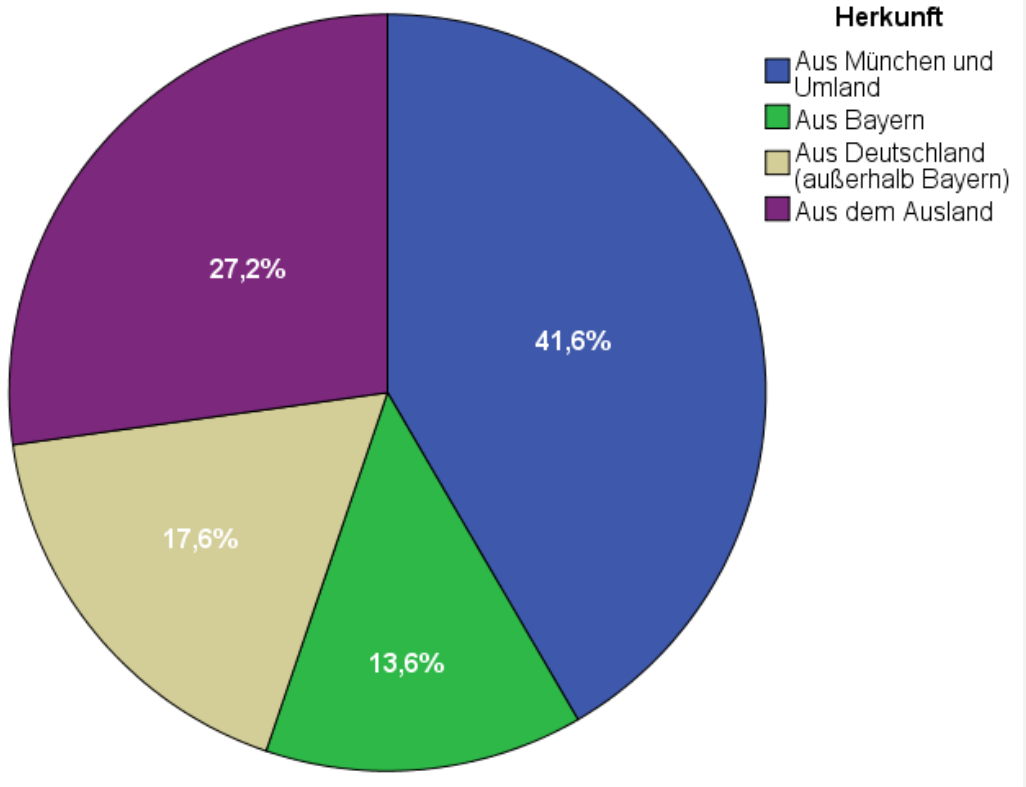






Diagrammtypen: Kreisdiagramm

- Werte für die einzelnen Merkmalsausprägungen werden Kreisanteilsflächen („Tortenstücke“) repräsentiert
- Ideal für nominalskalierte Daten
- Absolute Häufigkeiten, relative Häufigkeiten oder Prozente
- Auch: „Tortendiagramm“



Diagrammtypen: Balken-/Säulendiagramm

- Werte für die einzelnen Merkmalsausprägungen werden durch Säulen/Balken repräsentiert
- Können für nominal- und ordinalskalierte Daten verwendet werden
- Bei ordinalen Variablen muss die Reihenfolge der Kategorien in der Grafik erhalten bleiben
- Absolute Häufigkeiten, relative Häufigkeiten oder Prozente

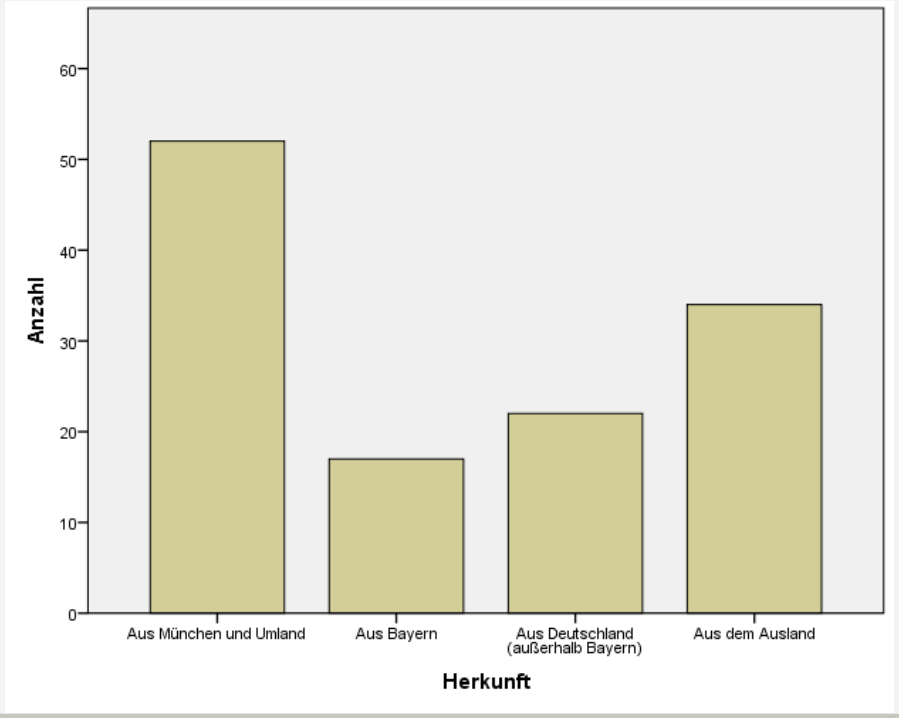
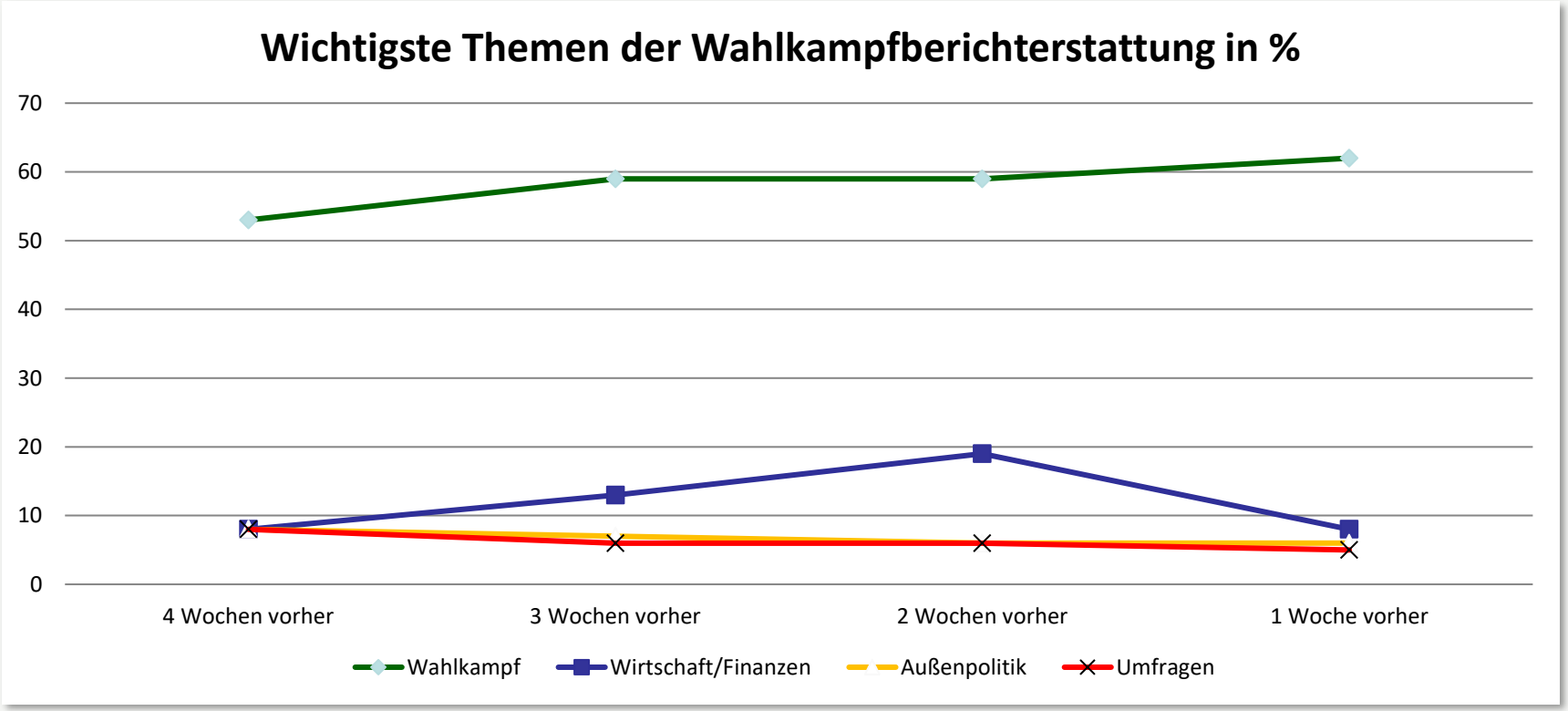


Diagramm-Typen: Liniendiagramm

- Geeignet für:
  - Visualisierung einer großen Zahl zeitabhängiger Daten
  - das generelle Verhalten (Trend) und nicht das absolute Ausmaß
  - Vergleich verschiedener Beobachtungsreihen miteinander
  - Sichtbarmachen von Prognosen und Interpolationen (Zwischenwerte)

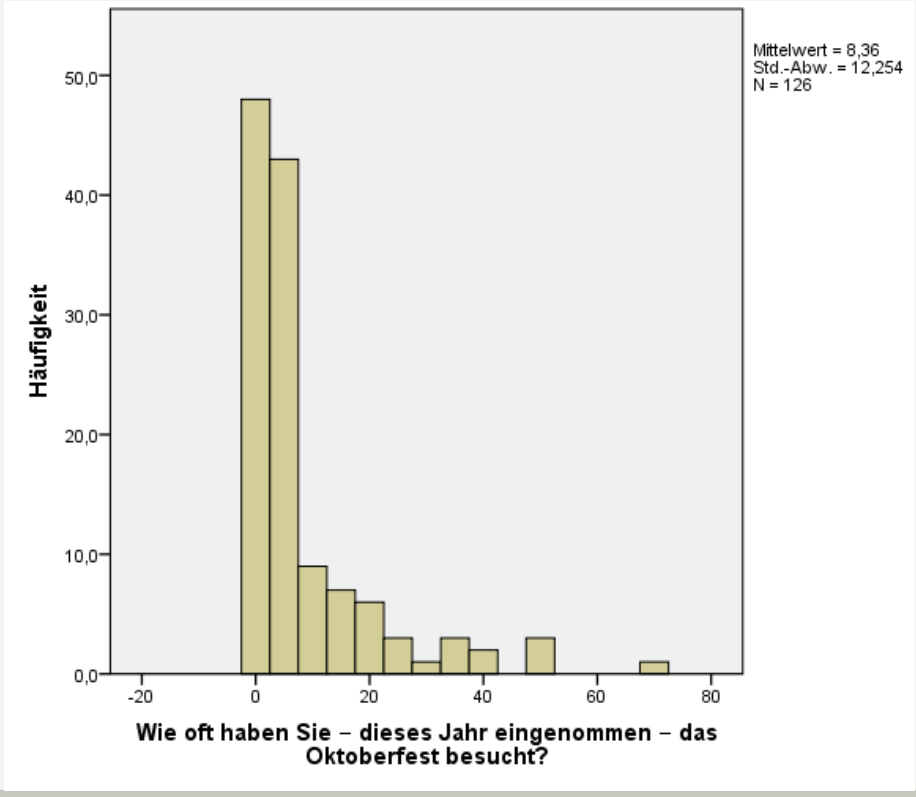
Diagramm-Typen: Liniendiagramm



Basis 567 Beiträge
Quelle: Wilke und Reinemann 2006, S. 318

Diagramm-Typen: Histogramm

- der „grafische Bruder“ der gruppierten Häufigkeitstabelle
- Zeigt Verteilung von Ausprägungen eines metrischen Merkmals auf Teilintervalle des Merkmalsbereichs
- Breite der Säulen ist proportional zur Klassenbreite
- Nur sinnvoll bei metrischen Daten



Wichtige Lage- und Streuungsparameter

- Lagemaße
  - Median
  - Modus
  - Arithmetisches Mittel (Mittelwert)
- Streuungsmaße
  - Spannweite
  - Varianz
  - Standardabweichung
  - Varianzkoeffizient

Lagemaße: Modus (Modalwert) $\overline{x}_{mod}$

- Merkmalswert mit der höchsten absoluten Häufigkeit
- Eher sinnvoll bei gruppierten Daten oder bei Merkmalen mit wenigen Ausprägungen
- Stabil gegenüber eindeutigen Transformationen
- Geeignet für alle Skalenniveaus

Lagemaße: Modus (Modalwert) $\overline{x}_{mod}$

- Beispiel: Mass Bier beim Oktoberfest

x_i	f_i	p_i	F_i
0	15	13,8%	13,8%
1	36	33,0%	46,8%
2	29	26,6%	73,4%
3	18	16,5%	89,9%
4	9	8,3%	98,2%
5	1	0,9%	99,1%
6	1	0,9%	100%

Lagemaße: Median $\bar{x}_{med}$

$$\bar{x}_{med} = \begin{cases} x_{\left(\frac{N+1}{2}\right)} & \text{, falls } N \text{ ungerade} \\ \frac{1}{2}\left(x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}\right) & \text{, falls } N \text{ gerade} \end{cases}$$

- Teilt die Daten in zwei gleiche Hälften
- Sehr anschaulich
- Stabil gegenüber monotonen Transformationen und Ausreißern
- Geeignet für ordinale und metrische Daten

Lagemaße: Median $\bar{x}_{med}$

- Beispiel 1: Zahl der täglich gesendeten Kurznachrichten ($N = 7$)

Geordnete Reihe:

Fall	1	2	3	4	5	6	7
x	0	5	5	10	15	20	30

➢ N ungerade: $\bar{x}_{med} = x_{\left(\frac{N+1}{2}\right)} = x_{\left(\frac{7+1}{2}\right)} = x_4 = \mathbf{10}$

- Beispiel 2: Tägliche Fernsehdauer in Stunden ($N = 8$)

Geordnete Reihe:

Fall	1	2	3	4	5	6	7	8
x	0	1	1	2	3	5	5	8

➢ N gerade: $\bar{x}_{med} = \frac{1}{2}\left(x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}\right) = \frac{1}{2}\left(x_4 + x_5\right) = \frac{1}{2}\left(2 + 3\right) = \mathbf{2,5}$

Lagemaße: arithmetisches Mittel $\overline{x}$

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

- Am häufigsten verwendetes Lagemaß
- Anfällig gegen extreme Werte
- Geeignet für metrische Daten

Lagemaße: arithmetisches Mittel $\overline{x}$

- Beispiel: Mass Bier beim Oktoberfest

x_i	f_i	$x_i \cdot f_i$
0	15	0
1	36	36
2	29	58
3	18	54
4	9	36
5	1	5
6	1	6
Summe	N=109	195

Vereinfachte Berechnung bei größerem N:

$$\overline{x} = \frac{\sum x_i \cdot f_i}{N}$$

$$= \frac{195}{109}$$

$$= 1{,}79$$

In a nutshell…

- Modus (Modalwert): $\overline{x}_{mod}$;
  am häufigsten vorkommender Wert
- Median: $\bar{x}_{med} = \begin{cases} x_{\left(\frac{N+1}{2}\right)} & \text{, falls } N \text{ ungerade} \\ \frac{1}{2}\left(x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}\right) & \text{, falls } N \text{ gerade} \end{cases}$
- arithmetisches Mittel: $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} = \frac{x_1+x_2+\cdots+x_N}{N}$

Übungsblatt 2: Aufgabe 1

Um sich einen Überblick über die Lesefreudigkeit der Studierenden zu machen, wurde von den KW-Professoren eine Befragung durchgeführt. Die Studierenden mussten angeben, an wie vielen Tagen der Woche sie das Institut zum Literaturstudium betreten. Die Stichprobe lieferte folgende Ergebnisse:

3	2	5	3	1	0	3	0	2	2
3	0	2	3	0	4	2	4	5	3
4	3	2	4	4	3	3	1	3	1

Ermitteln Sie ...

- absolute Häufigkeiten (mittels Strichliste) und daraus
- Modus
- Median
- arithmetisches Mittel

Tipp: Schneller ohne Taschenrechner !