Statistik – Einführung

Beschreibende Statistik
Kapitel 2

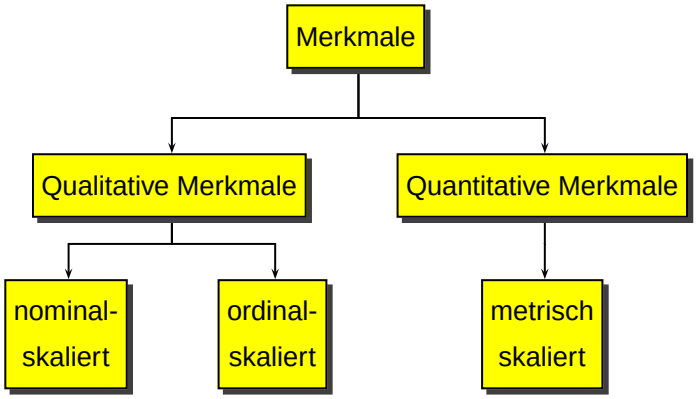
Statistik – WU Wien

Gerhard Derflinger · Michael Hauser · Jörg Lenneis · Josef Leydold ·
Günter Tirler · Rosmarie Wakolbinger

Lernziele

1. Charakterisieren Merkmale nach ihren Skalenniveaus.
2. Beschreiben qualitative Daten graphisch.
3. Beschreiben quantitative Daten graphisch.
4. Erzeugen und interpretieren graphische Darstellungen.
5. Erklären numerische Dateneigenschaften.
6. Beschreiben zusammenfassende Maßzahlen.
7. Analysieren numerische Daten mit Hilfe dieser Maßzahlen.

Skalenniveaus von Merkmalen



- **Nominalskala**
  Besitzt ein Merkmal für das eine Reihung der Ausprägungen nicht möglich bzw. sinnvoll ist.
  Die Merkmalsausprägungen heißen **Kategorien**.
  - E.g., Geschlecht (m,w), Religion (r.k., evang. HB, ...), Haarfarbe (blond, brünett, rot, ...), ...

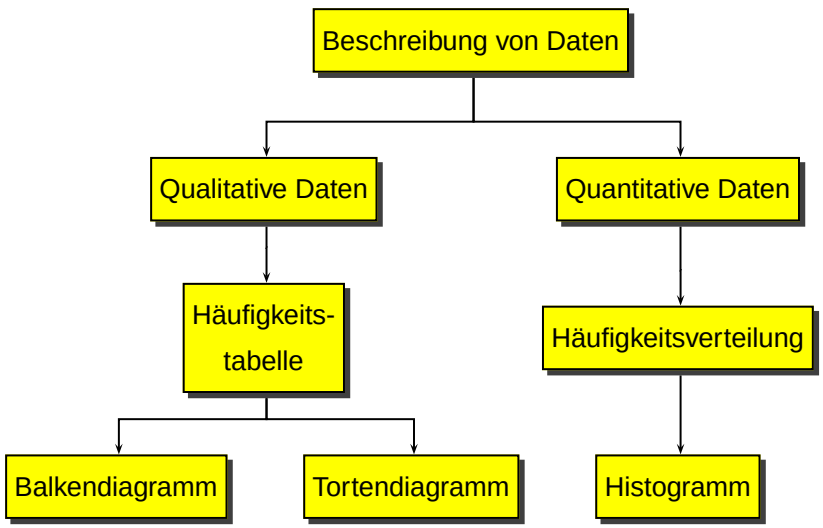
- **Ordinalskala**
  Zwischen einzelnen Merkmalsausprägungen gibt es eine natürliche Rangordnung. Differenzen lassen sich aber nicht quantifizieren.
  - E.g., Schulabschluß (Lehre, Matura, Studium), Prüfungsnoten (1,2,3,4,5), Güteklassen, ...

Skalenniveaus // Quantitative Merkmale

- **Metrische Skala**
  Merkmalsausprägungen lassen sich ordnen. Und es können Abstände numerisch angegeben werden.
  Es gibt meist viele (verschiedene) Merkmalsausprägungen.
  - Z.B.: Körpergröße, Alter, Einkommen, ...

Skalenniveaus bilden eine Hierarchie, i.e., ein metrisch skaliertes Merkmal ist auch ordinal- oder nominalskaliert, aber nicht umgekehrt.

Präsentation von Daten



Präsentation qualitativer Daten

Häufigkeitstabelle

1. Häufigkeitstabelle eines Merkmals gibt die Kategorien (Merkmalsausprägungen) und die Zahl der Elemente pro Kategorie an.
2. Erhält man, indem die Antworten den Kategorien zugeordnet werden (Strichliste).
3. Angegeben werden die absoluten oder relativen Häufigkeiten (Prozentsätze), oder beides.

Häufigkeitstabelle

Befragung von 200 Studenten:

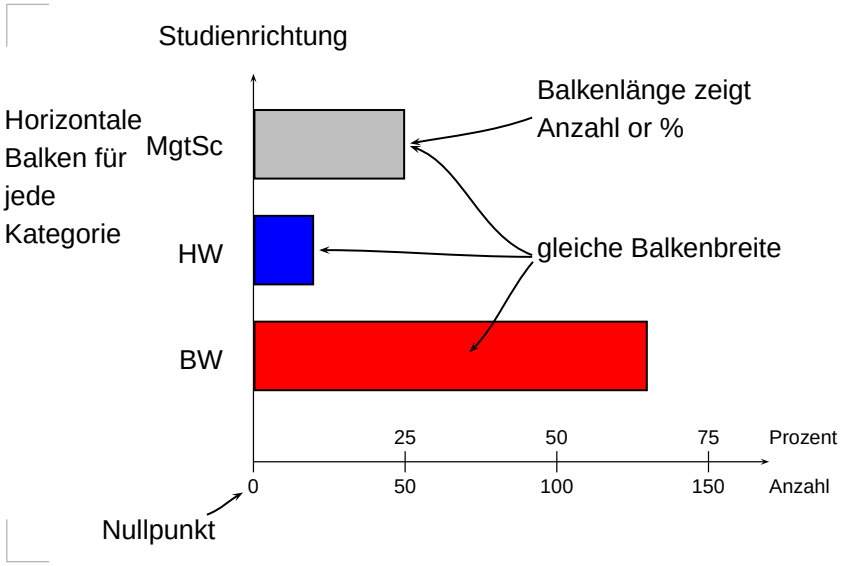
Merkmal

Studienrichtung	Anzahl	Anteil
Betriebswirtschaft	130	65%
Handelswissenschaft	20	10%
Management Science	50	25%
Gesamt	200	100%

$\frac{130}{200} \cdot 100\%$

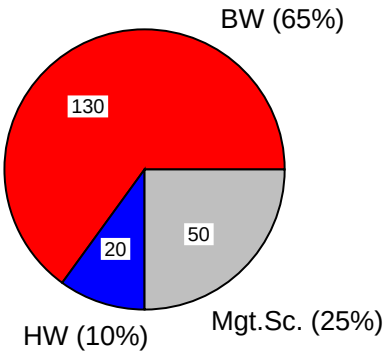
Kategorie (Ausprägung)

aus Strichliste



Graphische Darstellung – Tortendiagramm

- Zeigt die Aufteilung in einzelne Kategorien.
- Darstellung der Anteile.
- Winkel $=$ Anteil $\times$ 360°

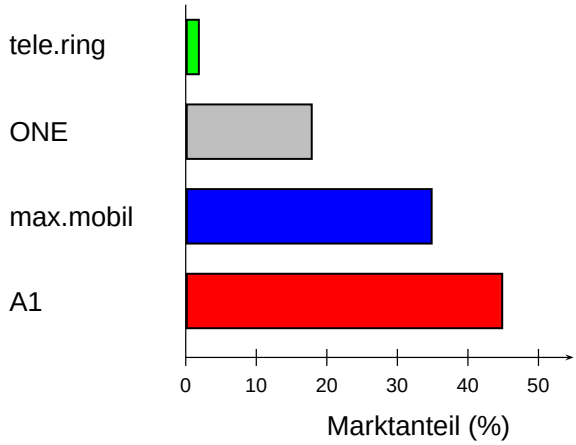


Beispiel // Marktanteil

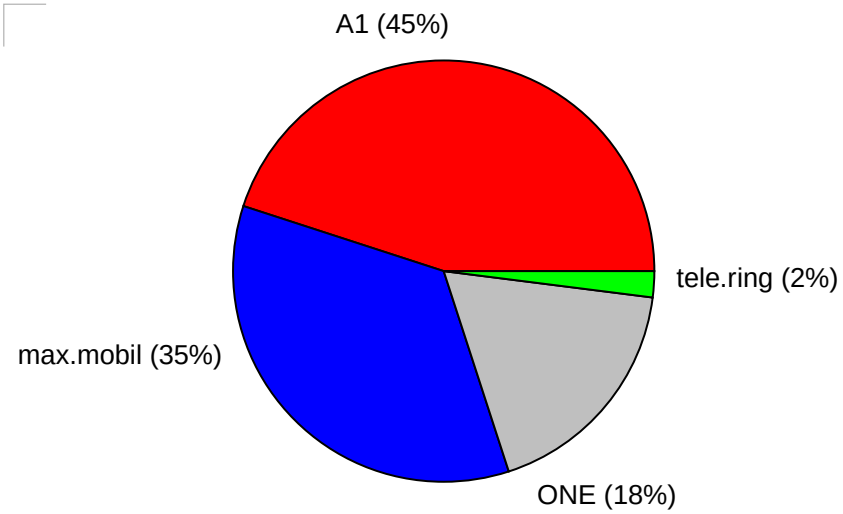
Sie analysieren den österreichischen Handymarkt und wollen die Marktanteile (des Jahres 2000) graphisch darstellen.

Die **Marktanteile** (in %):

A1	45
max.mobil	35
ONE	18
tele.ring	2



Beispiel // Lösung



Präsentation quantitativer Daten

Einteilung in Klassen

Beispiel:
Rohdaten: 24, 26, 24, 21, 27, 27, 30, 41, 32, 38

Klasse	Anzahl
15 – 25	3
25 – 35	5
35 – 45	2
Summe	10

Konvention: 15 bis unter 25, d.h., 15 gehört zur Klasse, 25 nicht.

Häufigkeitsverteilung // Tabellenerstellung

Vorgangsweise bei der Klasseneinteilung:

1. Bereich (kleinsten und größten Wert) bestimmen.
2. Geeignete Zahl an Klassen auswählen.
   (Typischerweise zwischen 5 und 15)
3. Klassenbreite bestimmen.
   Die Klassen sind in der Regel gleich breit.
4. Klassengrenzen bestimmen.
5. Klassenmitte berechnen.
6. Beobachtungen abzählen und den Klassen zuordnen.

Häufigkeitsverteilung // Tabellenbeispiel

Rohdaten: 24, 26, 24, 21, 27, 27, 30, 41, 32, 38

Klasse	Klassenmitte	Anzahl
15 – 25	20	3
25 – 35	30	5
35 – 45	40	2
Summe		10

Klassengrenzen

(obere + untere Grenze) / 2

Klassenbreite $= 45 - 35 = 10$

Rohdaten: 24, 26, 24, 21, 27, 27, 30, 41, 32, 38

	Häufigkeiten	
	absolute	relative

Klassen	Anzahl	Anteil	Prozent
15 – 25	3	0.3	30
25 – 35	5	0.5	50
35 – 45	2	0.2	20
Summe	10	1.0	100

Relative Summenhäufigkeitsverteilung

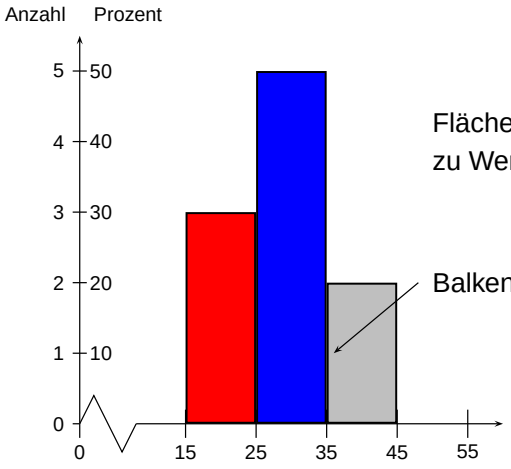
Gibt an, wieviele Beobachtungen – relativ oder prozentuell – einen Wert besitzen, der sich links der jeweiligen oberen Klassengrenze befindet.
(Links bedeutet: Kleiner als die obere Klassengrenze.)

Rohdaten: 24, 26, 24, 21, 27, 27, 30, 41, 32, 38

		Summenhäufig-keiten	
Klassen	Anteil	relativ	prozentuell
15 – 25	0.3	0.3	30
25 – 35	0.5	0.8	80 ← $30 + 50$
35 – 45	0.2	1.0	100

Histogramm

absolute oder relative Werte



Flächen proportional zu Werten

Balken berühren sich

Numerische Beschreibung quantitativer Daten

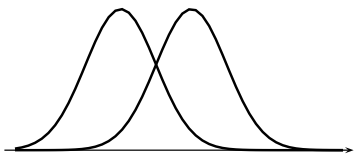
Notation

Maß	Stichprobe	Grundgesamtheit
Mittelwert	$\bar{x}$	μ
Standardabweichung	s	σ
Varianz	s^2	σ^2
Umfang	n	N

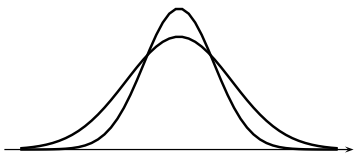
μ heißt auch Erwartungswert.

Numerische Charakterisierung

- **Lage**: Lagemaße
  - arithmetisches Mittel
  - Median
  - Modus



- **Streuung**: Streuungsmaße
  - Spannweite
  - Interquartilsabstand
  - Varianz
  - Standardabweichung



- **Form**
  - Schiefe

Lage und Lagemaße

(Arithmetisches) Mittel

- Maß für die Lage der Verteilung
  - Für metrisch skalierte Daten
- Am häufigsten verwendetes Maß
- Durchschnitt
  „Durchschnittlicher Wert“ der Daten
- Empfindlich gegen „Ausreißer“
- Definition (Formel für Stichprobenmittel)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Arithmetisches Mittel // Beispiel

Rohdaten: 10.3, 4.9, 8.9, 11.7, 6.3, 7.7

$$\begin{aligned}
\bar{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\
&= \frac{x_1 + x_2 + x_3 + x_4 + x_5 + x_6}{6} \\
&= \frac{10.3 + 4.9 + 8.9 + 11.7 + 6.3 + 7.7}{6} \\
&= 8.30
\end{aligned}$$

- Notwendig zur Berechnung des arithmetischen Mittels aus einer Häufigkeitstabelle.
- k verschiedene Werte $x_1, \ldots, x_k$ mit den Häufigkeiten $h_1, \ldots, h_k$.
- Berechnung

$$\bar{x} = \sum_{i=1}^{k} x_i \frac{h_i}{n} = \frac{1}{n} \sum_{i=1}^{k} x_i h_i$$

wobei $n = \sum_{i=1}^{k} h_i$

Gewichtetes Mittel // Beispiel

In einer Lehrveranstaltung wurde die Anzahl der Geschwister der Teilnehmer erhoben:

Anzahl Geschwister	x_i	0	1	2	3	4
Häuigkeit	h_i	12	14	5	2	0

Die durschnittliche Anzahl an Geschwister erhalten wir mit

$$n = 12 + 14 + 5 + 2 = 33$$

$$\bar{x} = \frac{1}{33} \left(0 \cdot 12 + 1 \cdot 14 + 2 \cdot 5 + 3 \cdot 2\right) = \frac{30}{33} = 0.909$$

Median

- Maß für die Lage der Verteilung
  - Für ordinalskalierte Daten
- Definition:
  Mittlerer Wert in der geordneten Liste
  - ungerades n: Mittlerer Wert
  - gerades n: Durchschnitt der beiden mittleren Werten
- Position des Medians $= \frac{n+1}{2}$
- Robust (unempfindlich) gegenüber Ausreißern.

Ungerade Anzahl von Daten:

Rohdaten	24.1	22.6	21.5	23.7	22.6
Sortiert	21.5	22.6	22.6	23.7	24.1
Position	1	2	**3**	4	5
			▲		

Position des Median $= \frac{n+1}{2} = \frac{5+1}{2} = 3$

Median = 22.6

Median // Beispiel für gerades n

Gerade Anzahl von Daten:

Rohdaten	10.3	4.9	8.9	11.7	6.3	7.7
Sortiert	4.9	6.3	7.7	8.9	10.3	11.7
Position	1	2	**3**	**4**	5	6
				▲		

Position des Median $= \frac{n+1}{2} = \frac{6+1}{2} = 3.5$

Median $= \frac{7.7+8.9}{2} = 8.3$

Modus

- Maß für die Lage der Verteilung
  - Auch für nominalskalierte Daten
- Definition:
  Häufigster Wert
- Kein, ein oder mehrere Modi sind möglich
- Kann auch bei qualitativen und quantitativen Daten verwendet werden
- Robust (unempfindlich) gegenüber Ausreißern

Kein Modus:

Rohdaten 10.3 4.9 8.9 11.7 6.3 7.7

Ein Modus:

Rohdaten 6.3 **4.9** 8.9 6.3 **4.9** **4.9**

Mehrere Modi:

Rohdaten 21 **28** **28** 41 **43** **43**

Lagemaße // Problem

Sie sind Finanzanalyst bei der BANK AUSTRIA. Sie haben Tagesendkurse von Neuemissionen von Aktien gesammelt:

17, 16, 21, 18, 13, 16, 12 und 11.

Beschreiben Sie die Lage der Verteilung der Kurse.

Lagemaße // Lösung

(Arithmetisches) Mittel:

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\ &= \frac{x_1 + x_2 + \ldots + x_8}{8} \\ &= \frac{17 + 16 + 21 + 18 + 13 + 16 + 12 + 11}{8} \\ &= 15.5\end{aligned}$$

Median:

Rohdaten	17	16	21	18	13	16	12	11
Sortiert	11	12	13	16	16	17	18	21
Position	1	2	3	**4**	**5**	6	7	8

▲

Position des Median $= \frac{n+1}{2} = \frac{8+1}{2} = 4.5$

Median $= \frac{16+16}{2} = 16$

Lagemaße // Lösung

Modus:

Rohdaten	17	16	21	18	13	16	12	11
Sortiert	11	12	13	**16**	**16**	17	18	21

Modus $= 16$

Lagemaße // Zusammenfassung

Begriff	Formel	Beschreibung	Skalenniveau
Mittel (arithm.)	$\bar{x} = \frac{\sum x_i}{n}$	„Durchschnitt“	metrisch
Median	Position $= \frac{n+1}{2}$	mittlerer Wert bei geordneten Daten	ordinalsk.
Modus		häufigster Wert	nominalsk.

Streuung und Streuungsmaße

Spannweite

- Maß für die Streuung metrisch skalierter Merkmale
- Definition:
  Differenz zwischen größter und kleinster Beobachtung

$$\text{Spannweite} = x_{\max} - x_{\min}$$

- Sehr empfindlich gegenüber Ausreißern

Varianz

- Maß für die Streuung für metrisch skalierte Merkmale
- Berücksichtigt die Verteilung
- Zeigt die Abweichung von Mittelwert, $\bar{x}$,
  bzw. Erwartungswert, μ
- Definition der Varianz:

Grundgesamtheit

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Stichprobe

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Die Standardabweichung ist die positive Quadratwurzel der Varianz.
Die Standardabweichung hat die gleiche Dimension wie Mittelwert.

- Standardabweichung der Grundgesamtheit:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

- Standardabweichung einer Stichprobe:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Varianz // Beispiel

Rohdaten: 10.3 4.9 8.9 11.7 6.3 7.7

Varianz in der Stichprobe:

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \quad \text{wobei} \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 8.300 \\ &= \frac{(10.3 - 8.3)^2 + (4.9 - 8.3)^2 + \ldots + (7.7 - 8.3)^2}{6-1} \\ &= 6.368 \end{aligned}$$

Standardabweichung in der Stichprobe:

$$s = \sqrt{s^2} = \sqrt{6.368} = 2.523$$

Varianz // Berechnung

Die Varianz der Grundgesamtheit berechnet sich leichter durch folgende Formeln (Verschiebungssatz):

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2$$

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2}{n-1} - \frac{n}{n-1}\bar{x}^2$$

Rohdaten: 10.3 4.9 8.9 11.7 6.3 7.7

Varianz in der Stichprobe:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2}{n-1} - \frac{n}{n-1}\bar{x}^2 \qquad \text{wobei} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 8.300$$

$$= \frac{10.3^2 + 4.9^2 + 8.9^2 + 11.7^2 + 6.3^2 + 7.7^2}{6-1} - \frac{6}{6-1}\,8.3^2$$

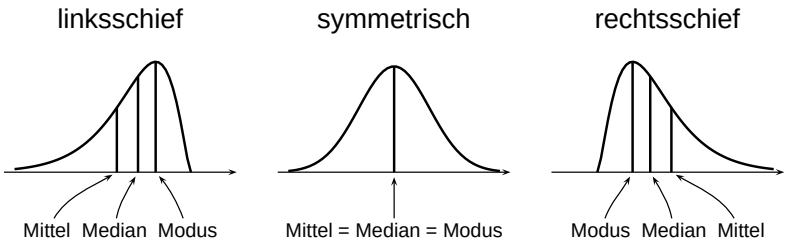
$$= 6.368$$

Streuungsmaße // Zusammenfassung

Begriff	Formel	Beschreibung
Spannweite	$x_{\max} - x_{\min}$	maximale Streuung
Standardabweichung (Stichprobe)	$\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$	Streuung um das Stichprobenmittel
Standardabweichung (Grundgesamtheit)	$\sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$	... um das Mittel in der Grundgesamtheit
Varianz (Stichprobe)	$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$	Quadrierte Streuung um das SP-Mittel
Varianz (Grundgesamtheit)	$\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$	... um das Mittel in der Grundgesamtheit

Form der Verteilung

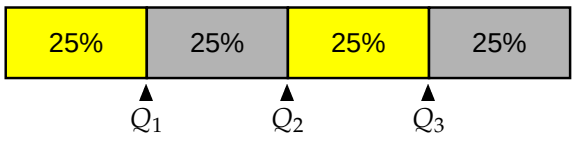
- Beschreibt die Verteilung der Daten
- Maß für die Form ist die **Schiefe**
  (= Abweichung von der Symmetrie)



Quartile und Box-Plots

Quartile

- Beschreiben die Verteilung der Daten x_j, $j = 1, \ldots, n$.
- Die **geordneten** Daten werden in vier gleiche Teile zerlegt.
  Das i-te Quartil ist der maximale x-Wert des i-ten Teiles:



- Position des i-ten Quartils $= i\,\dfrac{n+1}{4}$, $\quad i = 1, 2, 3$
- $1/4$ aller Daten sind kleiner oder gleich Q_1.
  $1/2$ aller Daten sind kleiner oder gleich Q_2 ($\widehat{=}$ Median).
  $3/4$ aller Daten sind kleiner oder gleich Q_3.

Rohdaten	10.3	4.9	8.9	11.7	6.3	7.7
Sortiert	4.9	6.3	7.7	8.9	10.3	11.7
Position	1	**2**	3	4	5	6

▲

$$Q_1\text{-Position} = \frac{1 \cdot (n+1)}{4} = \frac{1 \cdot (6+1)}{4} = 1.75 \cong 2$$

$$Q_1 = 6.3$$

Quartile // Beispiel – Q_2

Rohdaten	10.3	4.9	8.9	11.7	6.3	7.7
Sortiert	4.9	6.3	7.7	8.9	10.3	11.7
Position	1	2	**3**	**4**	5	6

▲

$$Q_2\text{-Position} = \frac{2 \cdot (n+1)}{4} = \frac{2 \cdot (6+1)}{4} = 3.5$$

$$Q_2 = \frac{7.7 + 8.9}{2} = 8.3$$

Quartile // Beispiel – Q_3

Rohdaten	10.3	4.9	8.9	11.7	6.3	7.7
Sortiert	4.9	6.3	7.7	8.9	10.3	11.7
Position	1	2	3	4	**5**	6

▲

$$Q_3\text{-Position} = \frac{3 \cdot (n+1)}{4} = \frac{3 \cdot (6+1)}{4} = 5.25 \cong 5$$

$$Q_3 = 10.3$$

- Streuungsmaß für metrisch skalierte Daten
- Definition:
  Differenz zwischen 3. und 1. Quartil

$$\text{Interquartilsabstand} = Q_3 - Q_1$$

- Zeigt den Bereich der mittleren 50% der Daten
- Robust (unempfindlich) gegenüber Ausreißern

Interquartilsabstand // Beispiel

Rohdaten	10.3	4.9	8.9	11.7	6.3	7.7
Sortiert	4.9	6.3	7.7	8.9	10.3	11.7
Position	1	**2**	3	4	**5**	6
		▲			▲	
		Q_1			Q_3	

$Q_1 = 6.3$
$Q_3 = 10.3$

Interquartilsabstand $= Q_3 - Q_1 = 10.3 - 6.3 = 4$

Interquartilsabstand // Beispiel

Sie sind Finanzanalyst der BANK AUSTRIA und wollen die Verteilung der folgenden acht Tagesendkurse von Aktien mit Hilfe der Quartile Q_1 und Q_3 sowie mit dem Interquartilsabstand beschreiben:

17, 16, 21, 18, 13, 16, 12 und 11.

Rohdaten	17	16	21	18	13	16	12	11
Sortiert	11	12	13	16	16	17	18	21
Position	1	**2**	3	4	5	6	**7**	8
		▲					▲	
		Q_1					Q_3	

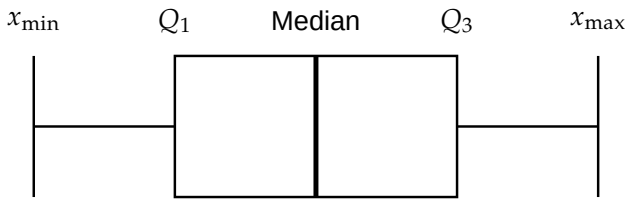
$Q_1 = 12$ (Position $= \frac{1 \cdot (n+1)}{4} = \frac{1 \cdot (8+1)}{4} = 2.25 \cong 2$)

$Q_3 = 18$ (Position $= \frac{3 \cdot (n+1)}{4} = \frac{3 \cdot (8+1)}{4} = 6.75 \cong 7$)

Interquartilsabstand $= Q_3 - Q_1 = 18 - 12 = 6$

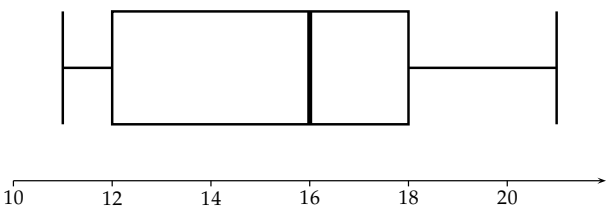
Box-Plot

- Graphische Darstellung der Daten;
  verwendet 5 Kenngrößen:



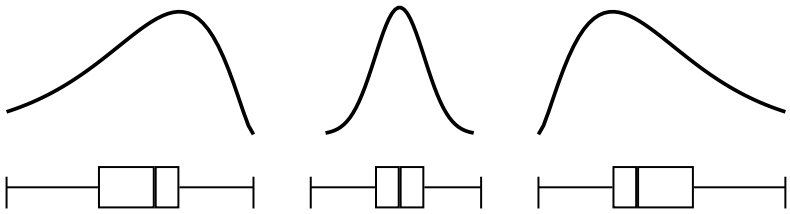
Box-Plot // Beispiel

Rohdaten	17	16	21	18	13	16	12	11
Sortiert	11	12	13	16	16	17	18	21
Position	**1**	**2**	3	**4**	**5**	6	**7**	8
	▲	▲		▲			▲	▲
	$x_{\min}$	Q_1		*Median*			Q_3	$x_{\max}$



Verteilung

linksschief symmetrisch rechtsschief



Zusammenfassung

1. Merkmale nach ihren Skalenniveaus charakterisiert.
2. Qualitative Daten graphisch beschrieben.
3. Quantitative Daten graphisch beschrieben.
4. Graphische Darstellungen erzeugt und interpretiert.
5. Numerische Dateneigenschaften erklärt.
6. Zusammenfassende Maßzahlen beschrieben.
7. Numerische Daten mit Hilfe dieser Maßzahlen analysiert.