LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR KOMMUNIKATIONSWISSENSCHAFT UND MEDIENFORSCHUNG

BA KW | Vorlesung

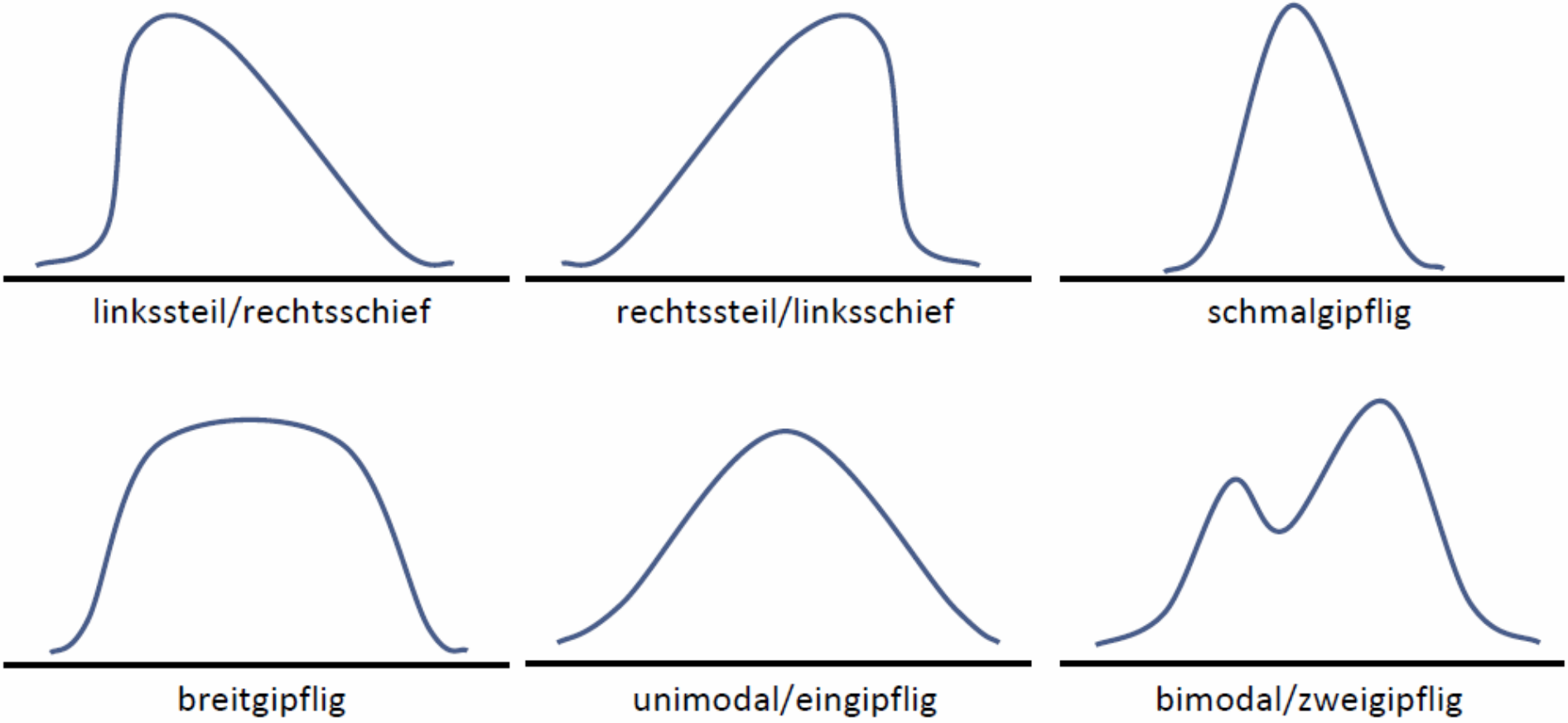
Einführung in die Statistik

Streuung

Prof. Thomas Hanitzsch

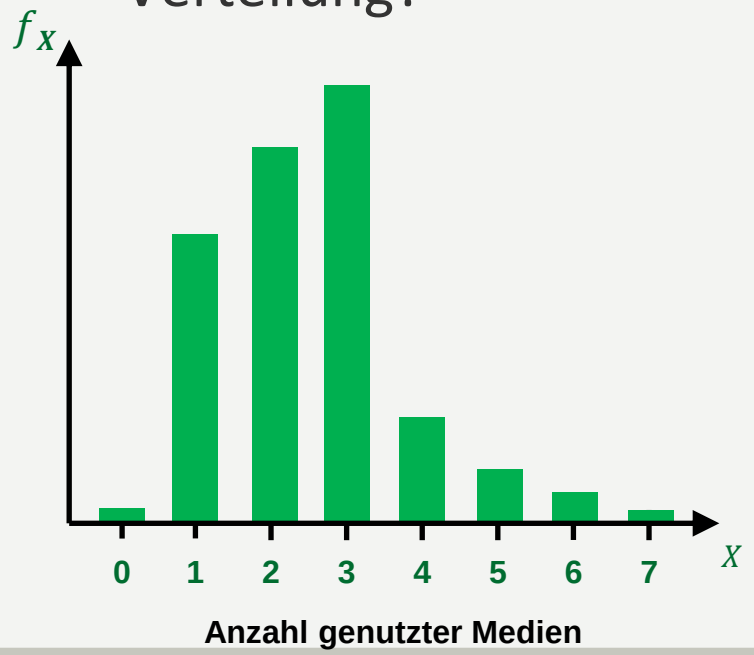
Empirische Verteilungen: Überblick

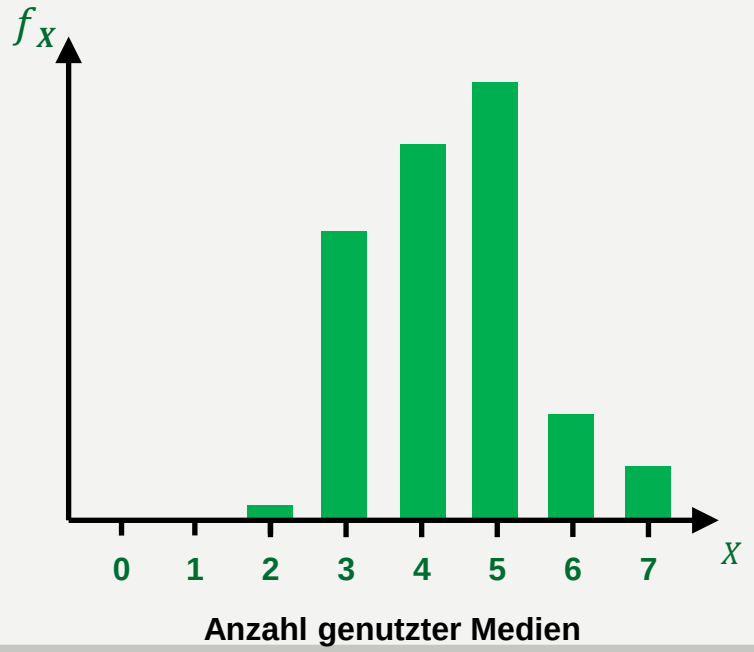
(Kuckartz et al. 2013: 47)



Eigenschaften von Häufigkeitsverteilungen

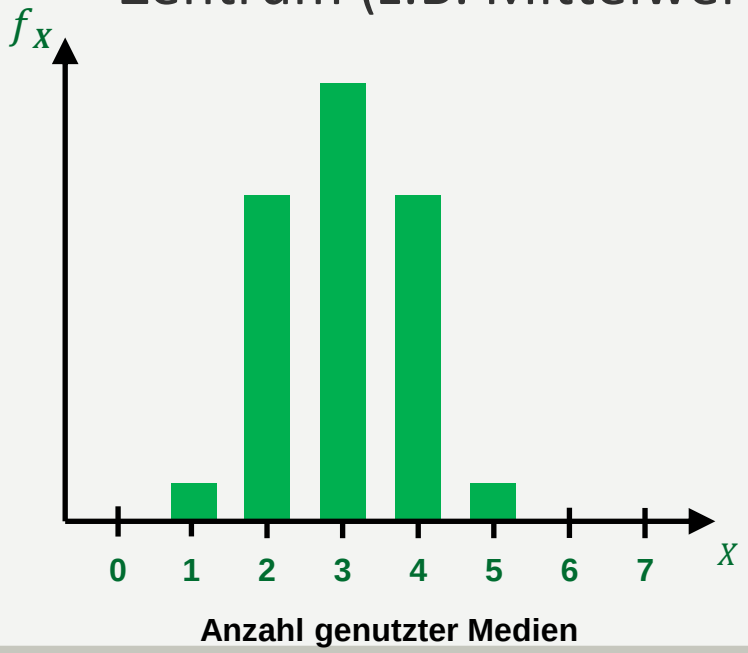
- Zentrale Tendenz:
  - Wo liegt das Zentrum bzw. der Schwerpunkt der Verteilung?

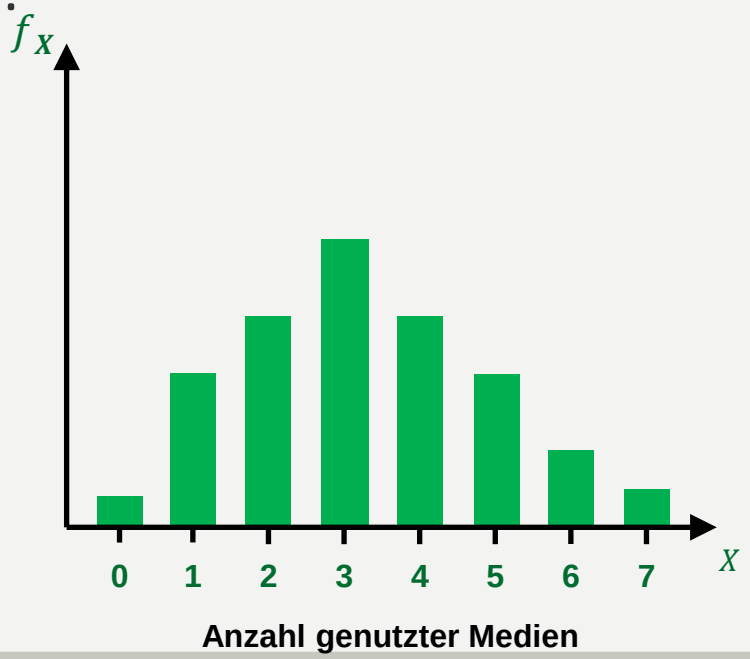




Eigenschaften von Häufigkeitsverteilungen

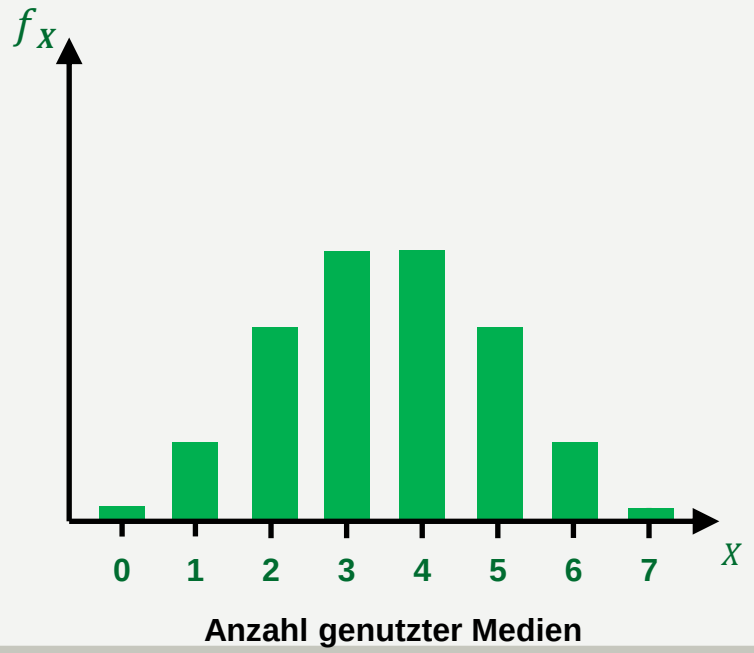
- Streuung:
  - Wie stark weichen die einzelnen Beobachtungswerte vom Zentrum (z.B. Mittelwert) ab?

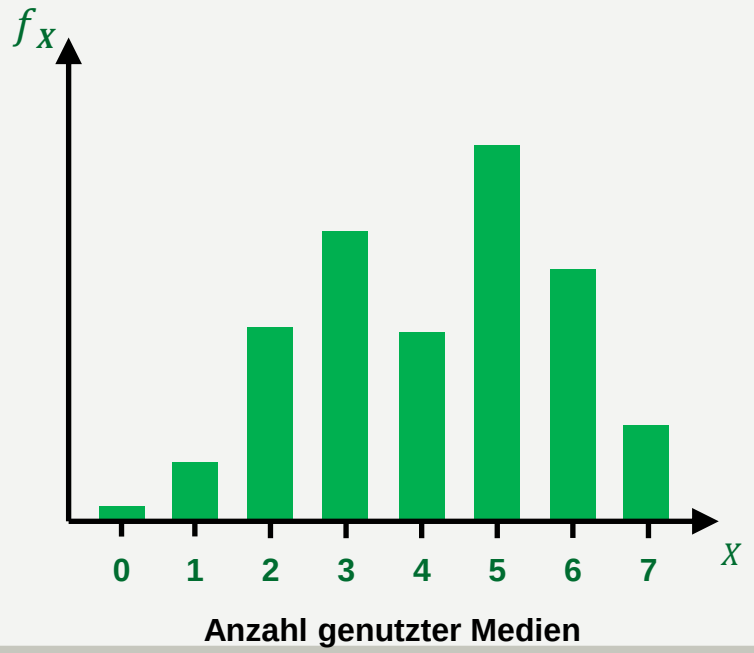




Eigenschaften von Häufigkeitsverteilungen

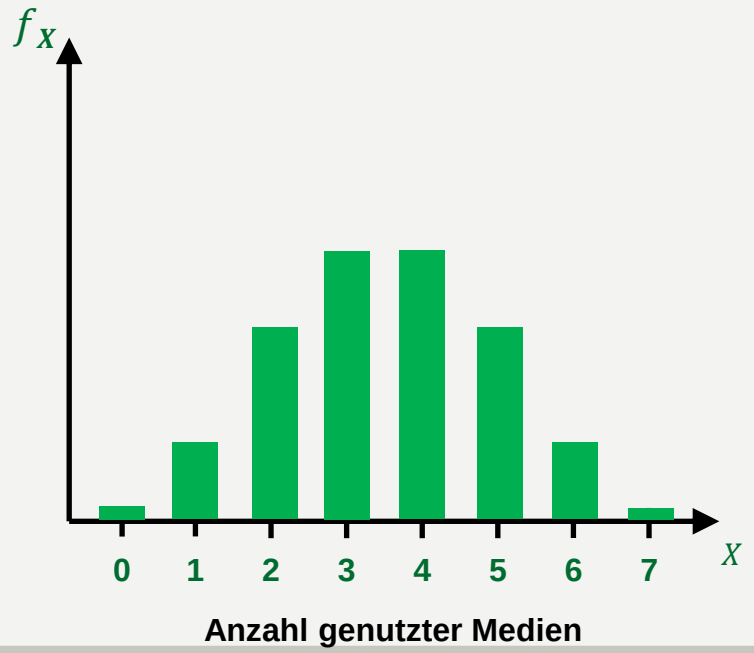
- Modalität:
  - Wie viele „Gipfel“ hat die Verteilung?

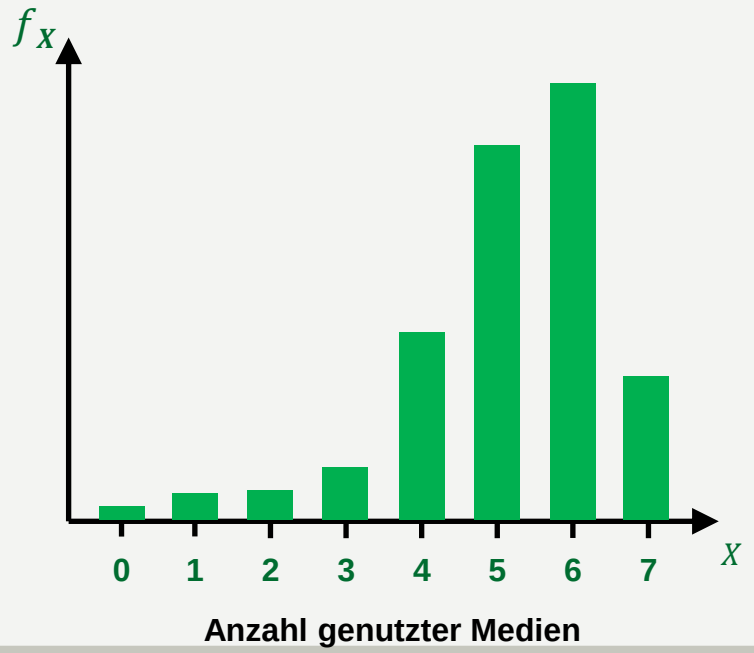




Eigenschaften von Häufigkeitsverteilungen

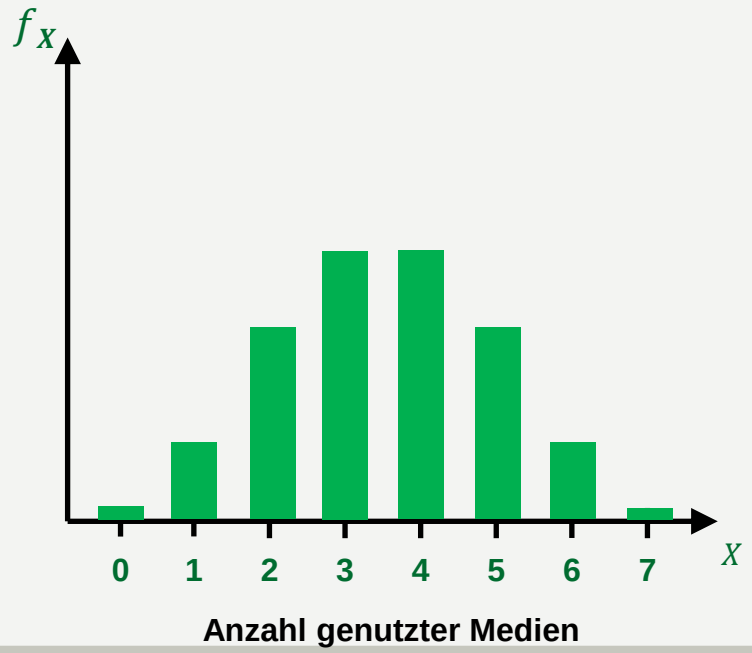
- Schiefe:
  - Wie symmetrisch ist die Verteilung?

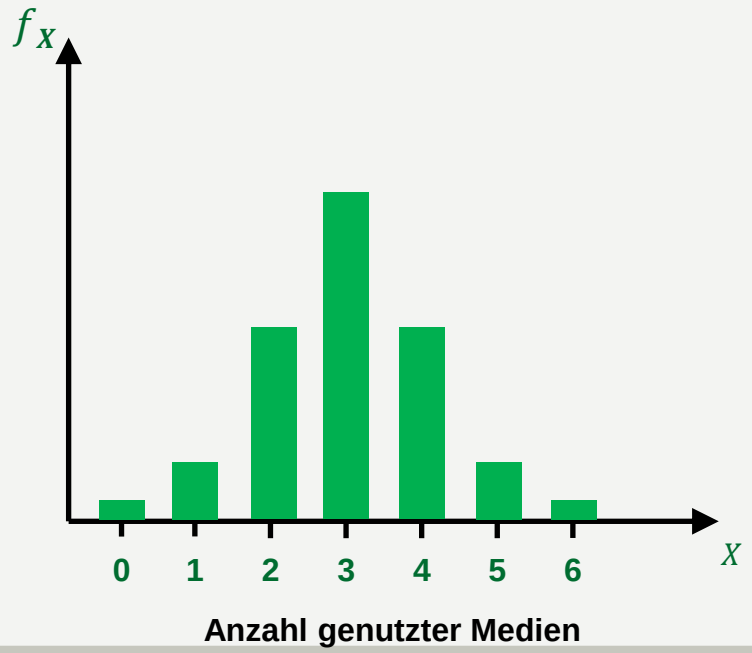




Eigenschaften von Häufigkeitsverteilungen

- Wölbung (Kurtosis):
  - Wie spitz bzw. gestaucht ist die Verteilung?





Streuungsmaße: Spannweite R

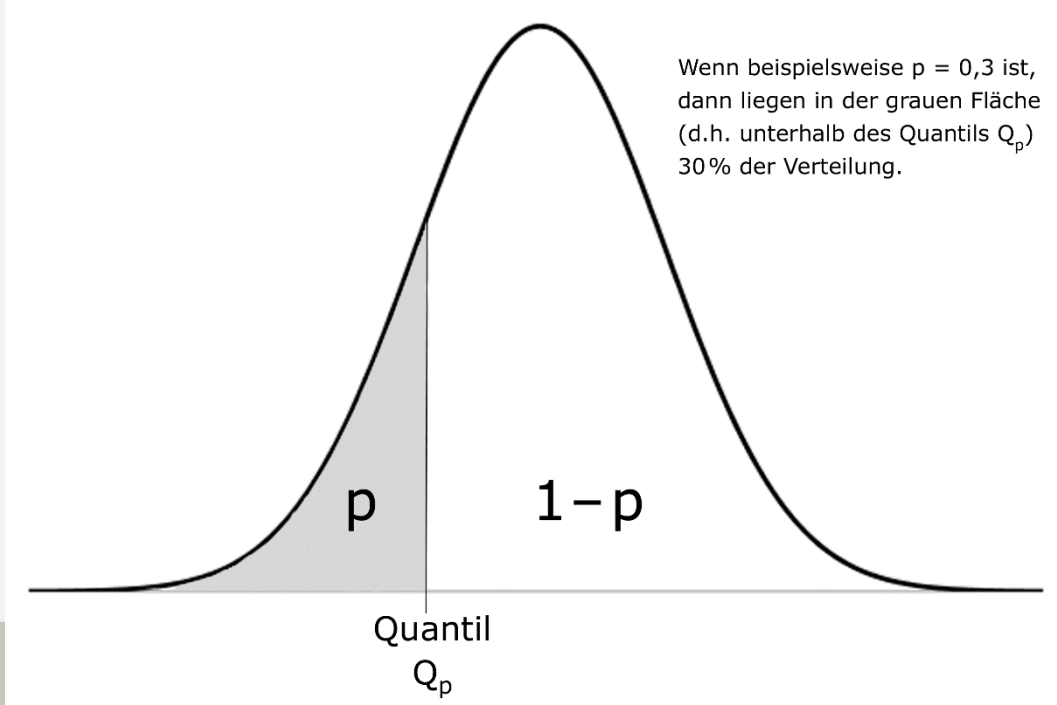
$$R = x_{max} - x_{min}$$

- Differenz von größtem (Maximum) und kleinstem (Minimum) vorkommendem Merkmalswert
- Ist der Bereich, in dem die empirischen Daten liegen
- Wichtig für die Datenkontrolle (z.B. zur Identifikation von fehlerhaften Dateneingaben)
- Anfällig gegen Ausreißer, da aus Extremwerten gebildet

Streuungsmaße: Quantile $\tilde{x}_p$

$$\tilde{x}_p = \begin{cases} \frac{1}{2}\left(x_{N\cdot p} + x_{Np+1}\right) & falls\ N \cdot p\ ganzzahlig \\ x_{\lceil N\cdot p \rceil} & falls\ N \cdot p\ nicht\ ganzzahlig \end{cases}$$

- Das p-Quantil teilt die geordnete Liste im Verhältnis p zu $1 - p$



Streuungsmaße: Quantile $\tilde{x}_p$

$$\tilde{x}_p = \begin{cases} \frac{1}{2}\left(x_{N\cdot p} + x_{Np+1}\right) & falls\ N \cdot p\ ganzzahlig \\ x_{\lceil N\cdot p \rceil} & falls\ N \cdot p\ nicht\ ganzzahlig \end{cases}$$

- Das p-Quantil teilt die geordnete Liste im Verhältnis p zu $1 - p$
- Sinnvoll für metrische Merkmale
- Typische Quantile:
  - *Median* (0,5-Quantil)
  - *Quartile*: 0,25-, 0,5- und 0,75-Quantil (gemeinsam mit Maximum und Minimum „Fünf-Punkte-Zusammenfassung“)
- Auch: *Perzentil* (0,5-Quantil = 50%-Perzentil)

Streuungsmaße: Quantile $\tilde{x}_p$

$$\tilde{x}_p = \begin{cases} \frac{1}{2}\left(x_{N\cdot p} + x_{N\cdot p+1}\right) & falls\ N \cdot p\ ganzzahlig \\ x_{\lceil N\cdot p \rceil} & falls\ N \cdot p\ nicht\ ganzzahlig \end{cases}$$

- Beispiel:

Fall	1	2	3	4	5	6	7	8	9	10	
x_i	1	1	1	2	2	3	4	5	6	6	(N=10)

- 0,5-Quantil (Median): $\tilde{x}_{0,5} = \frac{1}{2}\left(x_{10\cdot 0,5} + x_{10\cdot 0,5+1}\right) = \frac{1}{2}(x_5 + x_6) = 2,5$
- 0,25-Quantil (Quartil $Q_{0,25}$): $\tilde{x}_{0,25} = x_{\lceil 10\cdot 0,25 \rceil} = x_3 = 1$
- 0,75-Quantil (Quartil $Q_{0,75}$): $\tilde{x}_{0,75} = x_{\lceil 10\cdot 0,75 \rceil} = x_8 = 5$

Streuungsmaße: Interquartilabstand IQR

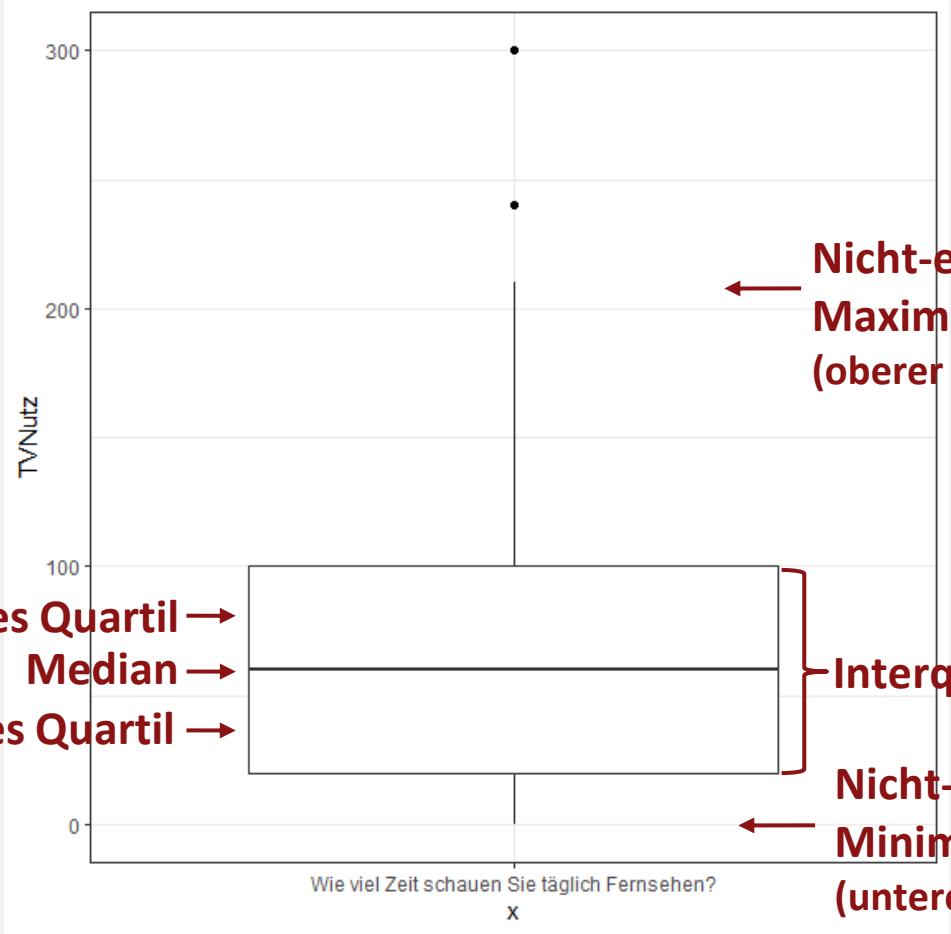
$$IQR = Q_{0,75} - Q_{0,25}$$

- Engl. *interquartile range* (IQR)
- Größe des Bereichs, in dem die mittlere Hälfte der Daten liegt
- Robuster gegenüber Ausreißern als die Spannweite
- Sinnvoll für metrische Daten

Grafische Darstellung: der Boxplot

- Auch: *Box-Whisker-Plot*
- Grafische Darstellung von Median und wichtigen Parametern der Streuung
- Eignet sich nur für unimodale Verteilungen und Daten auf metrischem Skalenniveau
- Boxplots vermitteln dem geübtem Auge eine schnelle Beurteilung der Verteilung
- Es lassen sich auch verschiedene Verteilungen anschaulich gegenüberstellen und vergleichen

Grafische Darstellung: der Boxplot



Nicht-extremes Maximum (oberer Whisker)

Spannweite

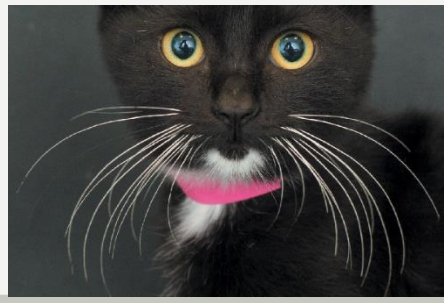
Oberes Quartil

Median

Unteres Quartil

Interquartilabstand

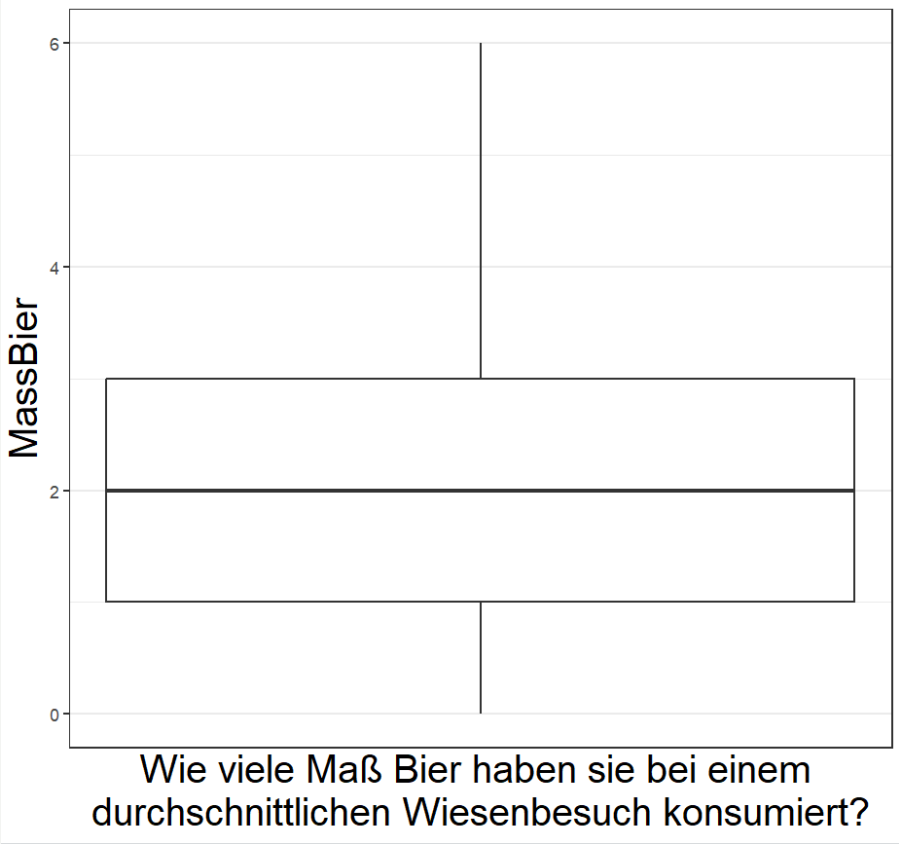
Nicht-extremes Minimum (unterer Whisker)

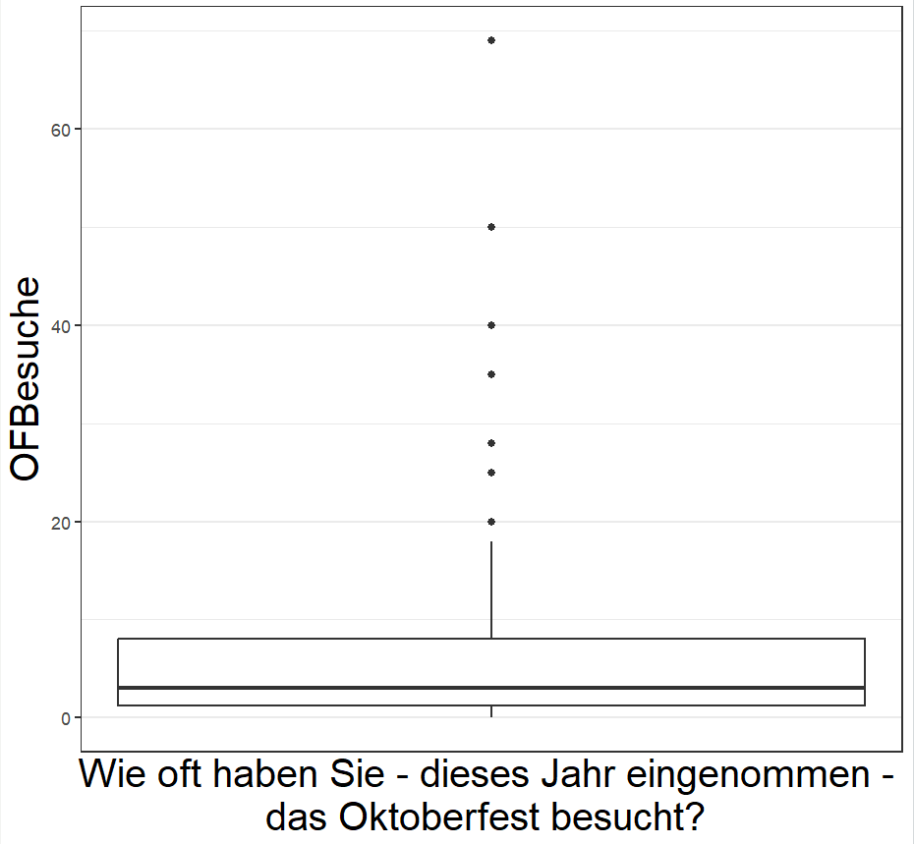


Grafische Darstellung: der Boxplot

Kennwert	Beschreibung	Lage im Boxplot
Unteres Quartil	25 % der Datenwerte sind kleiner oder gleich x_i	Unterkante der Box
Oberes Quartil	75 % der Datenwerte sind kleiner oder gleich x_i	Oberkante der Box
Interquartilabstand	Wertebereich, in dem sich die mittleren 50 % befinden	Vertikale Ausdehnung der Box
Median	Teilt den Datensatz in zwei Hälften	Waagerechter Strich in der Box
Nicht-extremes Minimum	Kleinster nicht-extremer Wert	Ende des unteren Whiskers oder kleinster Wert
Nicht-extremes Maximum	Größter nicht-extremer Wert	Ende des oberen Whiskers oder größter Wert
Spannweite	Gesamter Wertebereich	Zwischen kleinstem und größtem Wert

Boxplots: zwei Beispiele



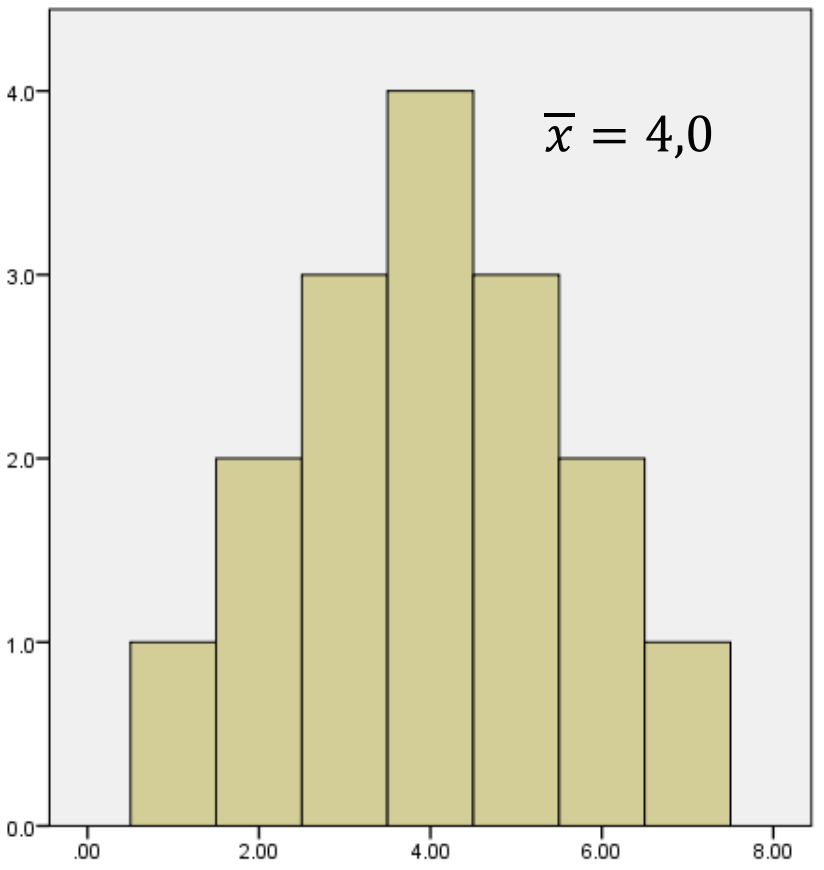


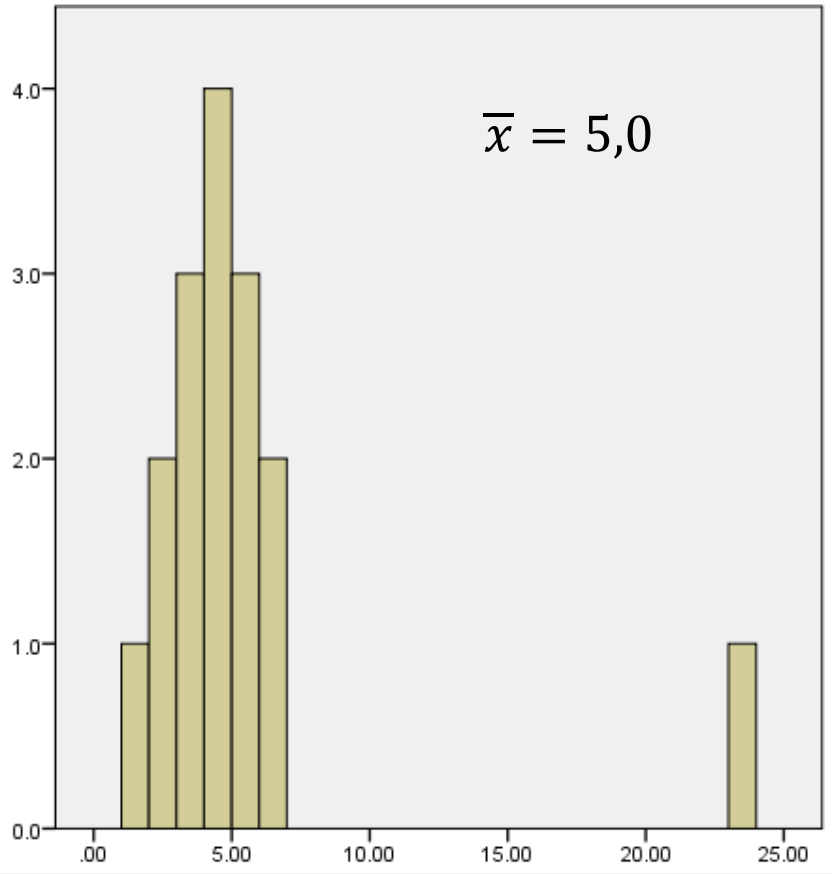
Extremwerte und Ausreißer

- **Ausreißer:**
  Werte, die weiter als 1,5 Interquartilabstände vom 1. bzw. 3. Quartil entfernt liegen
- **Extremwerte:**
  Werte, die weiter als 3 Interquartilabstände vom 1. bzw. 3. Quartil entfernt liegen

- Ausreißer und Extremwerte sollten durch eine sorgefältige Inspektion der Verteilung identifiziert werden, da sie wichtige Kennwerte verzerren (Mittelwerte!) und zu Verzerrungen bei statistischen Auswertungen führen können

Extremwerte und Ausreißer: Beispiel





Streuungsmaße: empirische Varianz s^2

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

- Beschreibt die Variabilität der empirischen Werte hinsichtlich eines Merkmals
- Definiert als die mittlere quadratische Abweichung der Werte vom arithmetischen Mittel
- Vorteil: alle Werte einer Verteilung werden berücksichtigt
- Sind alle Werte identisch ⍰ $s^2 = 0$
- Geeignet für metrische Skalen, wird in der Praxis aber auch für quasi-metrische Daten verwendet (mind. 5 Abstufungen)

Streuungsmaße: Standardabweichung s

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

- Definiert als Quadratwurzel der Varianz
- Ist ein Maß dafür, wie weit die einzelnen Messwerte im Durchschnitt um den Mittelwert streuen
- Hat die gleiche Dimension wie die Daten und beschreibt diese daher intuitiver
- Geeignet für metrische Skalen, wird in der Praxis aber auch für quasi-metrische Daten verwendet (mind. 5 Abstufungen)

Streuungsmaße: Variationskoeffizient v

$$v = \frac{s}{\bar{x}}$$

- Bemisst die Streuung relativ zum Mittelwert ($\bar{x} \neq 0$)
- Kann auch in Prozent angegeben werden, d.h. als Anteil der Streuung am arithmetischen Mittel
  - z.B. bei $v = 0{,}5$ beträgt die Streuung 50 % des Mittelwerts
- Somit lassen sich Merkmale unterschiedlicher Größenordnung (verschiedene Skalen und Maßeinheiten) hinsichtlich ihrer Streuung vergleichen
- Aber: Voraussetzung ist Ratio-Skala

Streuungsmaße: Beispiele

```
# A tibble: 2 x 16
  Variable      N Missing     M    SD   Min   Q25   Mdn   Q75   Max Range CI_95_LL CI_95_UL Skewness Kurtosis Variance
  <chr>     <int>   <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
1 OFBesuche   126       0  8.36 12.3      0  1.25     3     8    69    69     6.20    10.5      2.49     9.56   150.
2 MassBier    109      17  1.79  1.26     0  1        2     3     6     6     1.55     2.03     0.629    3.17     1.58
```

1.Wie oft haben Sie – dieses Jahr eingenommen – das Oktoberfest besucht?
2. Wie viele Maß Bier haben Sie bei einem durchschnittlichen Wiesenbesuch Konsumiert?

```
# A tibble: 4 x 16
  Variable      N Missing     M    SD Min                 Q25   Mdn   Q75 Max                 Range CI_95_LL CI_95_UL Skewness Kurtosis Variance
  <chr>     <int>   <int> <dbl> <dbl> <dbl+lbl>           <dbl> <dbl> <dbl> <dbl+lbl>         <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
1 OFgut       126       0  2.77 0.997 0 [Stimme nicht zu]     2     3     4 4 [Stimme …         4     2.59     2.95   -0.450     2.58    0.995
2 OFsauf      126       0  2.66 1.16  0 [Stimme nicht zu]     2     3     4 4 [Stimme …         4     2.45     2.86   -0.543     2.52    1.35
3 OFlaut      126       0  1.79 1.25  0 [Stimme nicht zu]     1     2     3 4 [Stimme …         4     1.57     2.01    0.361     2.08    1.56
4 OFsauraus   126       0  1.92 1.32  0 [Stimme nicht zu]     1     2     3 4 [Stimme …         4     1.69     2.15    0.0623    1.86    1.74
```

1.Ich finde das Oktoberfest gut
2. Für meinen Geschmack wird zu viel getrunken
3. Das Oktoberfest ist mir zu laut
4. Ich lasse auf dem Oktoberfest gern auch mal die Sau raus

Übungsblatt 3: Aufgabe 1

- Eine kleine Befragung hat ergeben, dass Personen nur eine begrenzte Anzahl an Fernsehsendern nutzen. Die gemessenen Werte für die Anzahl der Sender sind wie folgt (geordnet):

  1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
  3, 3, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 6, 6, 7, 9, 10, 15, 25

- Bestimmen Sie (ohne R)...
  - Spannweite
  - Quartile und Interquartilabstand
  - Varianz und Standardabweichung
  - 80% Perzentil