Korpuslinguistik und Statistik

(050370)

Fabian Barteld, M.A.
Fabian.Barteld@ruhr-uni-bochum.de

9. Sitzung

Urliste

	Variable 1	Variable 2	...
Beobachtung 1	x1	y1	
Beobachtung 2	x2	y2	
...			...

Beispiel: Worthäufigkeiten

Rang	Häufigkeit	Wort
1	13	der
2	6	den
3	6	und
4	5	Wörter
5	4	Altmann
...		
186	1	ältesten
187	1	überarbeitete
188	1	Übersetzungen

(http://de.wikipedia.org/wiki/Worth%C3%A4ufigkeit, 06.01.2013)

Beschreibende Statistik

- Ziel: Daten übersichtlich aufbereiten
- Methoden
  - Verteilung
  - Diagramme
  - Kennzahlen

(empirische) Verteilung

Häufigkeiten der Ausprägungen einer Variable

Beispiel: Ergebnisse einer Wahl

Person	Peter	Anne
Stimmen	16	19

einfache Darstellung: Tabelle

Variable	Ausprägung 1	Ausprägung 2	...
Häufigkeit	x	y	

Beispiel: Verteilung von Worthäufigkeiten

Worthäufigkeit	1	2	3	4	5	6	13
Häufigkeit	143	25	9	7	1	2	1

Vergleich von Verteilungen

Frage:
Kommen *hapax legomena* in dem Wikipedia-Artikel zu „Worthäufigkeit“ und in der „Göttlichen Komödie“ gleich häufig vor?

	Anzahl der *hapax legomena*
Wikipedia-Artikel	143
Göttliche Komödie	8.246

Absolute und relative Häufigkeiten

hapax legomena

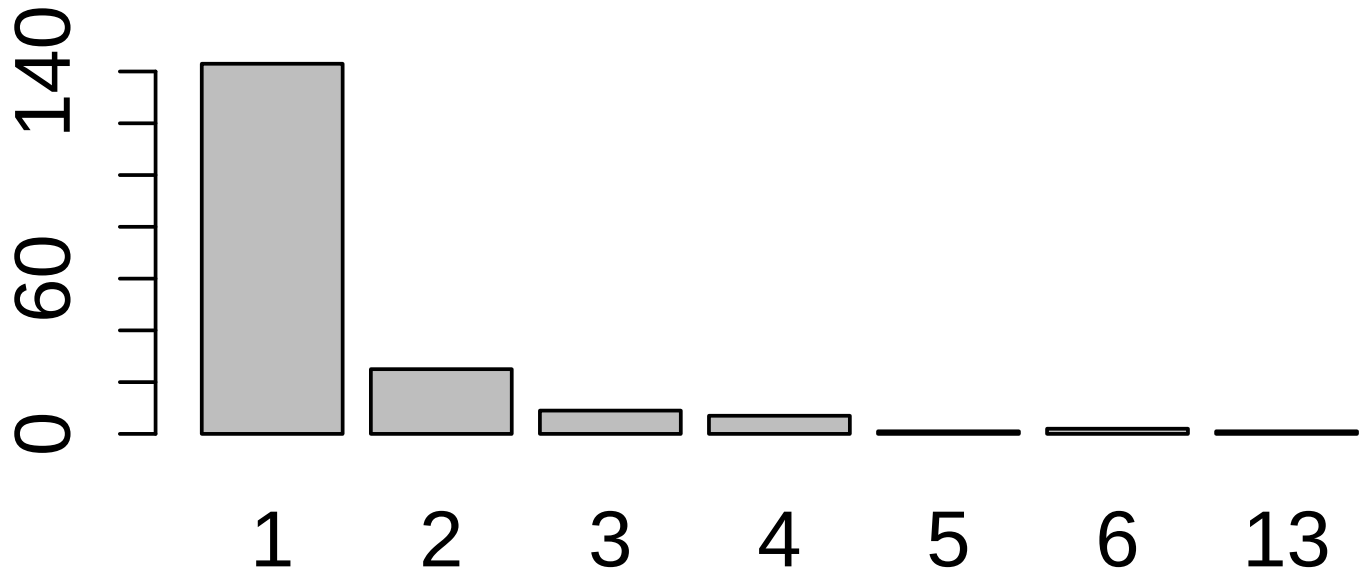
	abs.	Types	rel.
Wikipedia-Artikel	143	188	0,76
Göttliche Komödie	8.246	14.565	0,57

Relative Häufigkeiten

$$\text{relative Häufigkeit} = \frac{\text{absolute Häufigkeit}}{\text{Anzahl der Beobachtungen}}$$

$$\text{Prozent} = \text{relative Häufigkeit} * 100$$

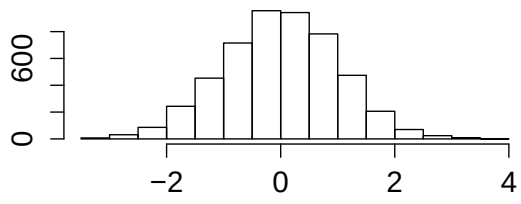
Darstellung von Verteilungen: Balkendiagramm

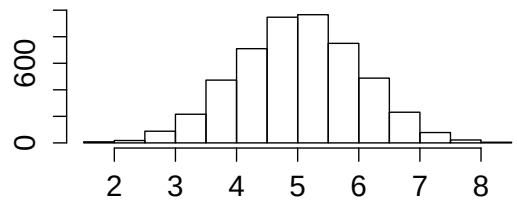


Kennzahlen

Ziel:
Reduktion einer Verteilung auf möglichst wenige, aussagekräftige Zahlen

Mittelwert





Drei Mittelwerte

1. Modalwert (Mode)
   - häufigster Wert
2. Median (Median)
   - mittlerer Wert
   - komparative oder metrische Daten
3. arithmetisches Mittel (Mean)
   - $\frac{\text{Beobachtung 1}+\text{Beobachtung 2}+\text{Beobachtung 3}+\ldots+\text{Beobachtung n}}{n}$
   - metrische Daten

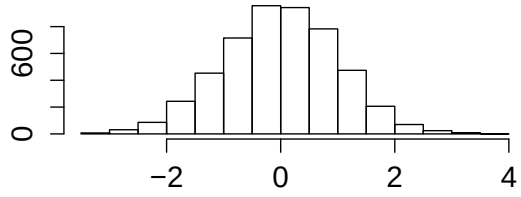
Robustheit

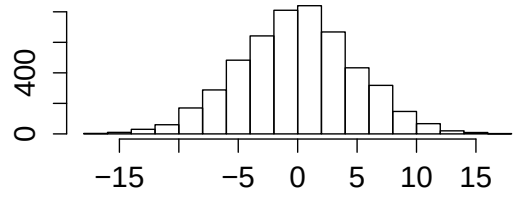
Der Median wird nicht so stark von Ausreißern beeinflusst wie das arithmetisch Mittel.

Worthäufigkeiten in der „Göttlichen Komödie“

Minimum	1
Maximum	2178
Median	1
arithmetisches Mittel	7,2

Streuungsmaß





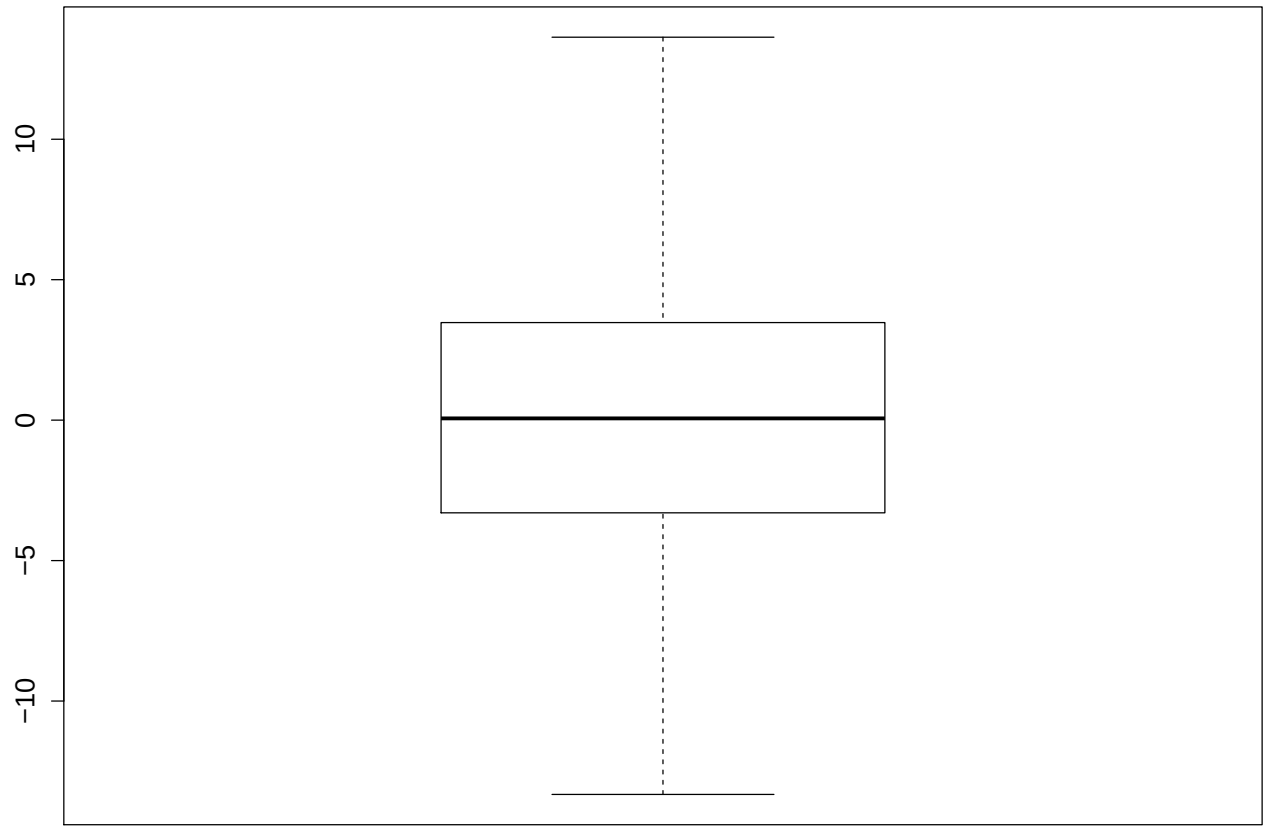
Drei Streuungsmaße

1. Quartile (Quartile)
   - Median gibt den mittleren Wert an
     also 50% der Daten links und 50% rechts
     Quartile geben dies für 25% und 75% an
2. Spannweite (Range)
   - Maximum − Minimum
   - metrische Daten
3. Varianz (Variance)
   - metrische Daten

Varianz (s^2)

1. Abweichung einer Beobachtung vom Mittelwert
   Beobachtung $-$ Mittelwert
   kann positiv oder negativ sein
   es interessiert aber nur die Größe der Abweichung, nicht die Richtung
2. Quadrat der Abweichung
   $(\text{Beobachtung} - \text{Mittelwert})^2$
3. „arithmetisches Mittel“ dieser Abweichungen
   $s^2 = \frac{1}{n} \sum (\text{Beobachtung} - \text{Mittelwert})^2$
   Standardabweichung (Standard Deviation)
   $s = \sqrt{\text{Varianz}}$

Darstellung von Kennzahlen: Boxplot



Gruppenarbeit

- säubern Sie Ihre Urliste
- verschaffen Sie sich mithilfe von „Sofa“ und der vorgestellten Statistiken einen Überblick über Ihre Daten

Literatur

- Worthäufigkeiten:
  Baayen, R. H. (2001). *Word frequency distributions*. Text, Speech and Language Technology. Dordrecht, Boston und London: Kluwer Academic Publishers
- Diagramme:
  10. Graphics in: Good, P. I. & Hardin, J. W. (2009). *Common errors in statistics (and how to avoid them)* (3rd edition). Hoboken, New Jersey: Wiley
- Krämer, W. (2011). *So lügt man mit Statistik* (2. Auflage). München: Piper