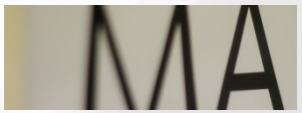
INSTITUT FÜR DEUTSCHE SPRACHE



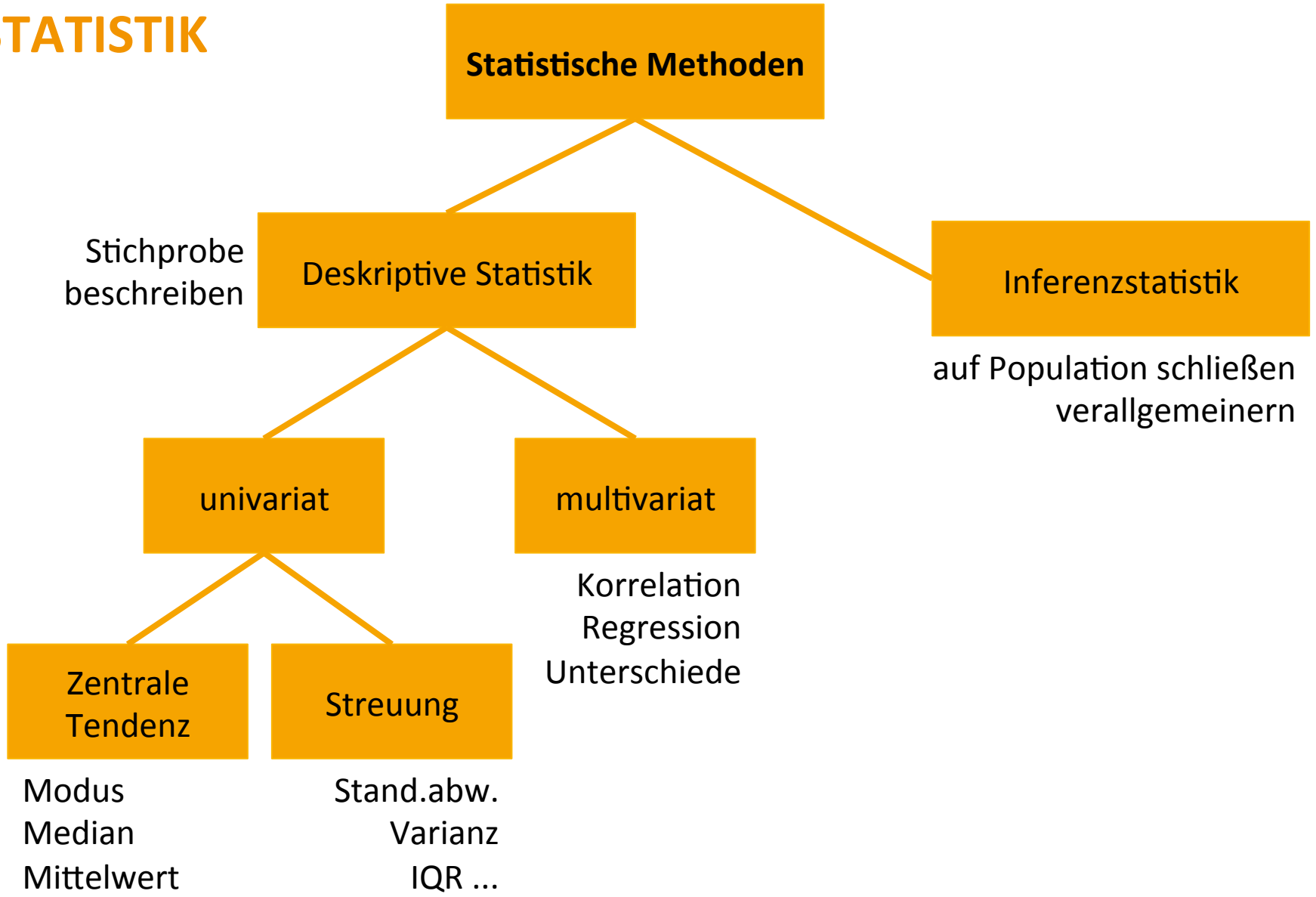
Sandra Hansen-Morath
Sascha Wolfer

STATISTIK MIT R

Maße der zentralen Tendenz & Streuungsmaße

Mitglied der Leibniz-Gemeinschaft

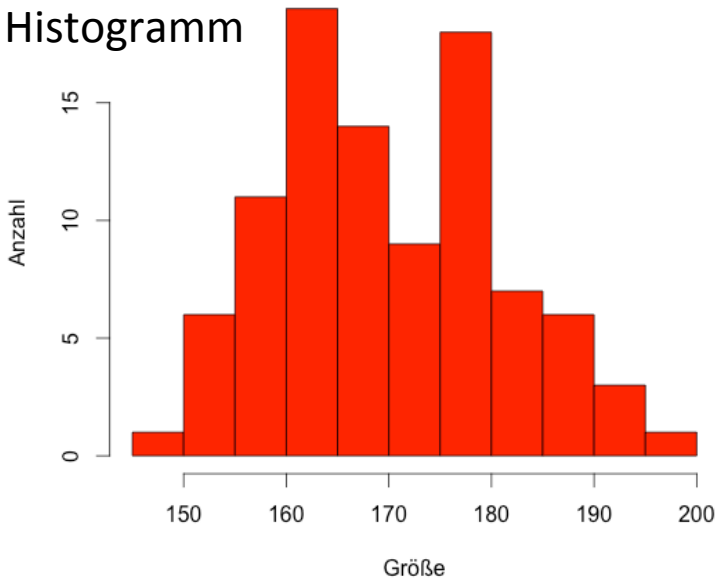
STATISTIK


Statistische Methoden
Stichprobe beschreiben
Deskriptive Statistik
Inferenzstatistik
auf Population schließen verallgemeinern
univariat
multivariat
Korrelation
Regression
Unterschiede
Zentrale Tendenz
Streuung
Modus
Median
Mittelwert
Stand.abw.
Varianz
IQR ...

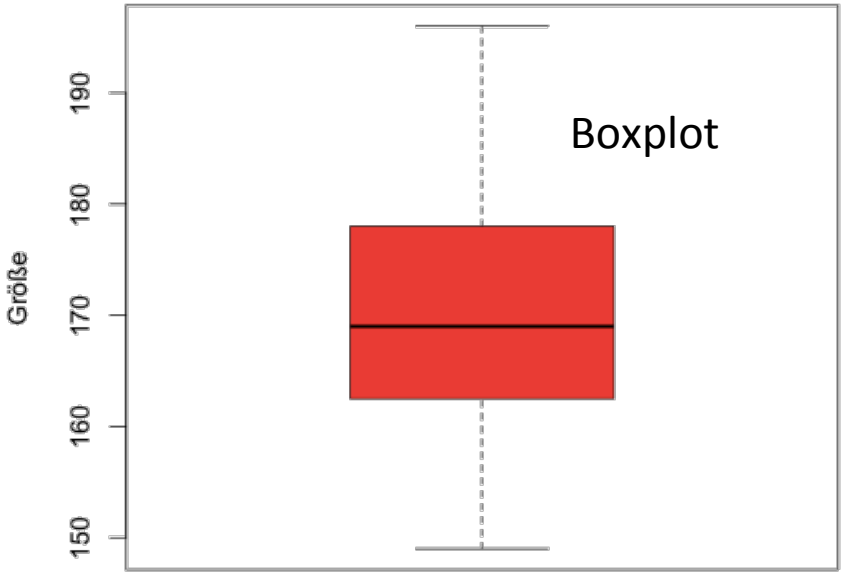
ZENTRALE TENDENZ

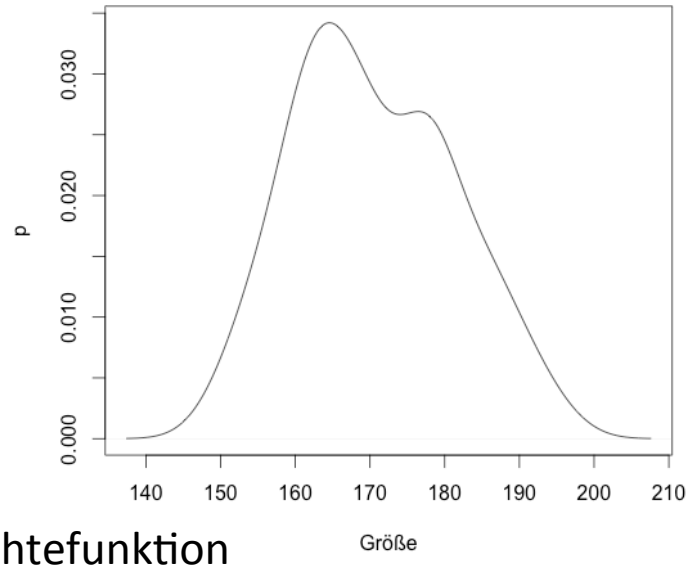
- Welche Möglichkeiten haben wir, die „Mitte“ eines Datensatzes zu beschreiben?
- Statistisch gesprochen: Wie können wir die zentrale Tendenz einer Verteilung beschreiben?
- Beispiel: Wir fragen 95 Personen nach ihrer Körpergröße.
  - Wir erhalten also eine bestimmte Verteilung von Körpergrößen.

Histogramm



Boxplot





Dichtefunktion

```
14 | 9
15 | 2234
15 | 55677899
16 | 0000011222233334444
16 | 5555566778888999
17 | 0001111223
17 | 55666777788888899
18 | 00012244
18 | 5567777
19 | 0122
19 | 6
```

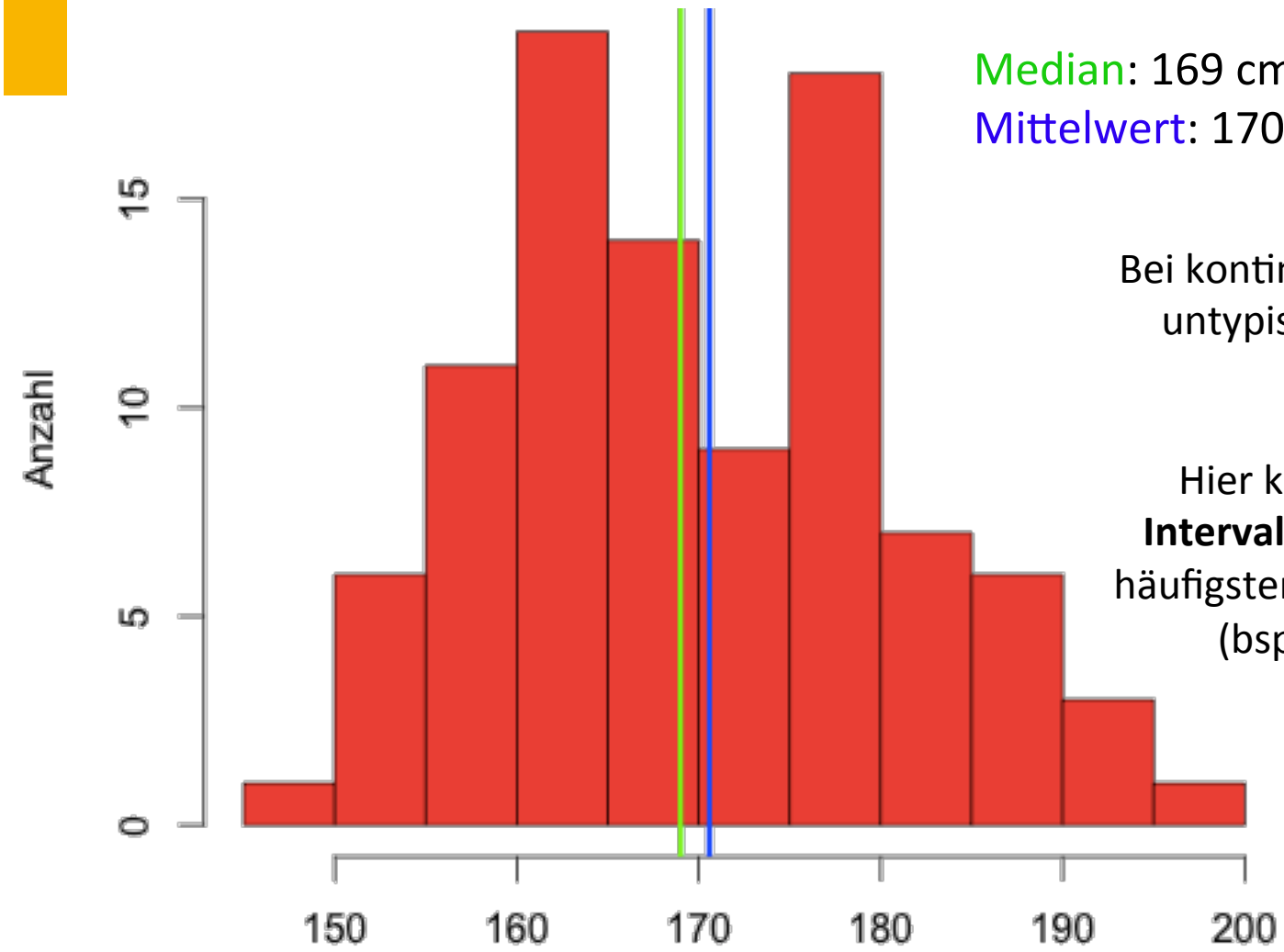
Stengel-Blatt-Diagramm

ZENTRALE TENDENZ

- Visualisieren ist eine gute Idee, aber wie können wir Kennwerte für die zentrale Tendenz unserer Verteilung berechnen?
- Die drei wichtigsten Maße:
  - Modalwert / Modus: Der am häufigsten vorkommende Wert
  - Median: Der Wert, der die Verteilung in zwei Hälften teilt
  - Arithmetischer Mittelwert ($\bar{x}$, „x quer“)
    - Summe aller Messwerte durch Anzahl Messwerte
    - Summe der Abweichungen ist 0.

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

ZENTRALE TENDENZ UNSERER VERTEILUNG



Median: 169 cm
Mittelwert: 170,6 cm

Bei kontinuierlichen Daten ist es untypisch einen Modalwert anzugeben.

Hier könnte man eher ein **Intervall** angeben, in dem am häufigsten Messwerte auftreten. (bspw. $160 < x < 165$)

ZENTRALE TENDENZ UND SKALENNIVEAUS

- In unserer Größen-Stichprobe sind die Daten von 45 Männern und 50 Frauen enthalten.
- Modalwert der Variable „biologisches Geschlecht“?
  - „weiblich“ mit *n* = 50 Fällen
- Mittelwert der Variable Geschlecht?

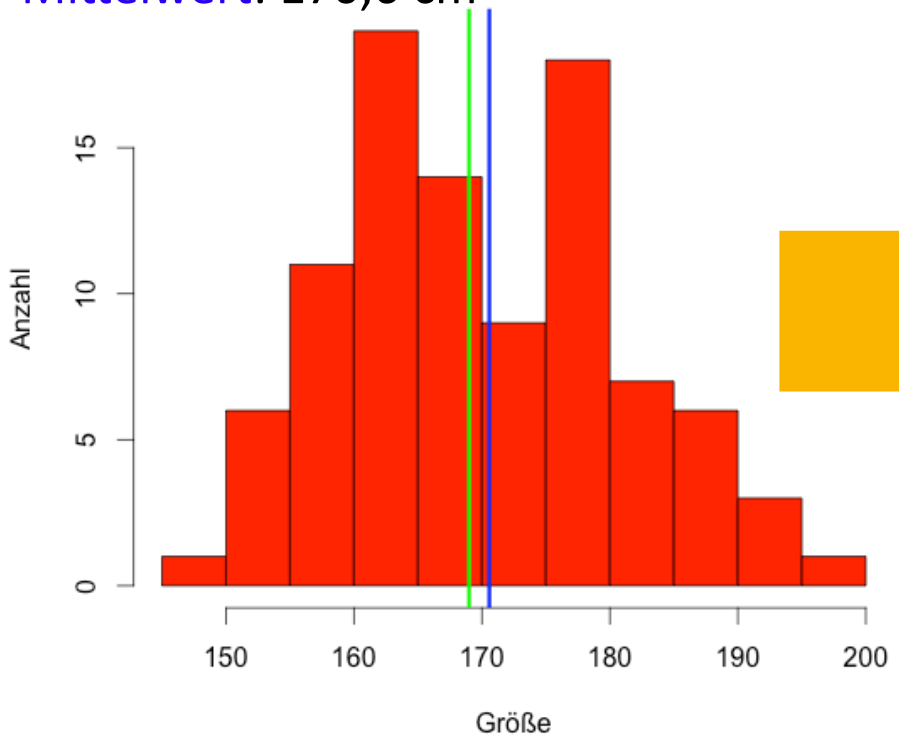
ZENTRALE TENDENZ UND SKALENNIVEAUS

- Mittelwert der Variable Geschlecht ist sinnlos ...
  - ... weil es auf die Kodierung der Variable ankommt, welchen Wert der Mittelwert annimmt.
  - ... weil ein „mittleres Geschlecht“ von (bspw.) 0,73 kein sinnvoller Wert ist.
- Je höher das Skalennivau, desto mehr Maße können berechnet werden:
  - Nominalskala: Modalwert
  - Ordinalskala: Modalwert, Median
  - Kardinalskala: (Modalwert), Median, Mittelwert

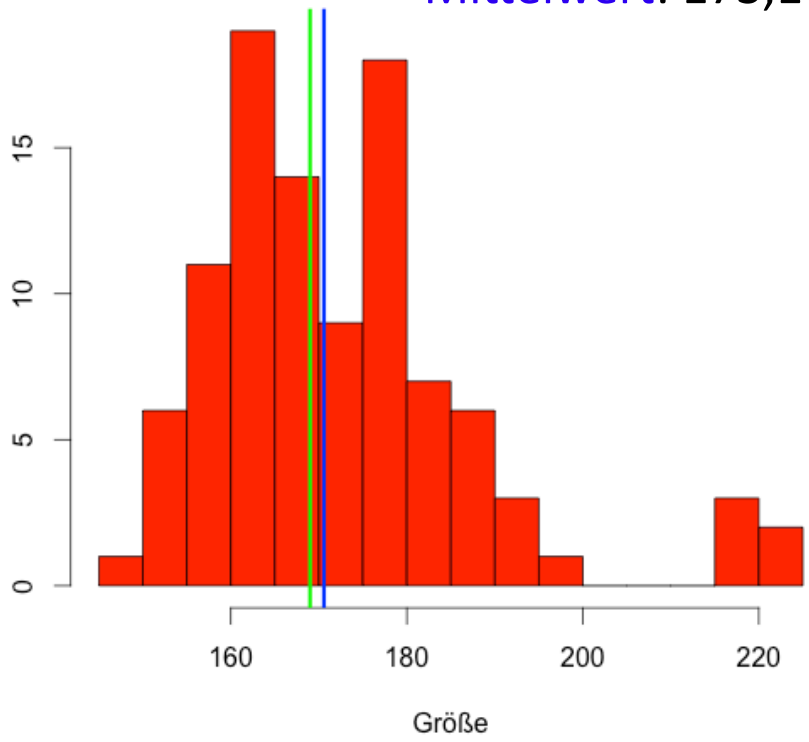
MITTELWERT UND MEDIAN

- Der Median ist weniger anfällig für Ausreißerwerte als der Mittelwert. Hier werden 5 hohe Werte hinzugefügt.

Median: 169 cm
Mittelwert: 170,6 cm

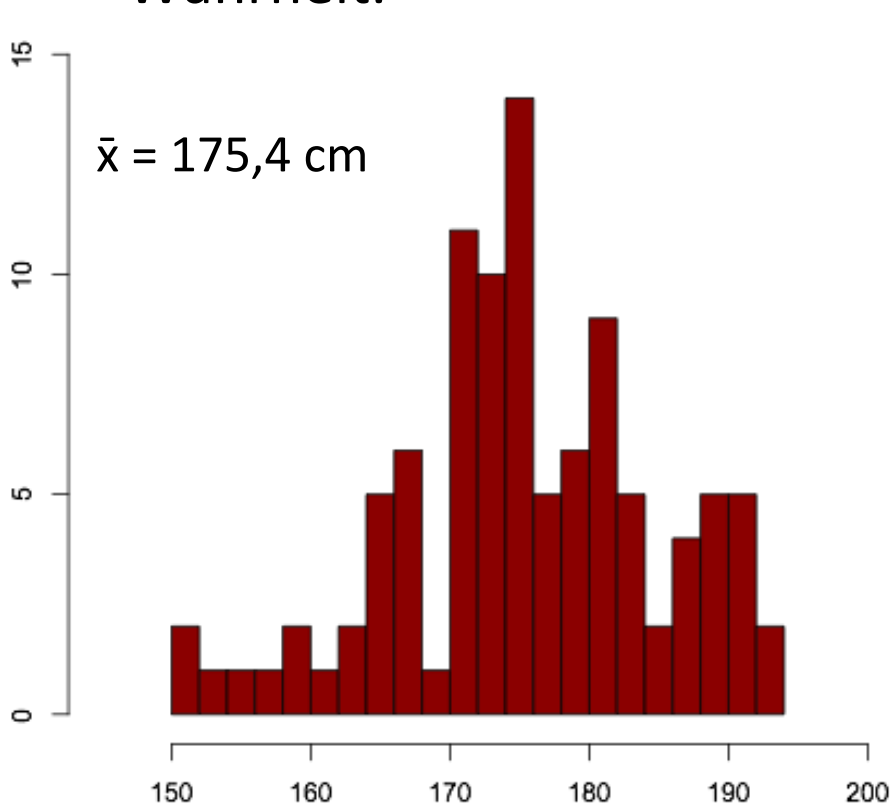


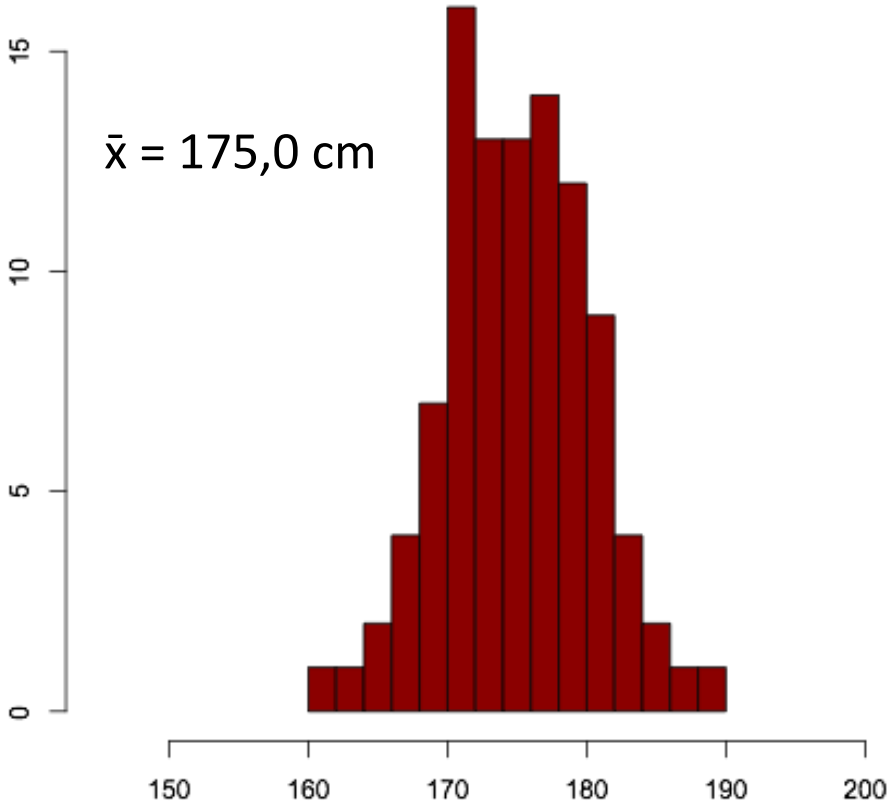
Median: 170 cm
Mittelwert: 173,1 cm



BESCHREIBEN VON VERTEILUNGEN

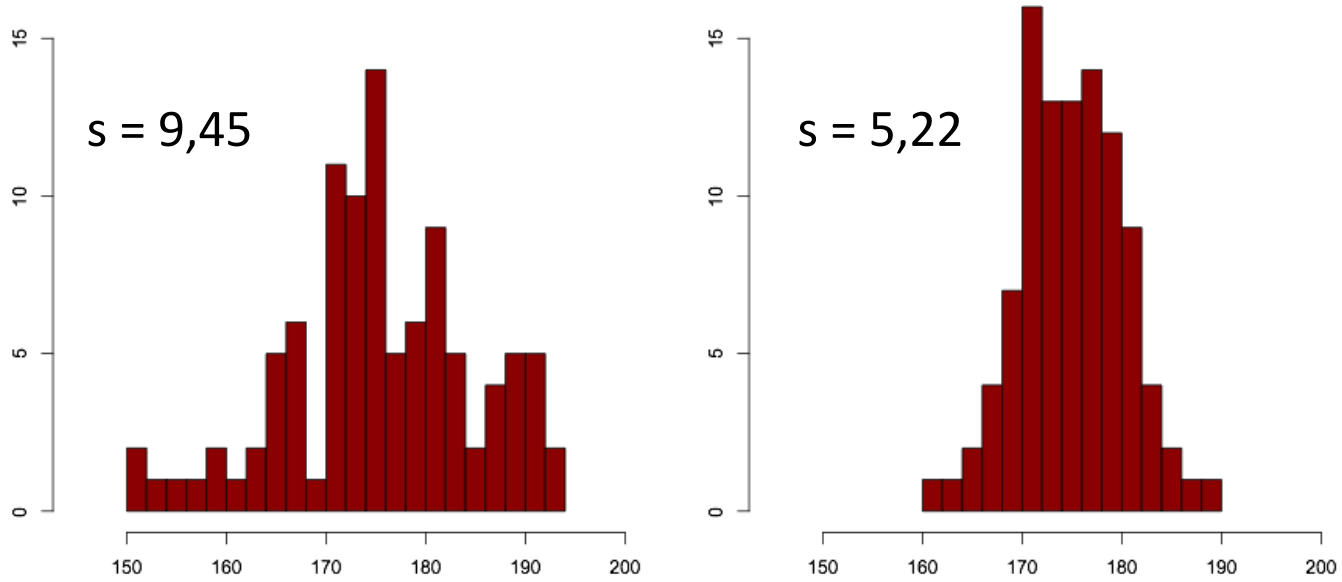
- Die zentrale Tendenz einer Verteilung ist nicht die ganze Wahrheit.





STREUUNG

- Um eine Verteilung zu beschreiben, müssen wir auch die Streuung der Messwerte beachten!
- Für metrisch skalierte Daten ist hier die **Standardabweichung** *s* das wichtigste Maß.



STANDARDABWEICHUNG

- Die Standardabweichung ist
  - die Summe der *quadrierten* Abweichungen vom Mittelwert,
  - geteilt durch die Größe der Stichprobe,
  - und daraus die Quadratwurzel.
- Große Abweichungen werden stärker bestraft!
- s^2: Varianz

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

DISPERSIONSMAßE

VARIANZ

- Die Varianz kennzeichnet die Verteilung von Werten um den Mittelwert
- Varianz = Summe der quadrierten Abweichungen aller Messwerte vom arithmetischen Mittel dividiert durch die Anzahl der Messwerte

Bsp. Länge von Nominalphrasen

5 3 6 2 3 1 3 2 2 24

BEISPIEL: NP AUS EINEM GERICHTSURTEIL DES BUNDESVERFASSUNGSGERICHTS

Bei **der Umsetzung der Vorgaben der Gerichte für eine verfassungskonforme Regelung der Überführung von Ansprüchen und Anwartschaften aus den Zusatz- und Sonderversorgungssystemen der ehemaligen DDR** lässt sich der Gesetzgeber von der befriedenden Wirkung dieser Entscheidungen leiten.

DISPERSIONSMAßE
VARIANZ

- Die Varianz kennzeichnet die Verteilung von Werten um den Mittelwert
- Varianz = Summe der quadrierten Abweichungen aller Messwerte vom arithmetischen Mittel dividiert durch die Anzahl der Messwerte

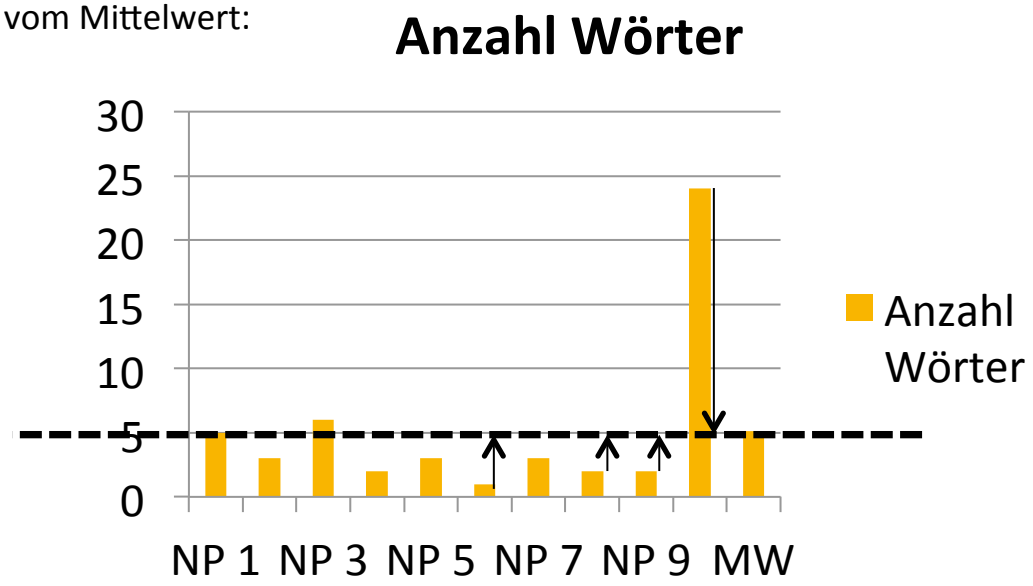
Bsp. Länge von Nominalphrasen

5 3 6 2 3 1 3 2 2 24

Mittelwert = 5,1

Abweichungen der einzelnen Messwerte vom Mittelwert:

1-5,1 = -4,1
2-5,1 = -3,1
2-5,1 = -3,1
2-5,1 = -3,1
3-5,1 = -2,1
3-5,1 = -2,1
3-5,1 = -2,1
5-5,1 = -0,1
6-5,1 = 0,9
24-5,1= 18,9



DISPERSIONSMAßE
VARIANZ

Abweichungen der einzelnen Messwerte vom Mittelwert:

1-5,1 = -4,1	→ quadriert = 16,81
2-5,1 = -3,1	→ quadriert = 9,61
2-5,1 = -3,1	→ quadriert = 9,61
2-5,1 = -3,1	→ quadriert = 9,61
3-5,1 = -2,1	→ quadriert = 4,41
3-5,1 = -2,1	→ quadriert = 4,41
3-5,1 = -2,1	→ quadriert = 4,41
5-5,1 = -0,1	→ quadriert = 0,01
6-5,1 = 0,9	→ quadriert = 0,81
24-5,1= 18,9	→ quadriert = 357,21

→ Summe = 416,9

→ Varianz = 416,9 / 10 = 41,69 Wörter²

Nachteil der Varianz:
- Aufgrund der Quadrierung andere Einheit als beobachtete Messwerte
- Keine konkreten Aussagen über die Streuungsbreite

Lösung:
- Standardabweichung

→ Ziehen der Quadratwurzel aus der Varianz

Standardabweichung = 6,46

DISPERSIONSMAßE
STANDARDABWEICHUNG

= Wurzel der Varianz

- Ist das meist verbreitete Streuungsmaß
- Standardabweichung besitzt immer die gleiche Maßeinheit wie das zu untersuchende Merkmal → Interpretation einfacher
- Kleinere Standardabweichung → gemessene Ausprägungen liegen eher enger um den Mittelwert; größer Standardabweichung → stärkere Streuung

Bsp.: Nominalphrasen

- Standardabweichung mit Ausreißer: **6,46**
- Standardabweichung ohne Ausreißer: **2,57**

1-5,1 = -4,1 → quadriert = 16,81
2-5,1 = -3,1 → quadriert = 9,61
2-5,1 = -3,1 → quadriert = 9,61
2-5,1 = -3,1 → quadriert = 9,61
3-5,1 = -2,1 → quadriert = 4,41
3-5,1 = -2,1 → quadriert = 4,41
3-5,1 = -2,1 → quadriert = 4,41
5-5,1 = -0,1 → quadriert = 0,01
6-5,1 = 0,9 → quadriert = 0,81

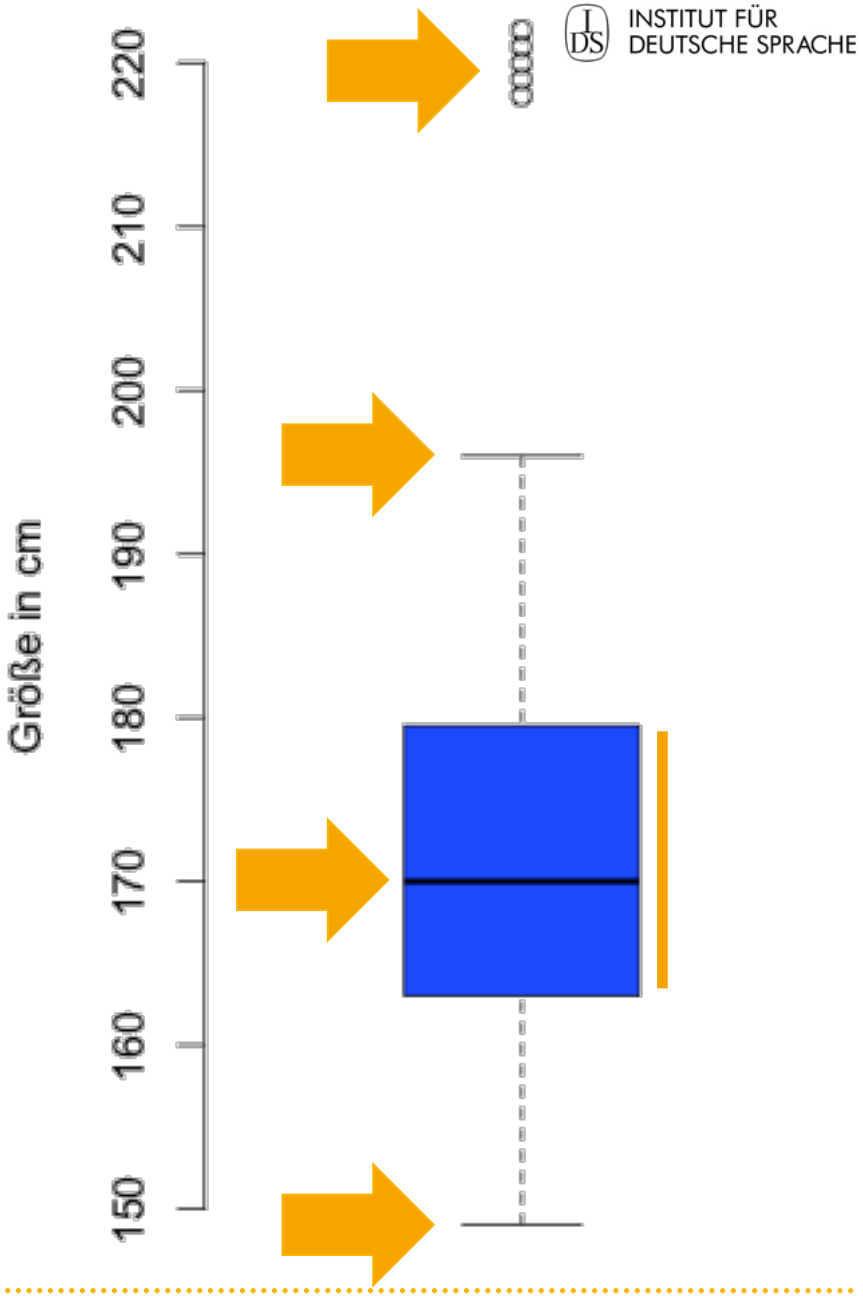
→ Summe = 59,69
→ Varianz = 59,69 / 9 = 6,63 Wörter²

STREUUNG

- Standardabweichung s
- Varianz s^2
- Range/Spannweite: max(x) – min(x)
- Perzentil/Quantil: Das x-te Perzentil ist jener Punkt der Daten, unter dem x Prozent aller Werte liegen.
- Quartile: Das 25., 50. und 75. Perzentil (trennen die Daten in Viertel)
- Interquartilabstand: Mittlere 50% der Daten

BOXPLOTS

- Boxplots bestehen aus den folgenden Elementen:
  - Median
  - Box: Interquartilabstand (1. bis 3. Quartil), Grenzen der Box: „Hinges“
  - „Whiskers“: Box +/- 1,5x Boxhöhe bis zum maximal zulässigen Wert
  - Outlier (falls vorhanden)



FUNKTIONEN IN R

x: numerischer Vektor (nur Zahlen)

- Median: `median(x)`
- Mittelwert: `mean(x)`
- Standardabweichung: `sd(x)`
- Varianz: `var(x)`
  - Ausprobieren: `sqrt(var(x))` ist gleich `sd(x)` ?!?
- Perzentile/Quantile: `quantile(x)`
- Histogramm: `hist(x)`, `MASS::truehist(x)`
- Dichtefunktion: `plot(density(x))`
- Boxplot: `boxplot(x)`

BEGRIFFE

- Modus
- Mittelwert
- Median
- Intervall
- Streuung
- Standardabweichung
- Varianz
- Range
- Perzentil / Quantil
- Quartil
- Interquartilabstand
- Boxplot
- Outlier / Ausreißer