Crashkurs
Einführung Biostatistik

Prof. Burkhardt Seifert

Abteilung Biostatistik, ISPM
Universität Zürich

- **Deskriptive Statistik**
- **Wahrscheinlichkeitsrechnung, Versuchsplanung**
- Statistische Inferenz
  - Prinzip statistischer Tests
  - Konfidenzintervalle
  - Stichprobengrösse, Power
- Korrelation und einfache lineare Regression

Deskriptive Statistik

- Wie beschreibe ich meine Daten richtig?
- Wie visualisiere ich meine Daten?

Wikipedia

Die empirische Wahrscheinlichkeitsverteilung der Körpergrößen großer Gruppen entspricht der Gauß'schen Normalverteilung.

Durchschnittliche Körpergröße	**Männer**	**Frauen**
Deutschland	180,2 cm	168,3 cm
Österreich	178,2 cm	165,5 cm
Schweiz	180,5 cm	167,2 cm
Frankreich	175,6 cm	162,5 cm
Australien	177,0 cm	164,3 cm
Brasilien	174,0 cm	161,2 cm
Bosnien und Herzegowina	186,0 cm	170,7 cm
China	169,7 cm	158,6 cm

- Woher weiss man das?
- Sind deutsche Männer kleiner als Schweizer? Ist es bei den Frauen umgekehrt?

Grundgesamtheit und Stichprobe

- Daten kommen aus einer **Stichprobe**.
- Daten von Stichproben variieren.
- Aussagen macht man für eine **Grundgesamtheit** (Population).

> Die **Grundgesamtheit** ist die Gesamtheit aller Individuen, für welche Aussagen gemacht werden sollen.

> Eine **Stichprobe** aus einer Grundgesamtheit ist die Menge der Individuen, die tatsächlich beobachtet wurden.

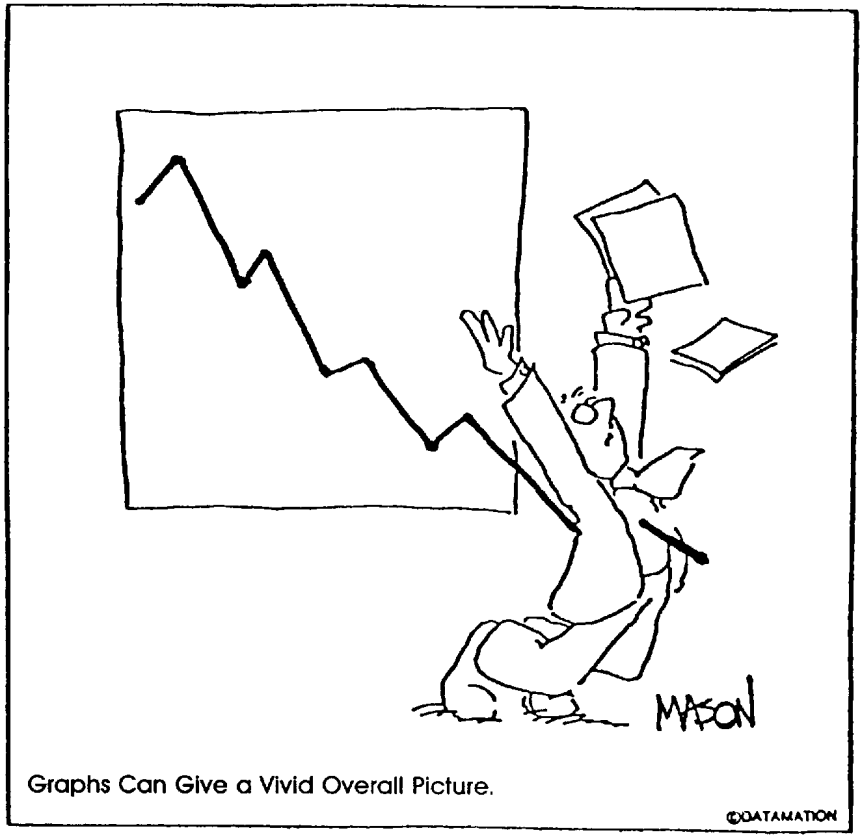
Beispiel:

Grundgesamtheit = alle Menschen (alle Schweizer)

Stichprobe = Studierende des 1. Studienjahres Human- und Zahnmedizin, die am Praktikum Statistik im Sommersemester 2006 teilgenommen haben

Deskriptive Statistik

- Daten mit wenigen charakteristischen Zahlen „gut“ beschreiben und visualisieren
  - durch statistische Kennwerte (Lage– und Streumasse)
  - durch Graphiken
- Ansatz „beschreibend“, ohne „Signifikanz“



Daten in einer Tabelle

SEX	Körpergrösse	HAND	GROUP	Tutor	Geschlecht
1	168	17.5	1	1	w
0	183.5	21	1	1	m
1	170	20	1	1	w
1	159	17	1	1	w
1	165	18	1	1	w
0	180	20	1	1	m
1	181	19.5	1	1	w
0	193	21.5	1	1	m
0	183	19.5	1	1	m
0	183	20.5	1	1	m
1	165	17.8	1	1	w
1	161	19.5	1	1	w
1	156	16.5	1	1	w
0	184	17	1	1	m
0	173	18.5	1	1	m
1	170	17.5	1	1	w
1	163	17.5	1	1	w
1	162	18	2	3	w
1	181	20.5	2	3	w
0	178	20	2	3	m
0	173	20	2	3	m
...	...	...	...	...	...

- insgesamt 245 Studenten in 16 Gruppen

Haupttypen von Daten

1) nominale, kategorielle Daten

- Zuordnung zu Kategorien
  $\longrightarrow$ Anzahlen und % sinnvoll
  Beispiele: Geschlecht, Blutgruppe

SEX	Körpergrösse	HAND	GROUP	Tutor	Geschlecht
1	168	17.5	1	1	w
0	183.5	21	1	1	m
1	170	20	1	1	w
1	159	17	1	1	w
1	165	18	1	1	w
0	180	20	1	1	m
1	181	19.5	1	1	w
0	193	21.5	1	1	m
0	183	19.5	1	1	m
0	183	20.5	1	1	m
1	165	17.8	1	1	w
1	161	19.5	1	1	w
1	156	16.5	1	1	w
0	184	17	1	1	m
0	173	18.5	1	1	m
1	170	17.5	1	1	w
1	163	17.5	1	1	w
1	162	18	2	3	w
1	181	20.5	2	3	w
0	178	20	2	3	m
0	173	20	2	3	m

- Output SPSS:

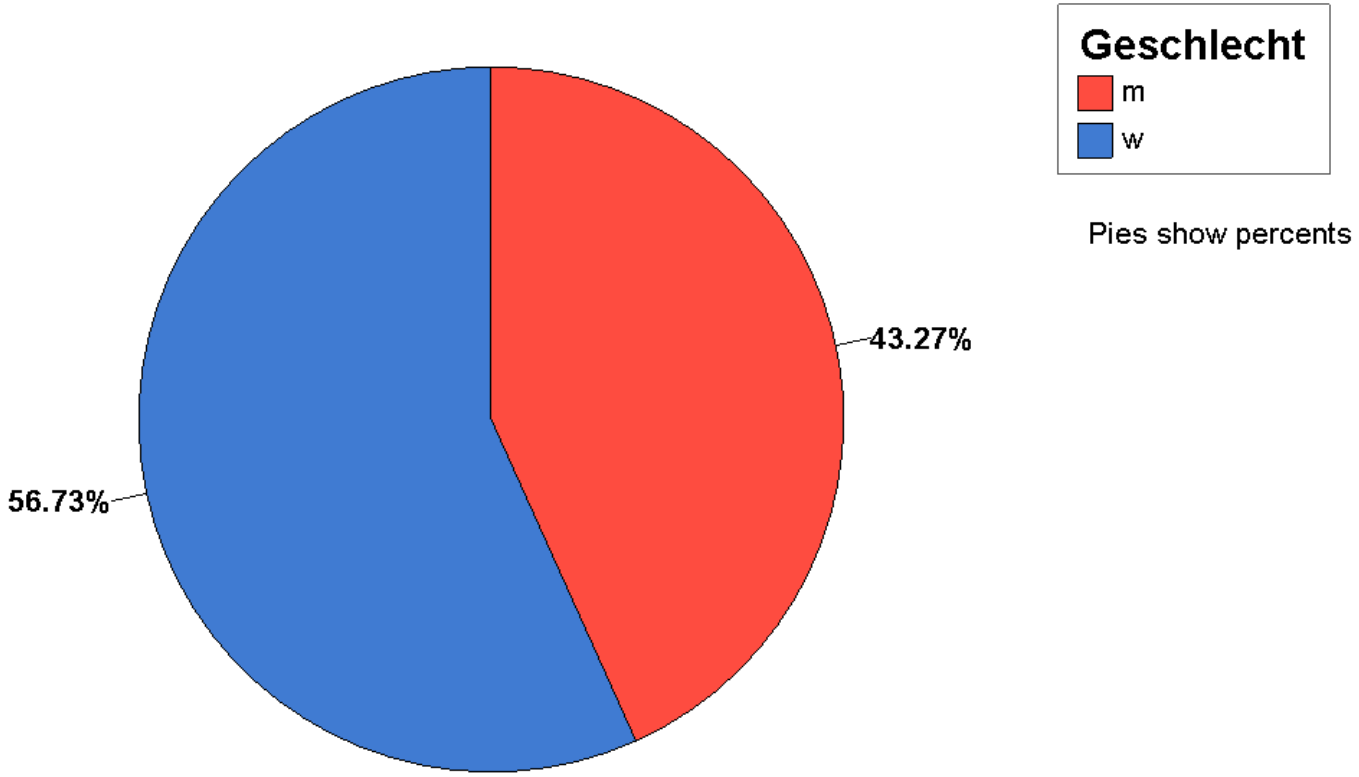
Geschlecht

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	m	106	43.3	43.3	43.3
	w	139	56.7	56.7	100.0
	Total	245	100.0	100.0	

1–2) ordinale Daten (geordnet kategoriell)

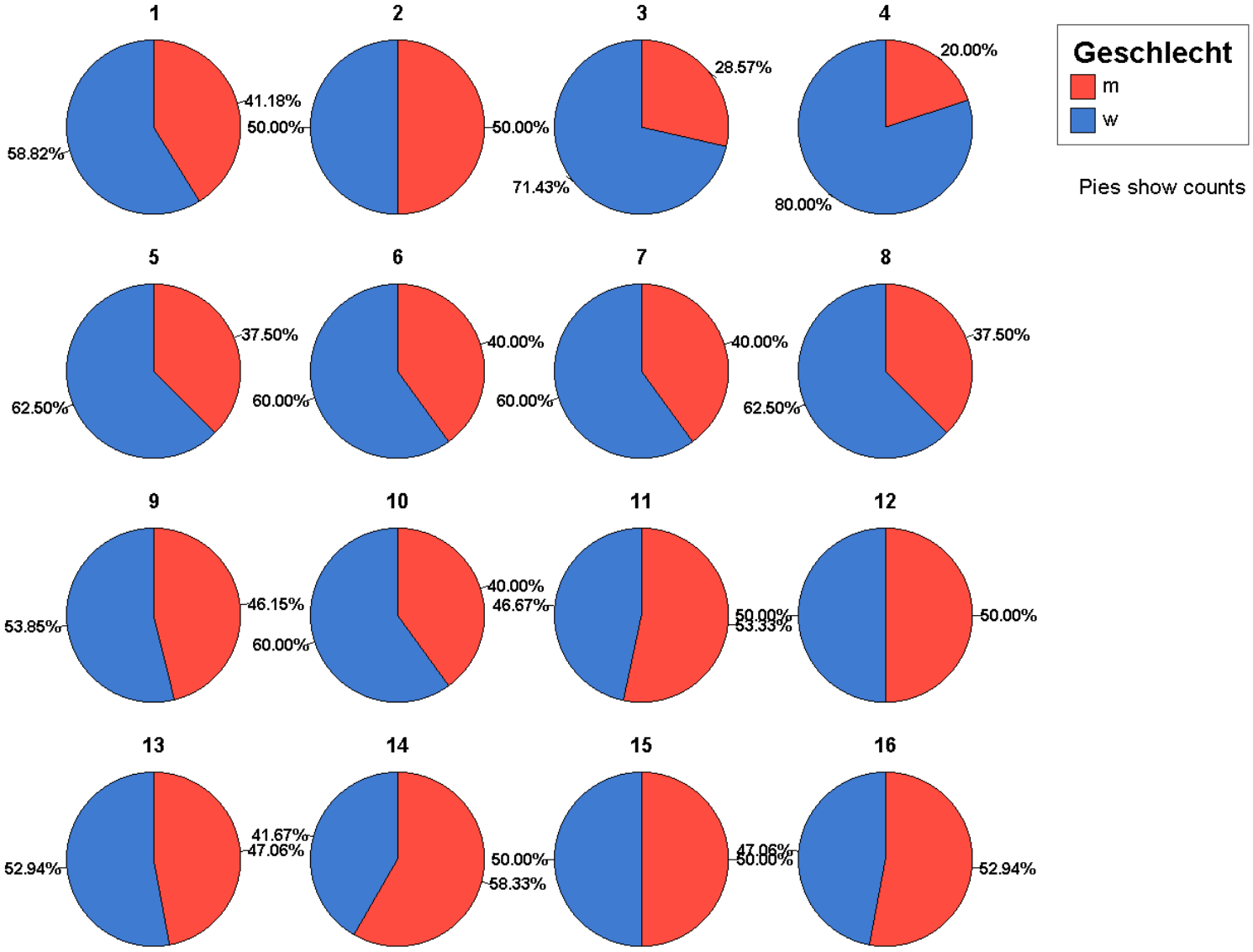
- haben Rangordnung
  Beispiel: Schweregrad einer Krankheit

Kuchendiagramm (piechart)

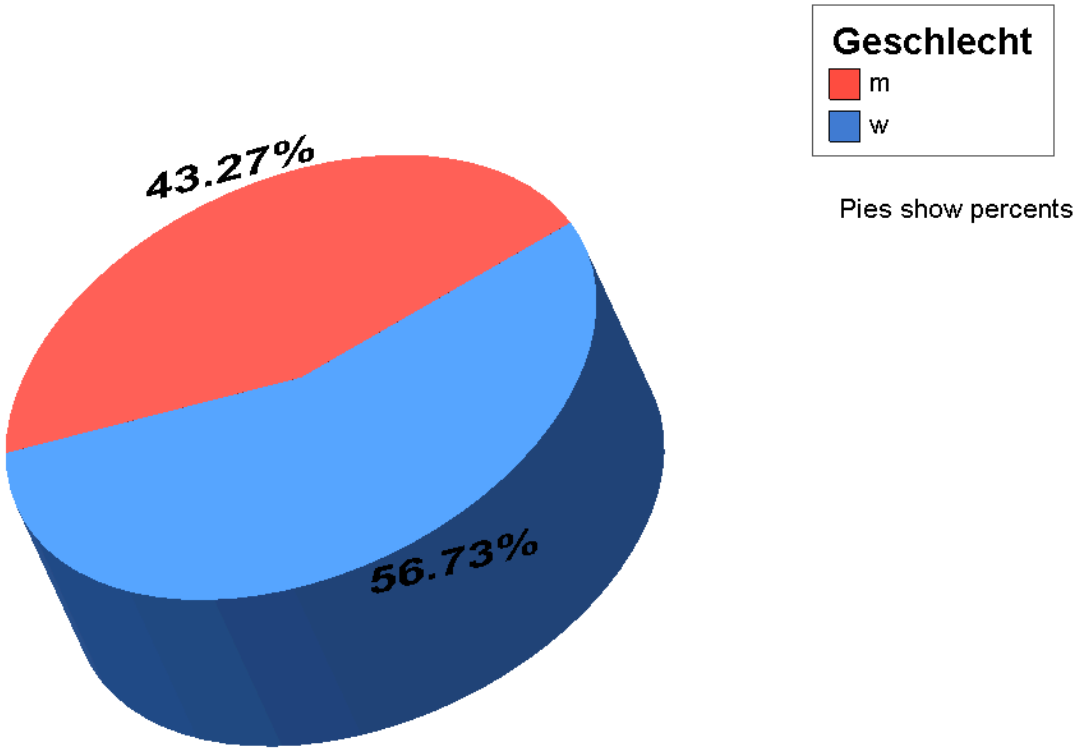


- Prozentzahlen ohne Dezimalstellen (maximal eine)
  1 StudentIn entspricht 0.4%!

Das Geschlechterverhältnis variiert von Stichprobe zu Stichprobe

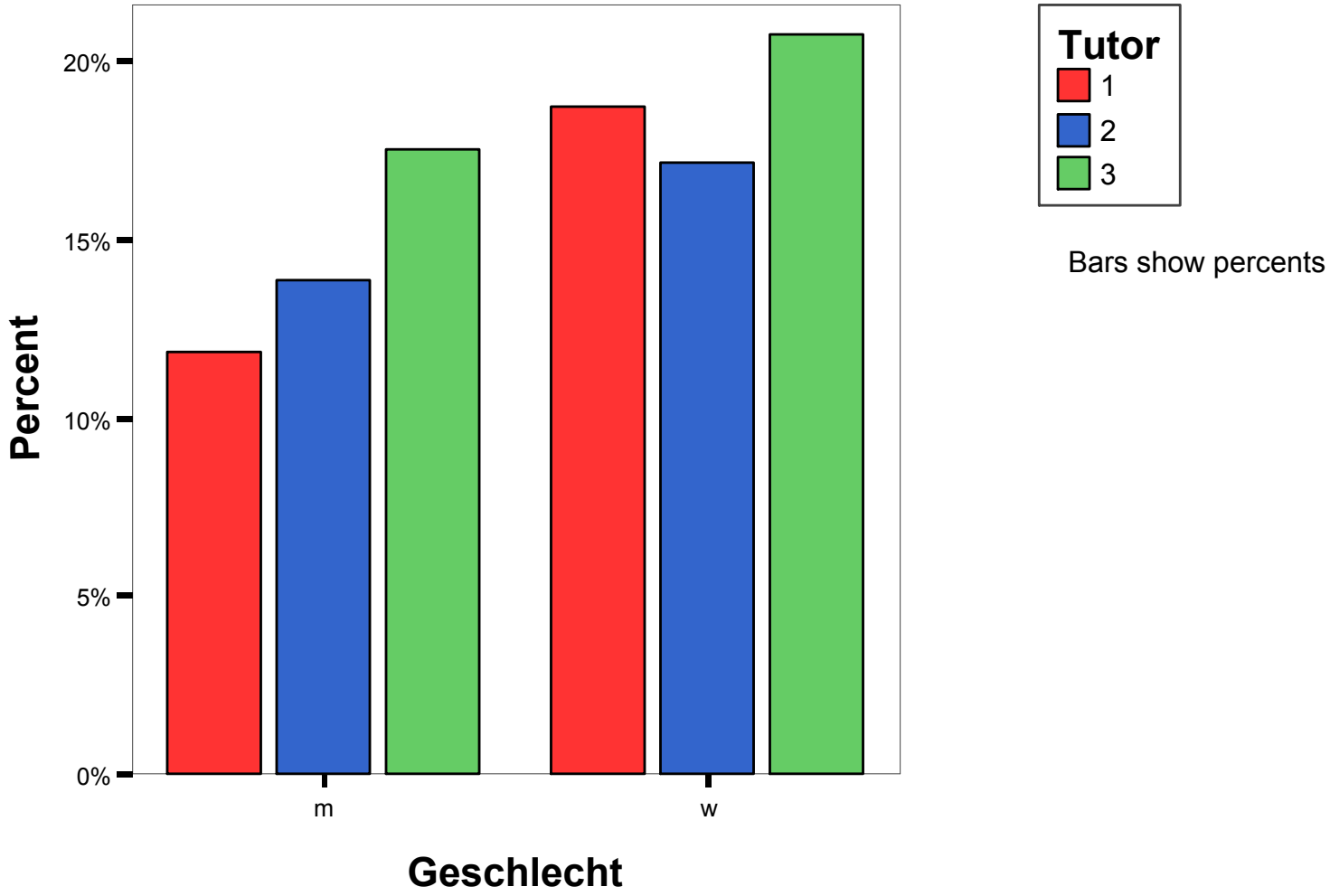


Kuchendiagramm (piechart)

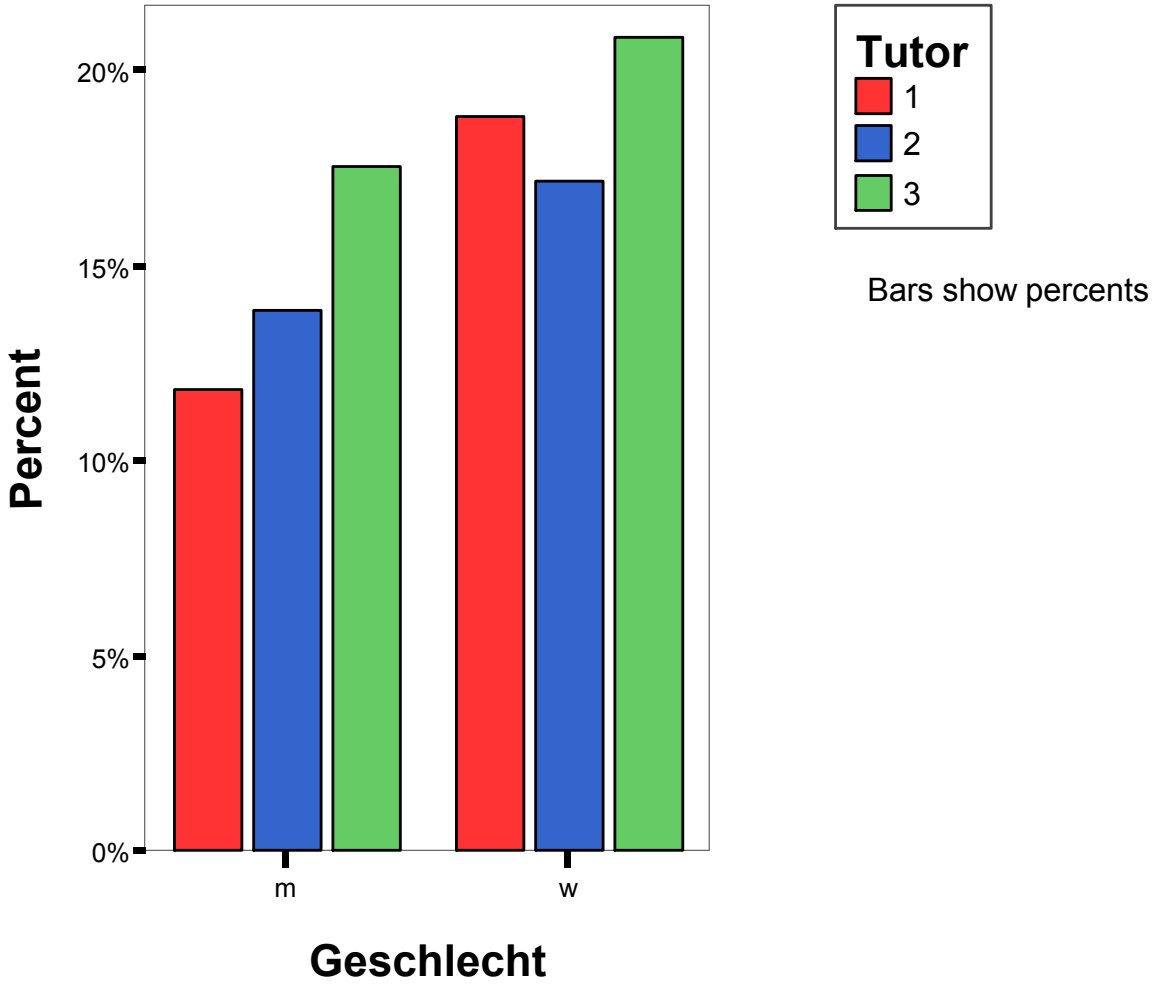


- Vorsicht vor 3–dimensionaler Darstellung

Balkendiagramm

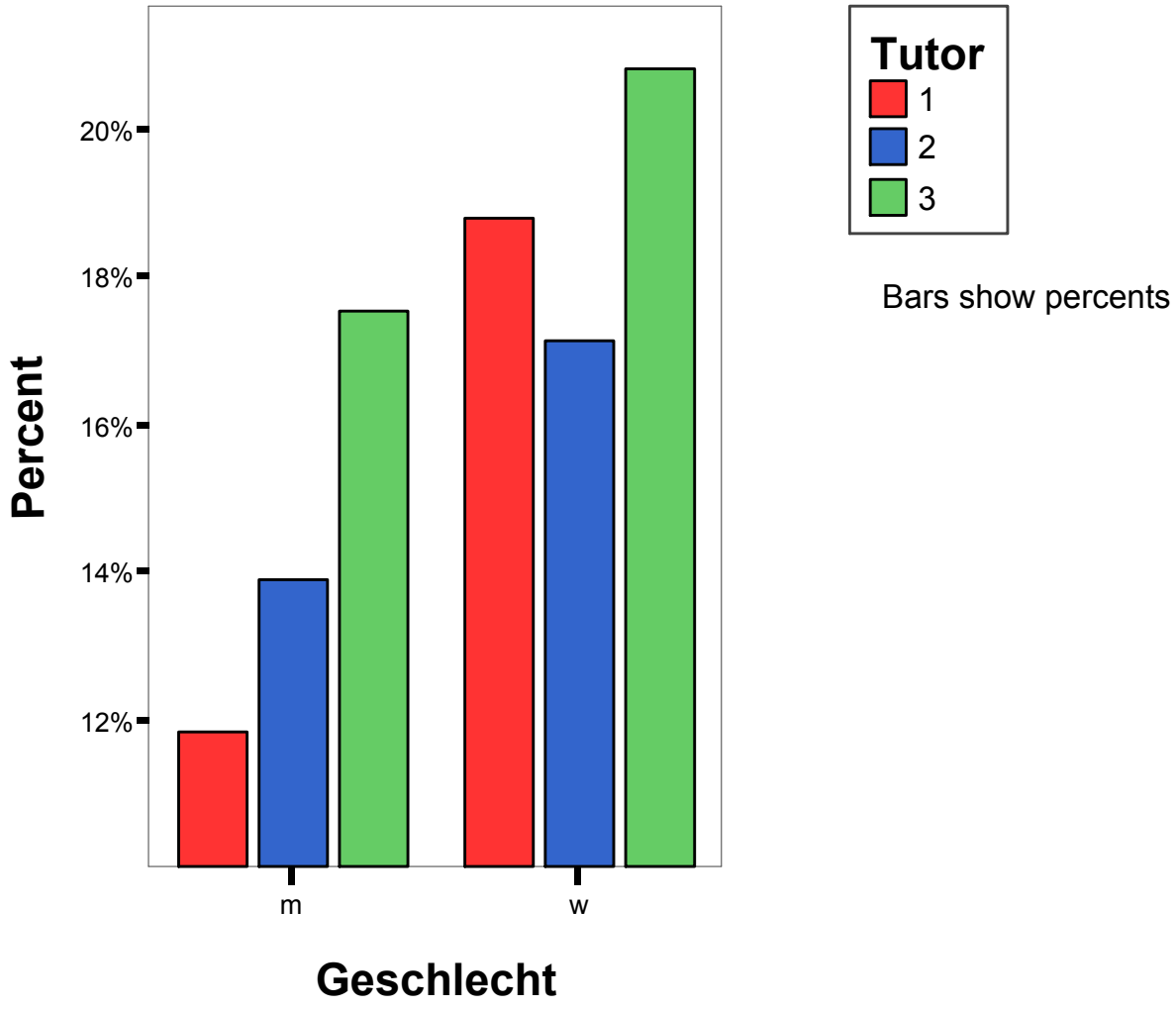


Balkendiagramm



- Traue keiner Graphik, die höher als breit ist: Durch das Strecken der y-Achse wird der Eindruck grosser Unterschiede erzeugt.

Balkendiagramm



- Traue keiner Graphik, die höher als breit ist.
- Balken stehen auf dem Boden, deshalb Nullpunkt beachten.

Haupttypen von Daten

2) stetige (numerische) Messdaten

- Differenzen und Mittelwerte sinnvoll
  Beispiel: Temperatur in Grad Celsius

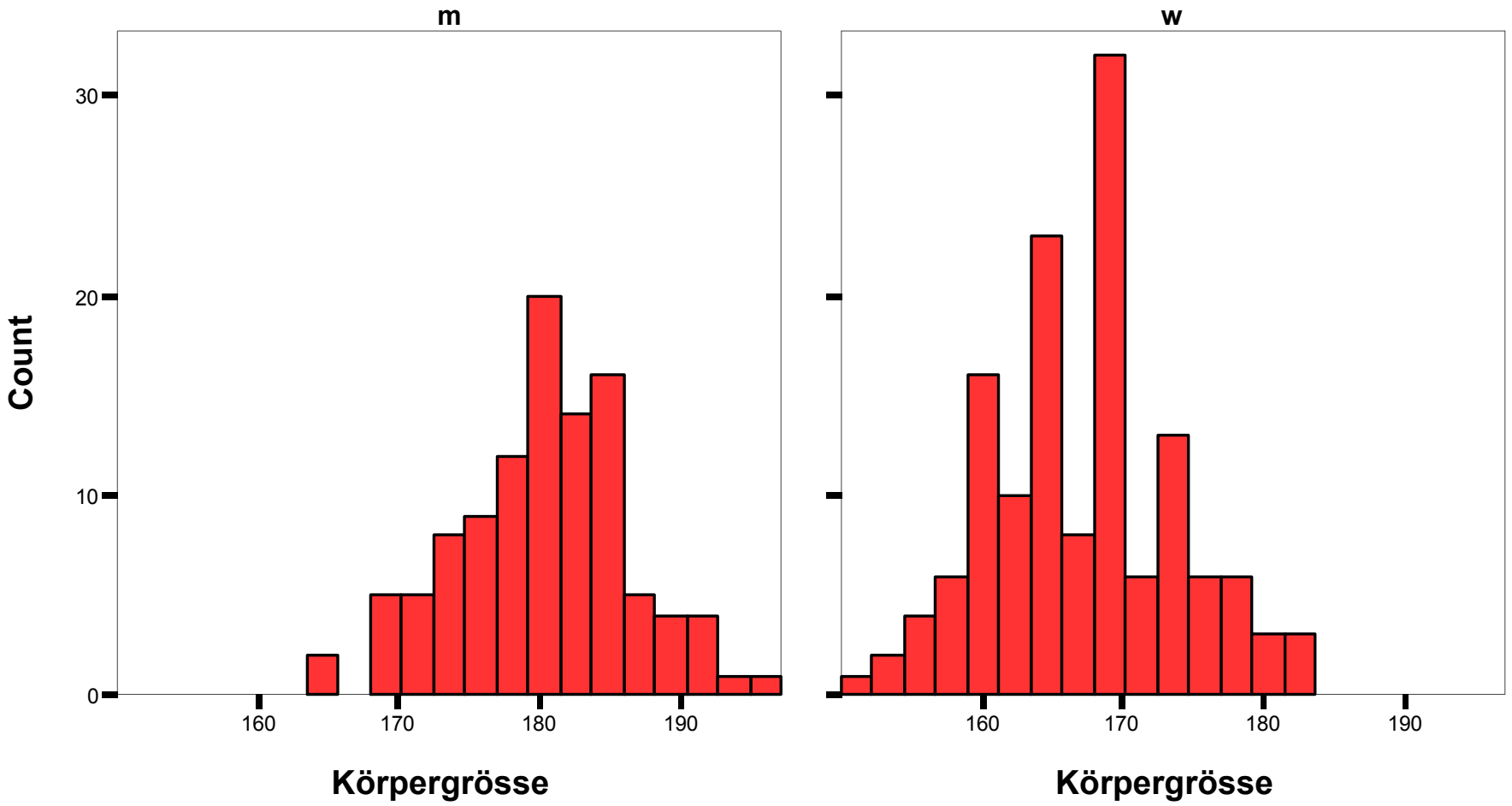
- Falls ein absoluter Nullpunkt existiert
  $\longrightarrow$ Quotienten machen Sinn
  Beispiele: Temperatur in Kelvin,
  Körpergrösse, Handlänge

SEX	Körpergrösse	HAND	GROUP	Tutor	Geschlecht
1	168	17.5	1	1	w
0	183.5	21	1	1	m
1	170	20	1	1	w
1	159	17	1	1	w
1	165	18	1	1	w
0	180	20	1	1	m
1	181	19.5	1	1	w
0	193	21.5	1	1	m
0	183	19.5	1	1	m
0	183	20.5	1	1	m
1	165	17.8	1	1	w
1	161	19.5	1	1	w
1	156	16.5	1	1	w
0	184	17	1	1	m
0	173	18.5	1	1	m
1	170	17.5	1	1	w
1	163	17.5	1	1	w
1	162	18	2	3	w
1	181	20.5	2	3	w
0	178	20	2	3	m
0	173	20	2	3	m

- nicht sinnvoll: „Es gab Zeitalter, in denen die Temperatur 60% über der jetzigen lag.“ *Film der BBC 2006*

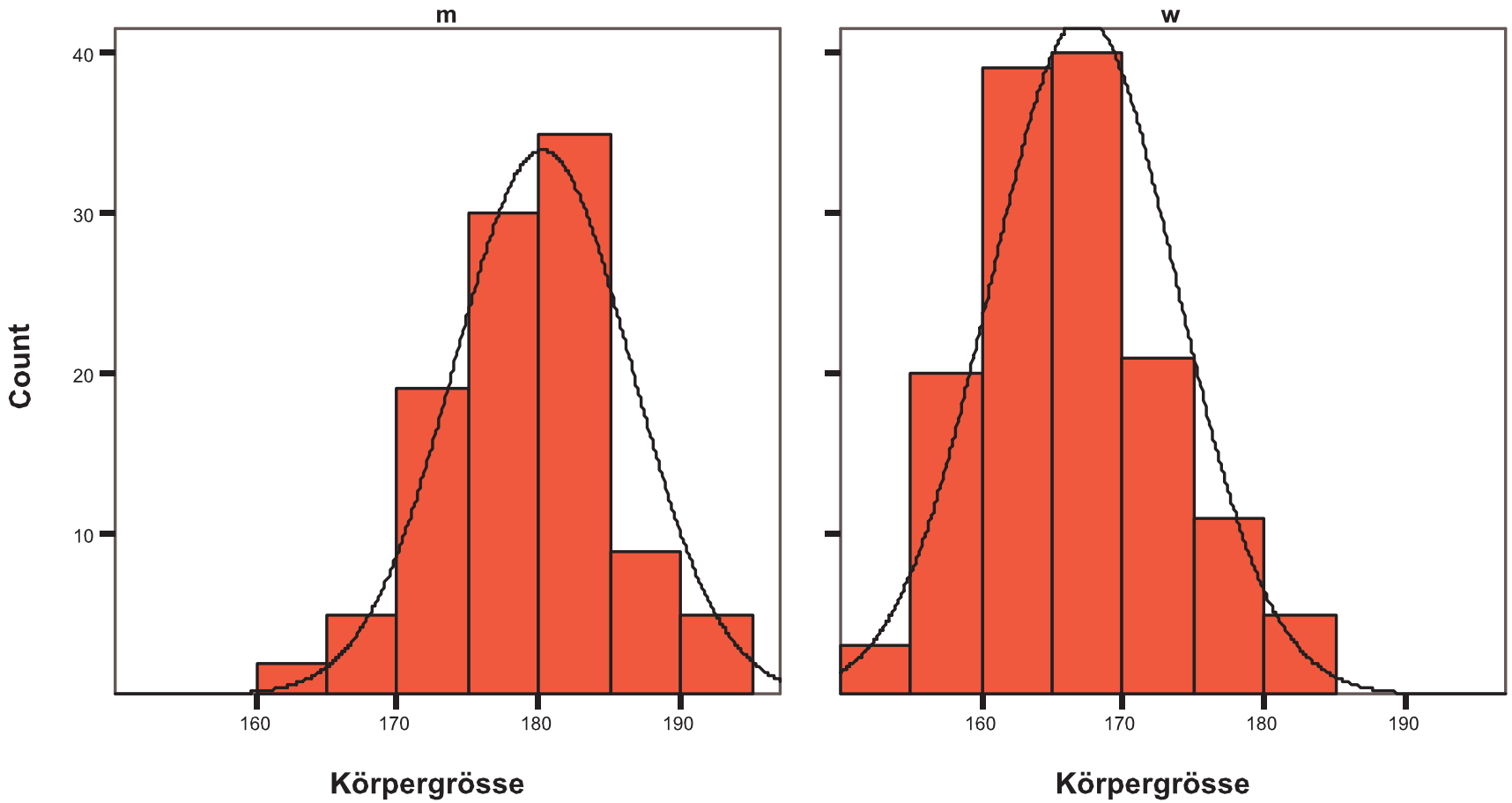
Jetzt	Damals
15^0C	24^0C
59^0F	94^0F = 34^0C
288 K	461 K = 188^0C

Histogramm



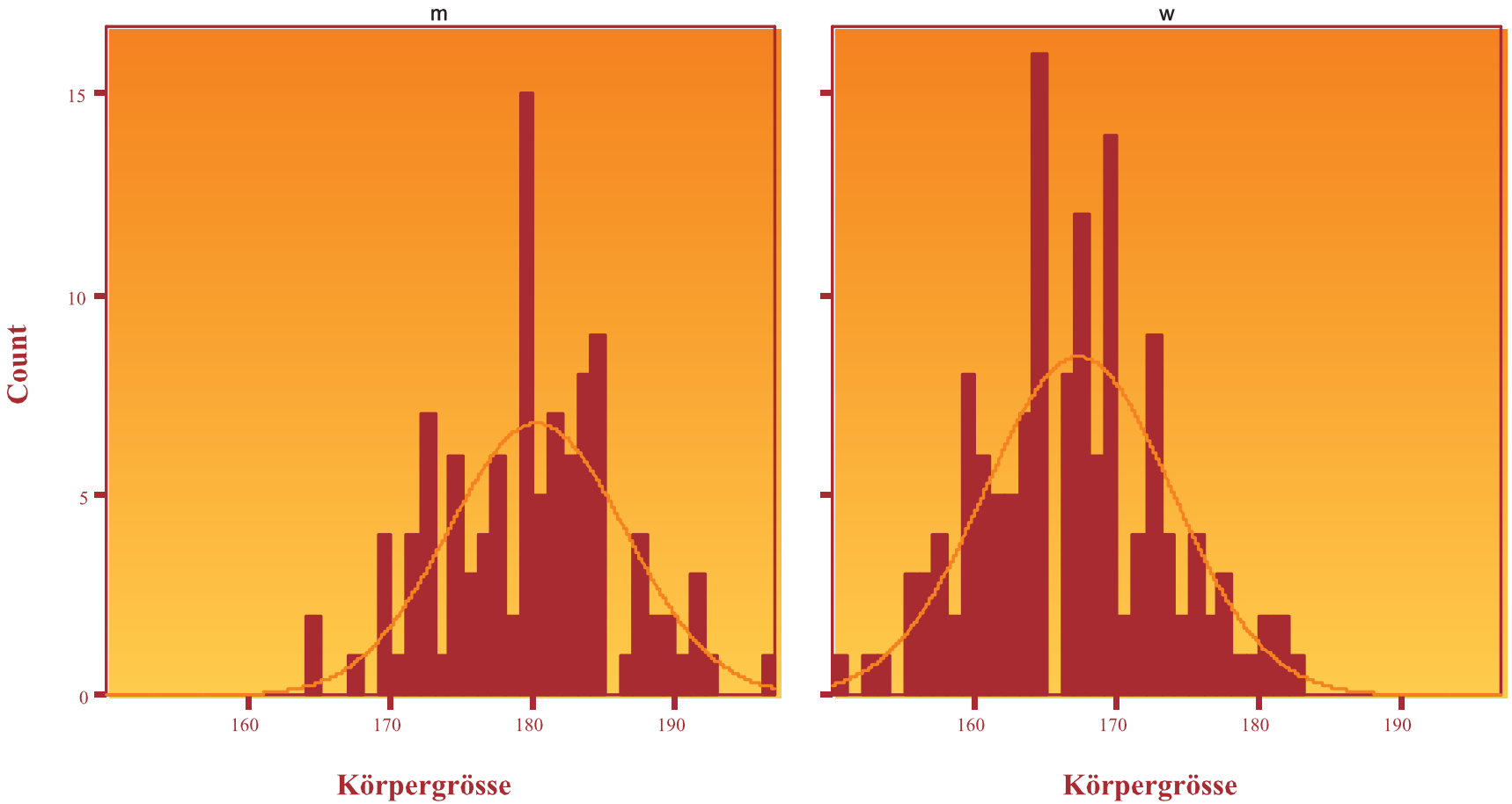
- visualisiert Verteilung in der Stichprobe
- Standardintervalllänge 2.24 cm

Histogramm



- sinnvolle Intervalllänge 5 cm
- Verteilung in der Population „Gauss'sche Normalverteilung“ angepasst

Histogramm



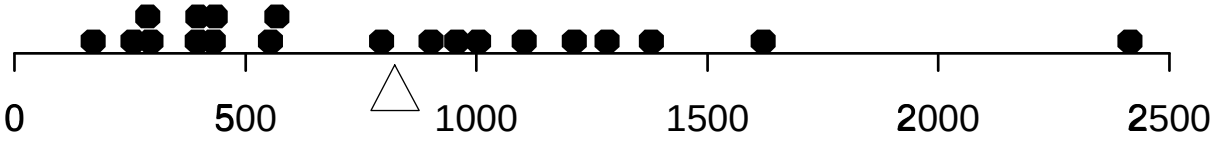
- Intervalllänge 1 cm: Histogramm sehr variabel

Charakterisierung des Zentrums der Daten

- Was ist ein typischer, mittlerer Wert ?

Mittelwert $\overline{x}$: Verhalten „im Mittel“ (mean, average)

$\overline{x} = (x_1 + x_2 + \ldots + x_n) \,/\, n$



Bei **normalverteilten Daten** ist der Mittelwert in der Stichprobe die beste Anpassung des Mittelwertes in der Population.

- empfindlich gegen Ausreisser

Streuung oder Variabilität einer Stichprobe

- Wie stark variieren die Daten um mittlere Lage?

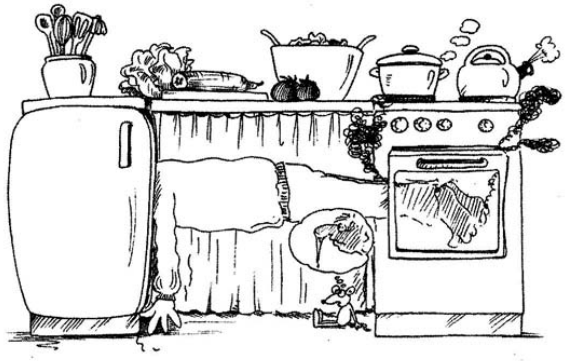
Varianz s^2:

Berechne Abweichungen $(x_1 - \overline{x}), \ldots, (x_n - \overline{x})$

Mittelwert? Nein — würde zu 0!

Also:

$s^2 = \{(x_1 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2\} \,/\, (n-1)$



A statistician is a person who, if you've got your feet in the oven and your head in the refrigerator, will tell you that, on average, you're very comfortable.

- Achtung: s^2 in quadrierten Einheiten (z. B. cm^2)

Standardabweichung: $s = \sqrt{\text{Varianz}}$ (in cm) (standard deviation, SD)

Bei **normalverteilten Daten** liegen 68% der Daten im Bereich Mittelwert $\pm$ SD, 95% der Daten im Bereich Mittelwert $\pm$ 2 SD.

- keine derartige Interpretation bei nicht normalverteilten Daten
- empfindlich gegen Ausreisser

- Daten werden oft als Mittelwert plus–minus Standardabweichung (mean $\pm$ SD) angegeben.

- Output SPSS:

Descriptive Statistics

Geschlecht		N	Minimum	Maximum	Mean	Std. Deviation
m	Körpergrösse	106	165	197	180.20	6.233
	Valid N (listwise)	106				
w	Körpergrösse	139	150	183	167.22	6.568
	Valid N (listwise)	139				

Mean $\pm$ **SD** oder **Mean** $\pm$ **SEM ?**

- Der **Standardfehler** des Mittelwertes (standard error of the mean, SEM) ist die Standardabweichung des Mittelwertes:

$$\text{SEM} = \text{SD}/\sqrt{n}$$

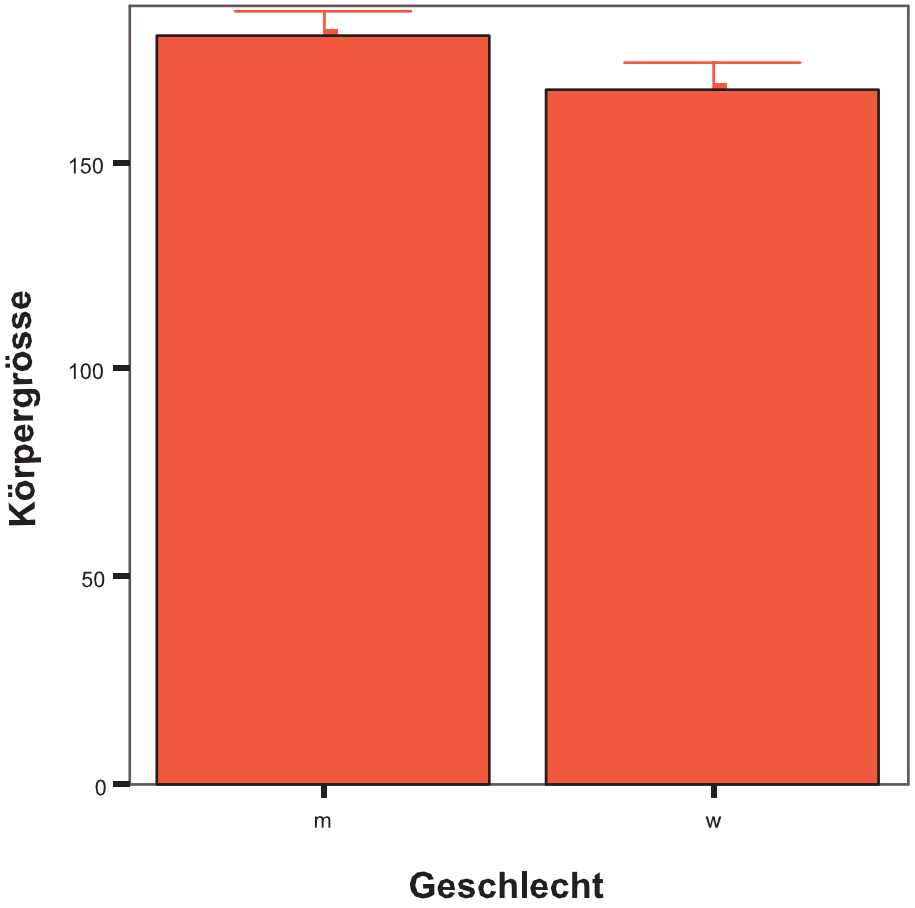
Der SEM hat in der deskriptiven Statistik nichts zu suchen!

- Output SPSS:

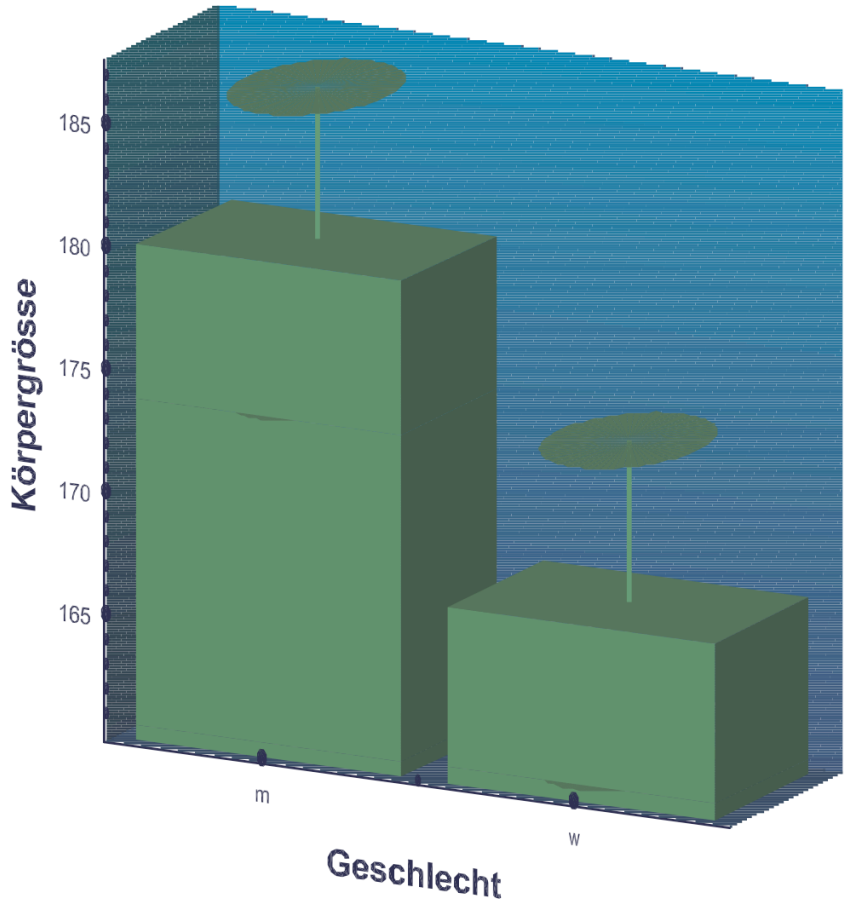
Descriptive Statistics

Geschlecht		N	Minimum	Maximum	Mean		Std. Deviation
		Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
m	Körpergrösse	106	165	197	180.20	.605	6.233
	Valid N (listwise)	106					
w	Körpergrösse	139	150	183	167.22	.557	6.568
	Valid N (listwise)	139					

Balkendiagramm

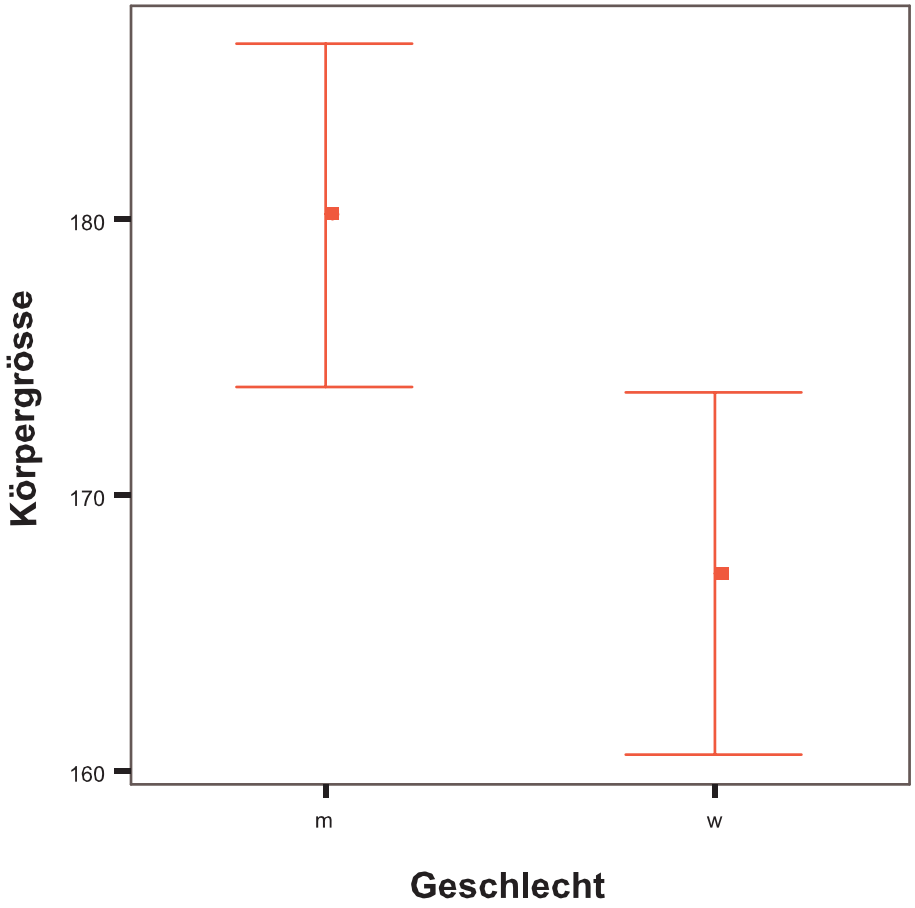


- Balken stehen auf dem Boden, deshalb Nullpunkt beachten
- Vorsicht vor 3–dimensionaler Darstellung



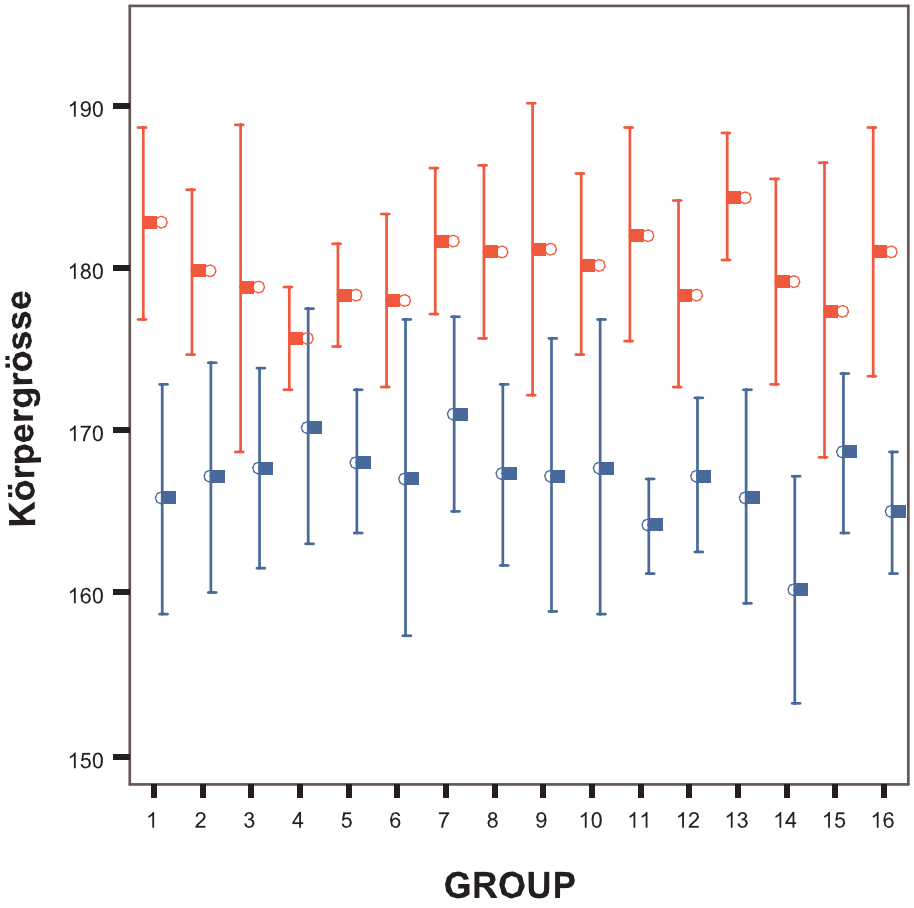
- Balken stehen auf dem Boden, deshalb Nullpunkt beachten
- Vorsicht vor 3–dimensionaler Darstellung

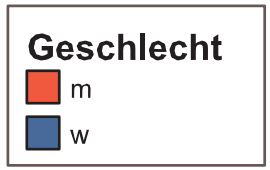
Punktdiagramm



- Nullpunkt hat hier keine Bedeutung

Die Körpergrösse variiert von Stichprobe zu Stichprobe





Error Bars show Mean +/- 1.0 SD

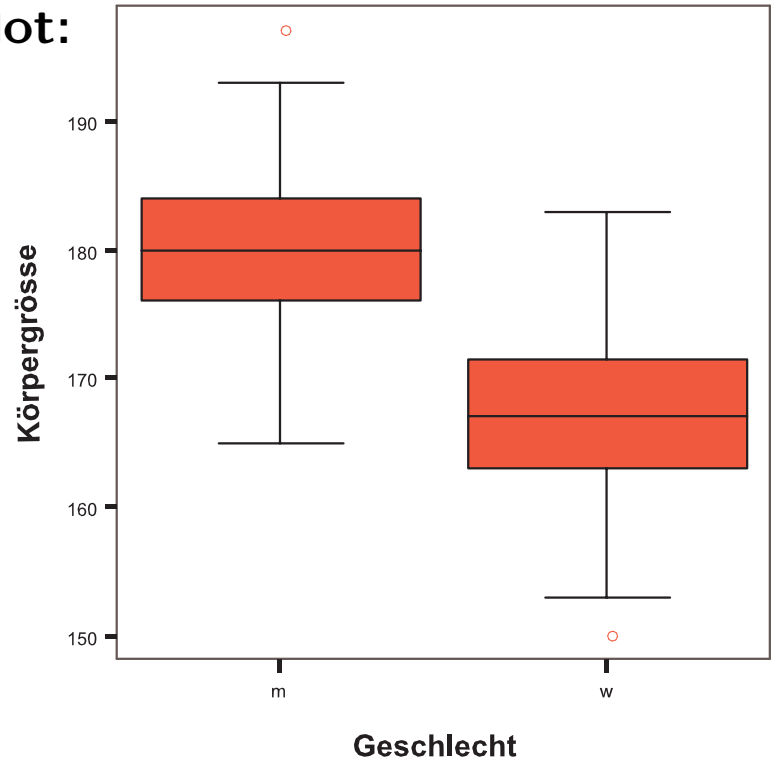
Dot/Lines show Means

Perzentile (Quantile)

α. – **Perzentil** (α% – Quantil):
α% der Daten sind kleiner oder gleich dem α. – Perzentil
und $(100 - \alpha)$% sind grösser oder gleich.

Beispiele:
- **Median** = 50. Perzentil
- **Quartile** = 25. und 75. Perzentile

Boxplot:



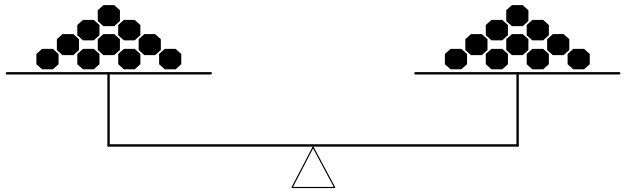
← Maximum (ohne Ausreisser)

← oberes Quartil
← Median
← unteres Quartil

← Minimum (ohne Ausreisser)

Charakterisierung des Zentrums der Daten

Median: „Zentrum“ der Daten, 50. Perzentil,
d.h. Hälfte der Stichprobe über Median, und Hälfte darunter



- Output SPSS:

Statistics

Körpergrösse

m	N	Valid	106
		Missing	0
	Minimum		165
	Maximum		197
	Percentiles	25	175.75
		50	180.00
		75	184.00
w	N	Valid	139
		Missing	0
	Minimum		150
	Maximum		183
	Percentiles	25	163.00
		50	167.00
		75	172.00

Variabilität einer Stichprobe

Spannweite = Maximum − Minimum

- gibt den Bereich (range) aller Daten an
- stark durch Extremwerte beeinflusst
- aber: Minimum und Maximum sehr einfach zu verstehen

$\longrightarrow$ Daten dennoch oft als „median[range]“ angegeben
„Median–Körpergrösse bei männlichen Studenten 180cm[165 − 197cm]“

Interquartilsabstand (interquartile range, IQR)
= 75. Perzentil − 25. Perzentil
= Boxlänge im Boxplot, umfasst zentrale 50% der Daten

- wie Standardabweichung ein Mass für Grösse des Bereichs der zentralen Daten

Bei der **Normalverteilung** ist der halbe Interquartilabstand 0.67 SD.

- „Median(IQR)“ sagt nichts über Schiefe

$\longrightarrow$ Daten oft als „Median [unteres Quartil, oberes Quartil]“ angegeben.

Wahrscheinlichkeitsrechnung

- Verbindung zwischen Stichprobe und Population

- „**wahre**“ (Populations–) **Kennzahlen**:

Wahrscheinlichkeit ($\approx$ relative Häufigkeit p) π

Erwartungswert ($\approx$ Mittelwert $\bar{x}$): μ

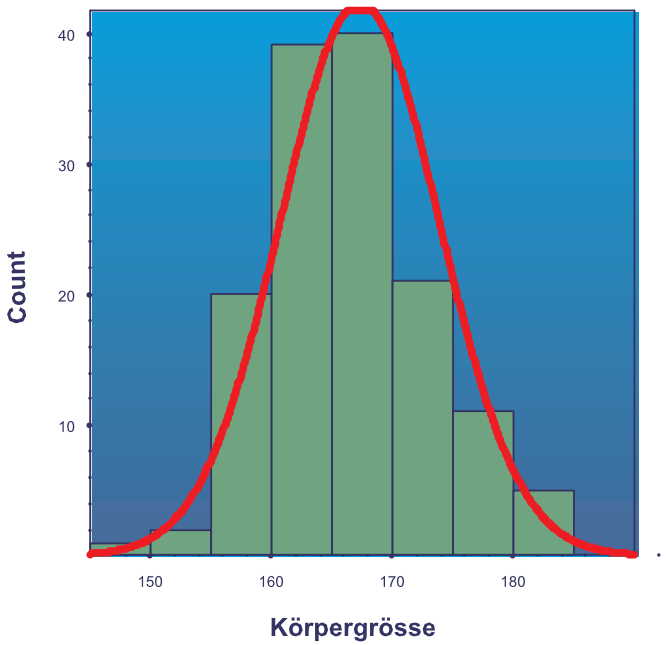
Standardabweichung ($\approx s$): σ

Perzentile

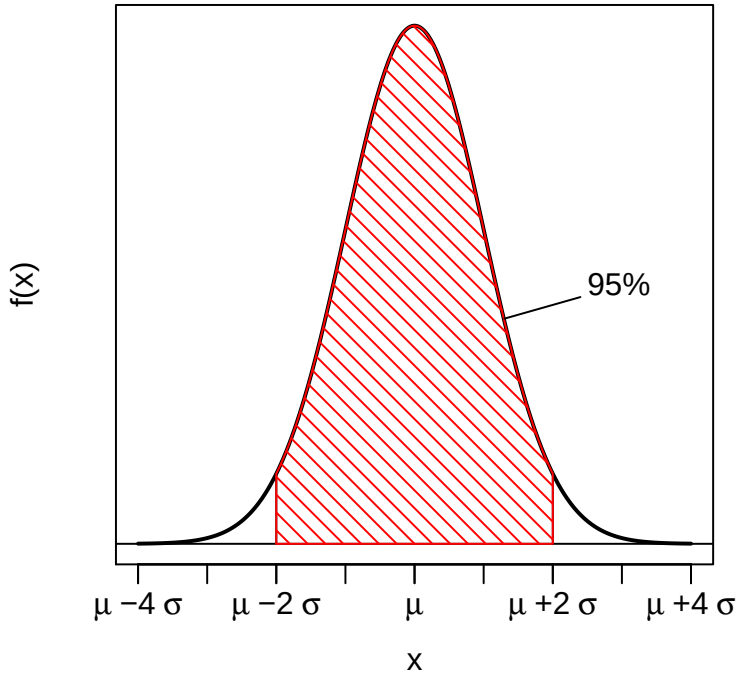
- benötigt für Testen und Konfidenzintervalle

Was ist eine Normalverteilung?

Stichprobe: Histogramm



Population: Wahrscheinlichkeitsdichte



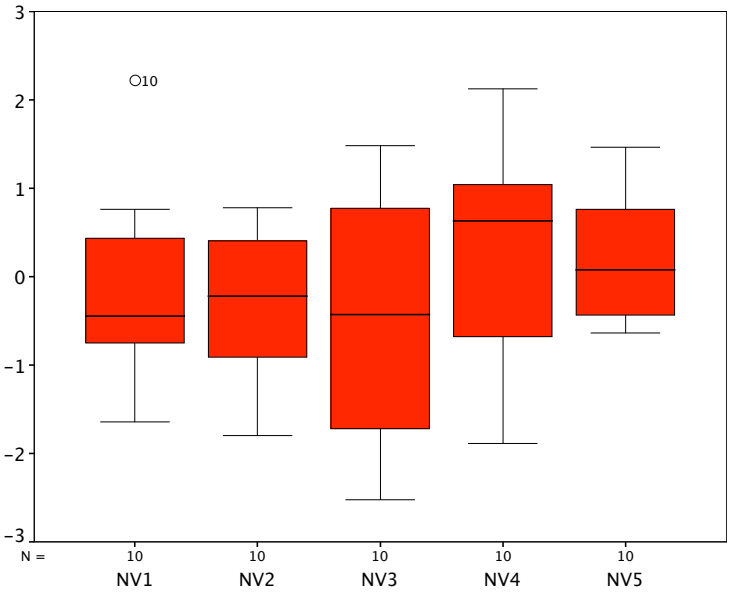
- Dichte:

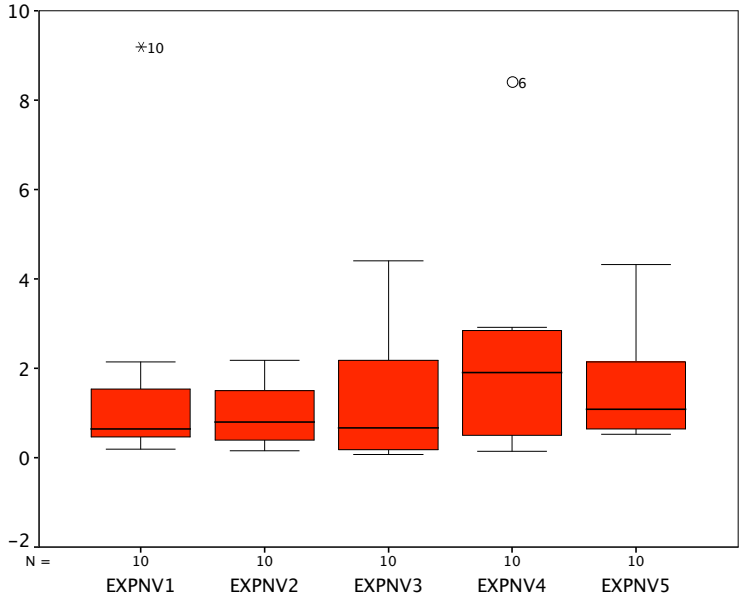
Wahrscheinlichkeit im Intervall $[a, b]$
= Fläche unter der Kurve von a bis b

Normalverteilung: symmetrisch, „keine“ Ausreisser

Woran erkennt man eine Normalverteilung ?

- kein Testproblem: die Nullhypothese kann man nicht beweisen
- graphisch überprüfen

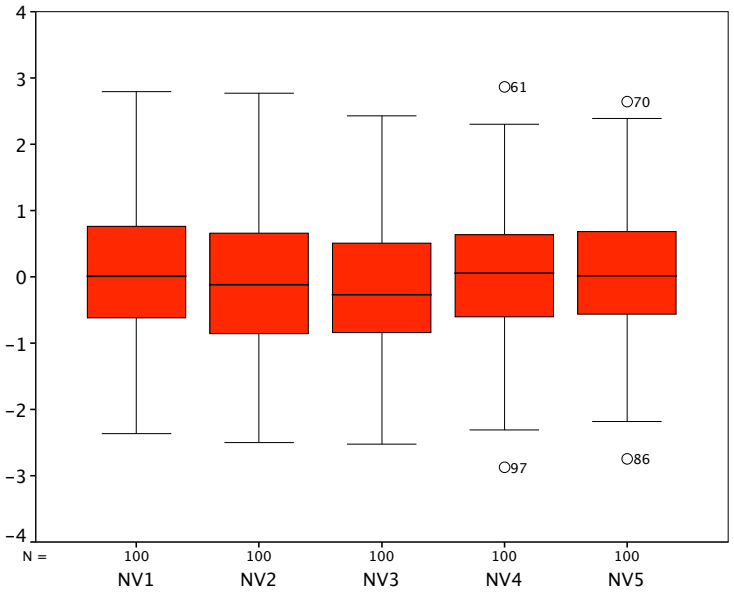


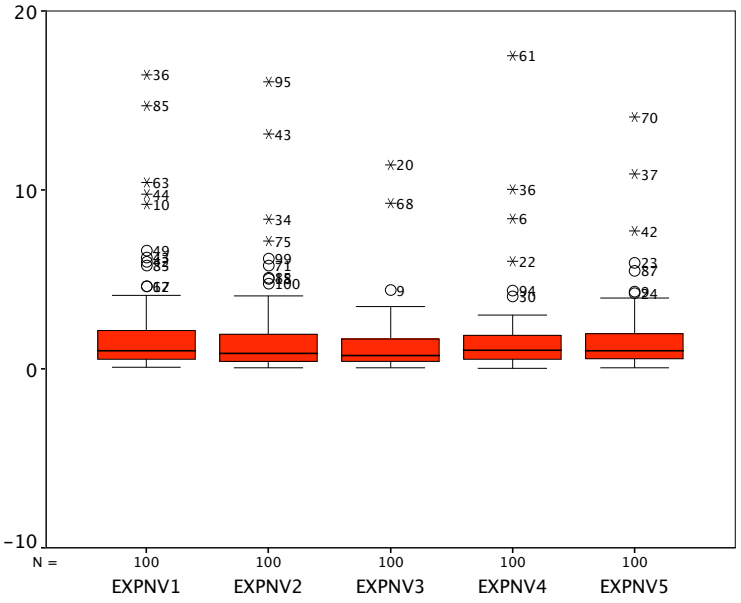


- **Normalverteilung** symmetrisch, „keine“ Ausreisser $\longrightarrow$ Median $\approx$ Mittelwert
- bei nichtnegativen Variablen: SD $<$ mean / 2 (besser: SD $<$ mean / 3)

Woran erkennt man eine Normalverteilung ?

- kein Testproblem: die Nullhypothese kann man nicht beweisen
- graphisch überprüfen

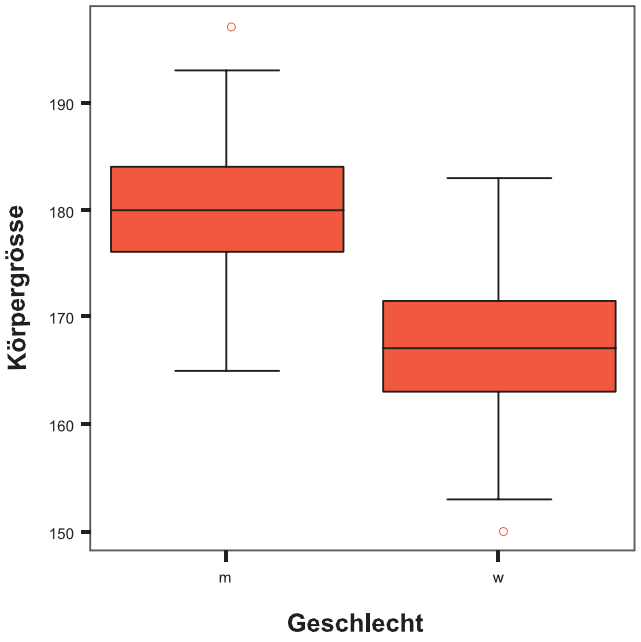




- **Normalverteilung** symmetrisch, „keine“ Ausreisser $\longrightarrow$ Median $\approx$ Mittelwert
- bei nichtnegativen Variablen: SD $<$ mean / 2 (besser: SD $<$ mean / 3)

Woran erkennt man eine Normalverteilung ?

- kein Testproblem: die Nullhypothese kann man nicht beweisen
- graphisch überprüfen



Statistics

Körpergrösse

m	N	Valid	106
		Missing	0
	Mean		180.20
	Median		180.00
	Std. Deviation		6.233
	Skewness		-.006
	Std. Error of Skewness		.235
	Kurtosis		-.041
	Std. Error of Kurtosis		.465
w	N	Valid	139
		Missing	0
	Mean		167.22
	Median		167.00
	Std. Deviation		6.568
	Skewness		.114
	Std. Error of Skewness		.206
	Kurtosis		-.201
	Std. Error of Kurtosis		.408

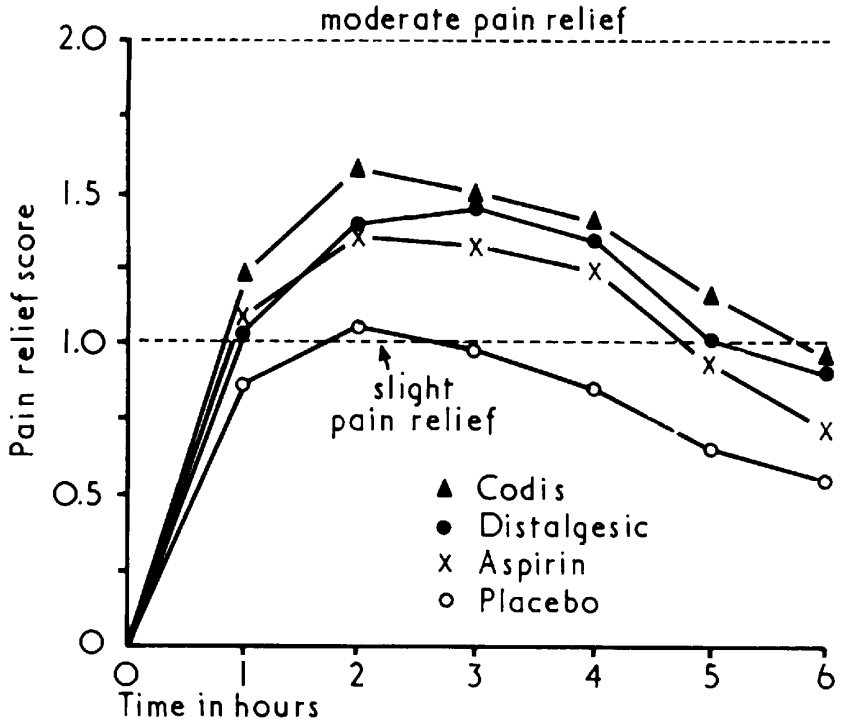
- **Normalverteilung** symmetrisch, „keine“ Ausreisser $\longrightarrow$ Median $\approx$ Mittelwert
- bei nichtnegativen Variablen: SD $<$ mean / 2 (besser: SD $<$ mean / 3)

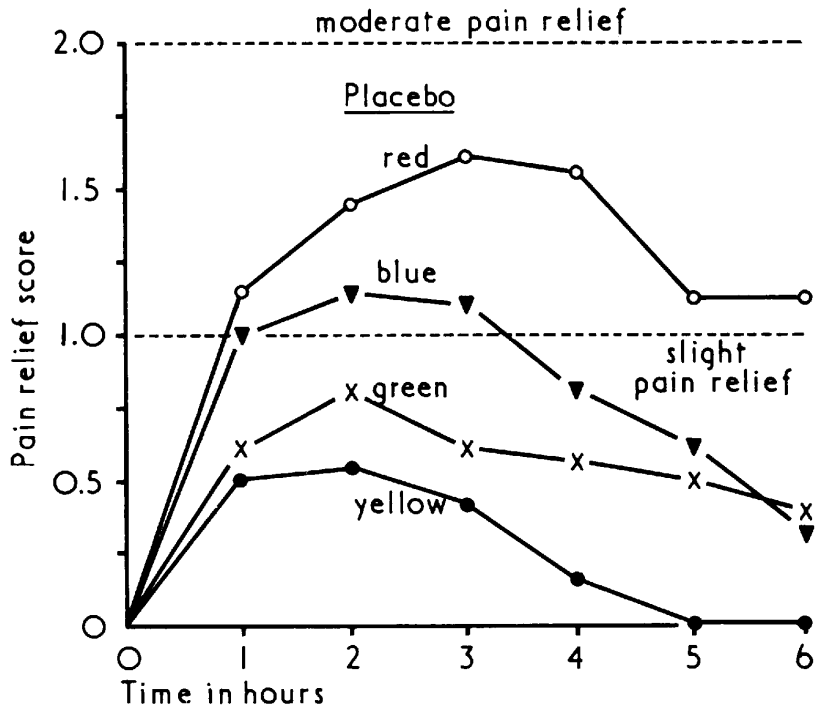
Versuchsplanung

- **Repräsentativität**: gleiche Chance für alle (einer Population), in die Stichprobe zu kommen
- **Randomisierung**: gleiche Chance für alle (einer Stichprobe), in eine Gruppe zu kommen
- **Standardisiertes Vorgehen:** klare Ein–/Ausschlusskriterien, experimentelle Bedingungen
- **Doppelverblindung:** Verfälschung durch Subjektivität vermeiden
- **Kontrolle:** neue Methode mit Placebo oder Standardtherapie vergleichen
- **Unabhängigkeit der Versuchseinheiten:** Beine eines Versuchstieres sind nicht unabhängig.
- **Einfache Versuche: zwei** Gruppen oder **zwei** Zeitpunkte vergleichen
- **Adäquate Stichprobengrösse:** Sowohl zu kleine als auch zu grosse Stichproben sind unethisch.
- **Informed consent**

Wichtigkeit von Placebo und Standardisierung

Beispiel: Huskisson EC (1974). Simple analgesics for arthritis. *BMJ* 4, 196–200.





- Rotes Placebo ist eines der wirkungsvollsten Schmerzmittel.