Applied Data Science (Prof. Dr. Kauffeldt)

Inhalt

- 1 Deskriptive Methoden
- 2 Testmethoden
  - 2.1 Ablauf statistischer Test
  - 2.2 Testen von Lageparametern

1 Deskriptive Methoden

1.1 Statistiken

```
Analysen -> Exploration -> Deskriptivstatistik
```

Deskriptivstatistik

Deskriptivstatistik

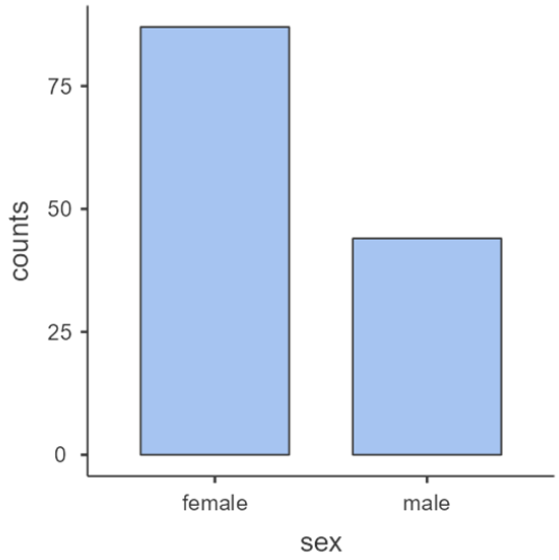
	spend_food
N	128
Fehlend	5
Mittelwert	183
Median	150
Modalwert	200
Standardabweichung	129
Varianz	16642
IQR	150
Wertebereich	800
Minimum	0
Maximum	800

Kann auch nach einer Gruppenvariable (bspw. Geschlecht) aufgeteilt werden.

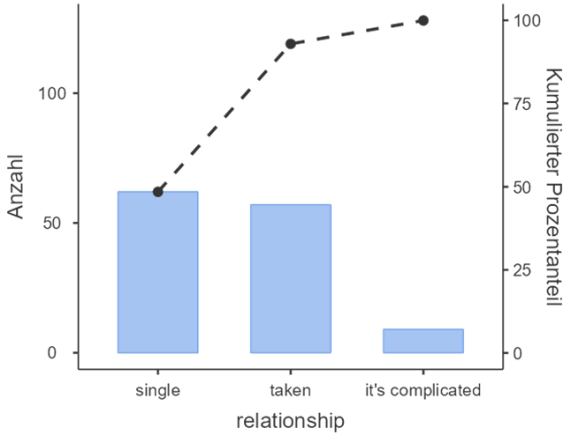
1.2 Graphiken

Nominale und Ordinale Daten: Häufigkeiten

```
Analysen -> Exploration -> Deskriptivstatistik -> Diagramme -> Balkendiagramm
```

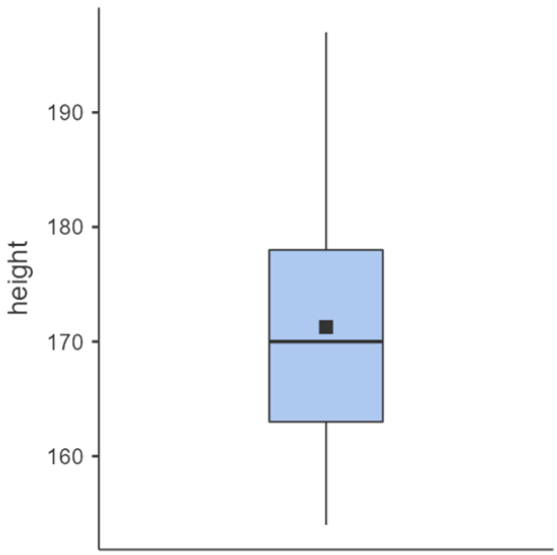


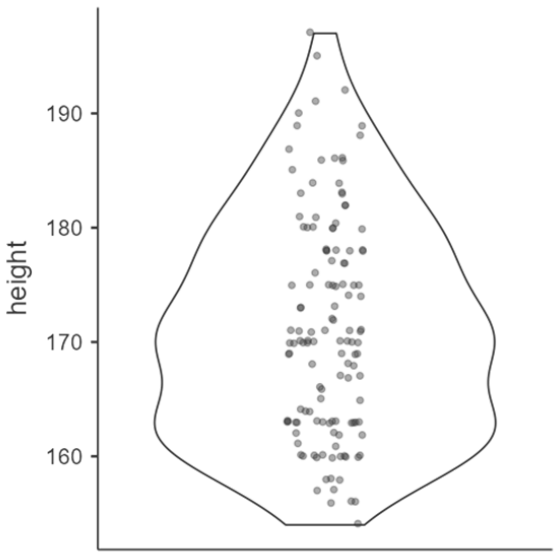
`Analysen -> Exploration -> Deskriptivstatistik -> Pareto-Diagramm`



Numerische Daten und Ordinale Daten: Boxplot und Violinplot

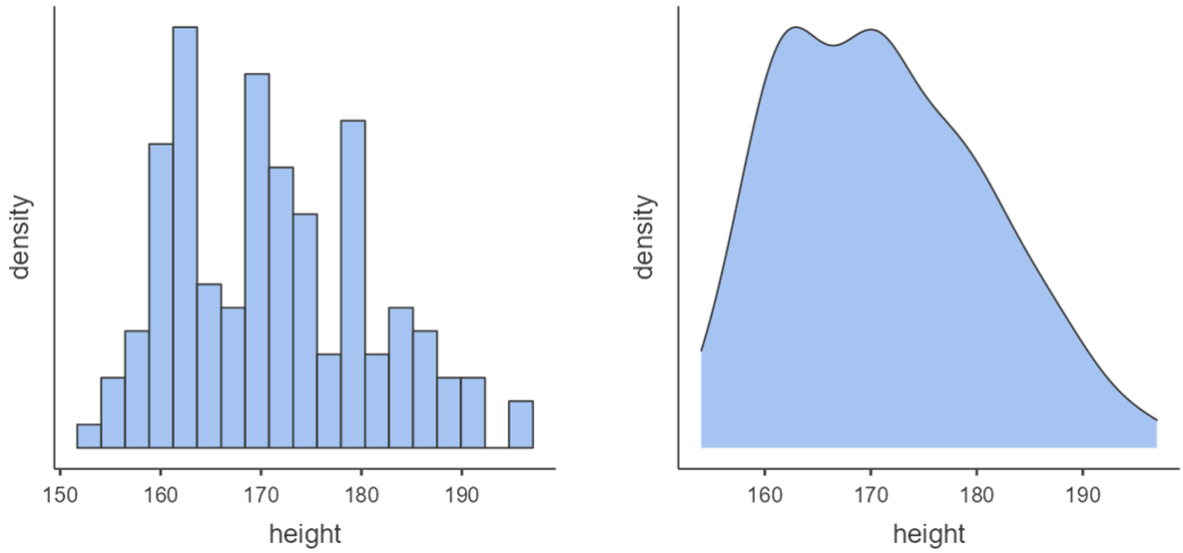
`Analysen -> Exploration -> Deskriptivstatistik -> Diagramme -> Boxplots`





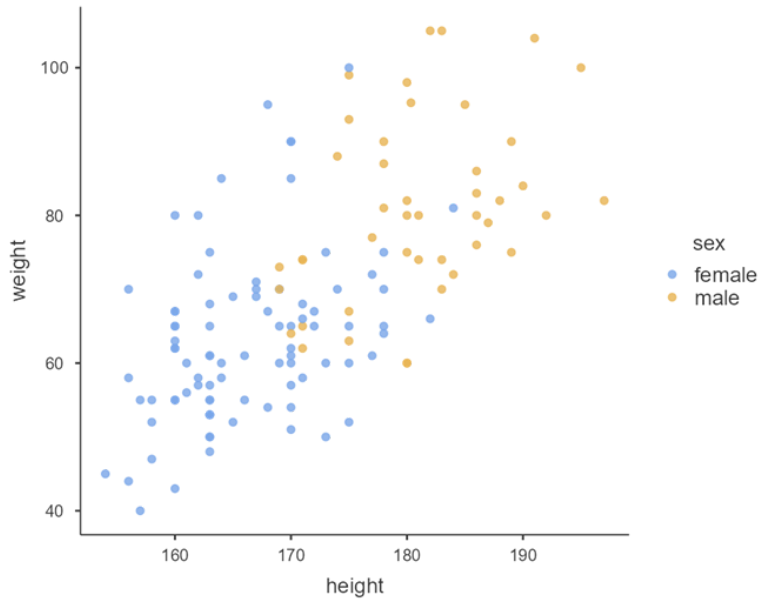
Numerische Daten: Histogramm und Dichte

```
Analysen -> Exploration -> Deskriptivstatistik -> Diagramme -> Histogramme
```



Bivariate numerische Daten: Streudiagramm

```
Analysen -> Exploration -> Deskriptivstatistik -> Streudiagramm
```



2 Testmethoden

2.1 Ablauf statistischer Test

1. **Problemstellung und Hypothesen formulieren**
   Nullhypothese H_0 ("Status Quo") und Alternativhypothese H_1 ("Forschungshypothese")
2. **Passenden statistischen Test auswählen**
3. **Voraussetzungen des Tests prüfen**
   bspw. Varianzhomogenität, Normalverteilung

4. **Ggf. Voranalyse**
5. **Ggf. Data Engineering**
   bspw. Codierung
6. **Test durchführen und interpretieren**

2.2 Testen von Lageparametern

Übersicht:

Messniveau	Test auf	Einstichprobentest	Zweistichprobentest	
			Unabhängig	*Abhängig*
Numerisch	Mittel-wert	t-Test	t-Test (Varianz-homogenität) Welch-Test	Gepaarter t-Test
Ordinal	Median	Vorzeichen-Test (Wilcoxon W) (*)	Mann-Whitney-U-Test (*)	Wilcoxon-Vorzeichen-Rang-Test (*)

(*) Nichtparametrische Tests

Beispiel: Zweistichproben t-Test (unabhängig)

Schritt 1: Problemstellung und Hypothesen formulieren

$$H_0 : Durschnittsgr\ddot{o}\ss eMann \leq Durchschnittsgr\ddot{o}\ss eFrau$$

$$H_1 : Durschnittsgr\ddot{o}\ss eMann > Durchschnittsgr\ddot{o}\ss eFrau$$

Schritt 2: Passenden Test auswählen
Unabhängiger Zweistichproben t-Test

Schritt 3: Voraussetzungen des Tests überprüfen

Voraussetzungen t-Test

- **T1. Numerische abhängige Variable.**
- **T2. Normalität.** Die Population(en) sind normalverteilt.
- **T3. Unabhängigkeit.** Die Messungen innerhalb und zwischen den Gruppen sind unabhängig.
- **T4. Binäre Gruppenvariable.** Es werden genau zwei Gruppen verglichen. [*]
- **T5. Homoskedastizität.** Varianzhomogenität: Varianz Gruppe 1 = Varianz Gruppe 2. [*]

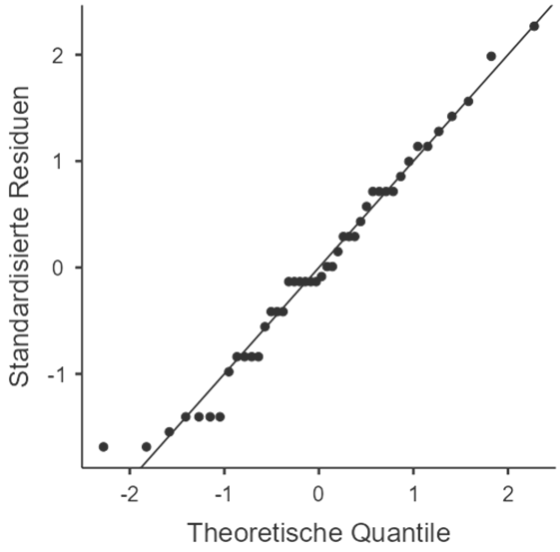
[*] Nur für Zweistichprobentest

T1.
Körpergröße ist numerisch. ✔

T2.
Überprüfung: Shapiro-Wilk-Test (H_0 : Normalverteilung, H_1 : Keine Normalverteilung) und QQ-Plot:

Erst nach Gruppe filtern, dann `Analysen -> Exploration -> Deskriptivstatistik -> Shapiro-Wilk und Q-Q`

Gruppe Männer:

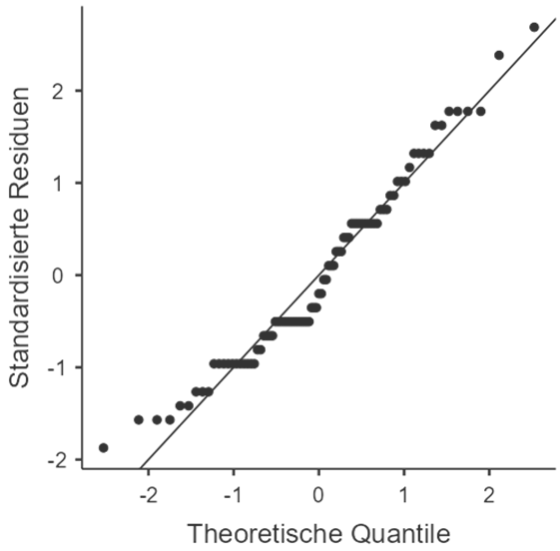


Deskriptivstatistik

	height
N	44
Fehlend	0
Mittelwert	181
Median	180
Standardabweichung	7.08
Minimum	169
Maximum	197
Shapiro-Wilk W	0.975
Shapiro-Wilk p	0.464

→ Erfüllt ✔

Gruppe Frauen:



Deskriptivstatistik

	height
N	87
Fehlend	0
Mittelwert	166
Median	165
Standardabweichung	6.58
Minimum	154
Maximum	184
Shapiro-Wilk W	0.964
Shapiro-Wilk p	0.017

→ Nicht erfüllt ✗

T3.
Messungen sind unabhängig. ✔

T4.
Nur 2 Gruppen. ✔

T5.
Überprüfung: Levenes Test (H_0 : Varianzen aller Gruppen sind gleich,

H_1 : Varianzen mindestens zweier Gruppen unterscheiden sich)

`Analysen -> t-Test für unabhängige Stichproben -> Homogenitätstest`

Levene's Test auf Varianzhomogenität

	F	df	df2	p
height	0.0233	1	129	0.879

Anmerkung. Ein niedriger p-Wert deutet auf eine Verletzung der Annahme gleicher Varianzen hin

→ Erfüllt ✔

Was tun, wenn die Voraussetzungen des Tests verletzt sind?

Abhängig von der Art der Verletzung:

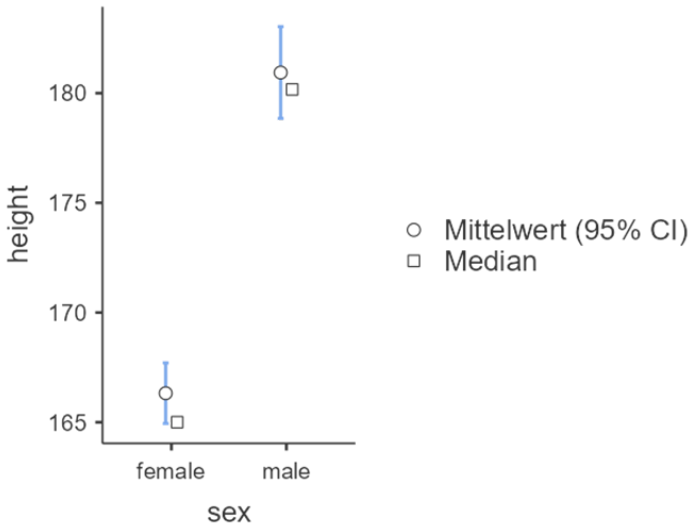
- Bei gewissen Verletzungen (bspw. abhängige Variable nicht-numerisch) kann der Test nicht durchgeführt werden
  - Bspw. bei ordinaler Variable Mann-Whitney-U-Test verwenden.
- Bei anderen Verletzungen erhalten wir weniger robuste Resultate.
  - Bei Verletzungen der Verteilungsannahme (Normalität), verwenden eines nichtparametrischen Tests zur Überprüfung der Resultate

Schritt 4: Voranalyse

`Analysen -> t-Test für unabhängige Stichproben -> Deskriptivstatistik und Deskriptive Diagramme`

Deskriptivstatistik für die Gruppen

	Gruppe	N	Mittelwert	Median	Std.-abw.	Std.-fehler
height	female	87	166	165	6.58	0.705
	male	44	181	180	7.08	1.07



Schritt 5: Test durchführen und interpretieren

`Analysen -> t-Test für unabhängige Stichproben`

t-Test für unabhängige Stichproben

		Statistik	±%	df	p	Mittlere Differenz	Std.-fehler der Differenz	95% Konfidenzintervall			Effektstärke
								Untere	Obere		
height	Student's t	-11.7		129	< .001	-14.6	1.25	-Inf	-12.5	Cohens d	-2.17
	Bayes-Faktor$_{10}$	1.12e+19	NaN								

Anmerkung. H_a $\mu_{female} < \mu_{male}$

→ p-Wert < 5% → H_0 kann abgelehnt werden → Statistisch signifikant → Beleg für H_1

Effektstärke:

$$Cohens\ d = \frac{Mittelwert_1 - Mittelwert_2}{gepoolte\ Standardabweichung}$$

Cohen´s d effect size	Interpretation	Differences in SD
d = .0 – .19	Trivial effect	<1/5 from a SD
d = .20	Small effect	1/5 from a SD
d = .50	Medium effect	1/2 from a SD
d = .80 or higher	Large effect	8/10 from a SD

Bayes-Faktor:

$$BF_{10} = \frac{P(\text{Beobachtete Daten} \mid \text{H1 wahr})}{P(\text{Beobachtete Daten} \mid \text{H0 wahr})}$$