

# Applied Data Science (Prof. Dr. Kauffeldt)

## Inhalt

- 1 Deskriptive Methoden
- 2 Testmethoden
  - 2.1 Ablauf statistischer Test
  - 2.2 Testen von Lageparametern
  - 2.3 Testen von Zusammenhängen
  - 2.4 Multiples Testen

## 1 Deskriptive Methoden

### 1.1 Statistiken

Analysen -> Exploration -> Deskriptivstatistik

## Deskriptivstatistik

Deskriptivstatistik

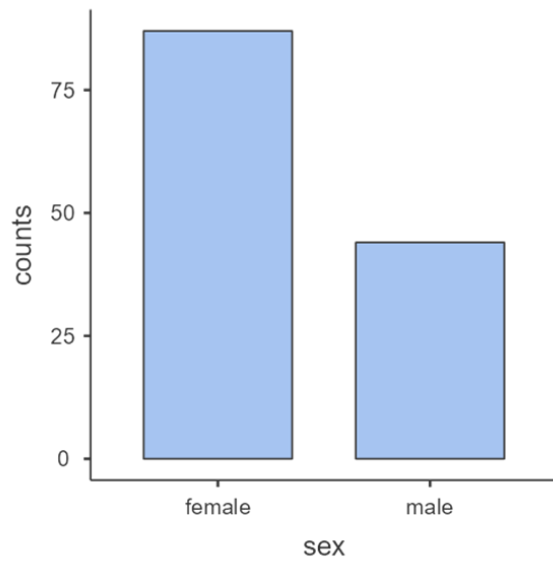
	spend_food
N	128
Fehlend	5
Mittelwert	183
Median	150
Modalwert	200
Standardabweichung	129
Varianz	16642
IQR	150
Wertebereich	800
Minimum	0
Maximum	800

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

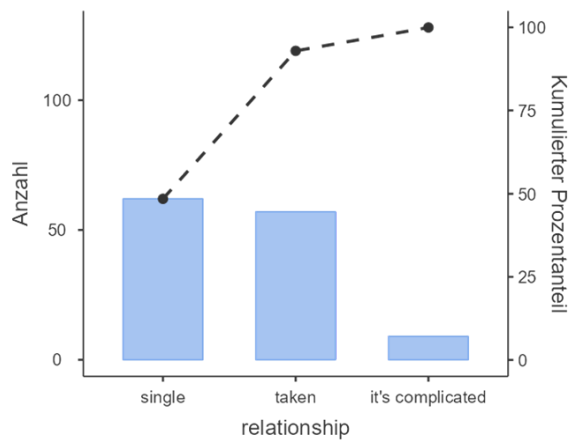
### 1.2 Graphiken

## Nominale und Ordinale Daten: Häufigkeiten

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

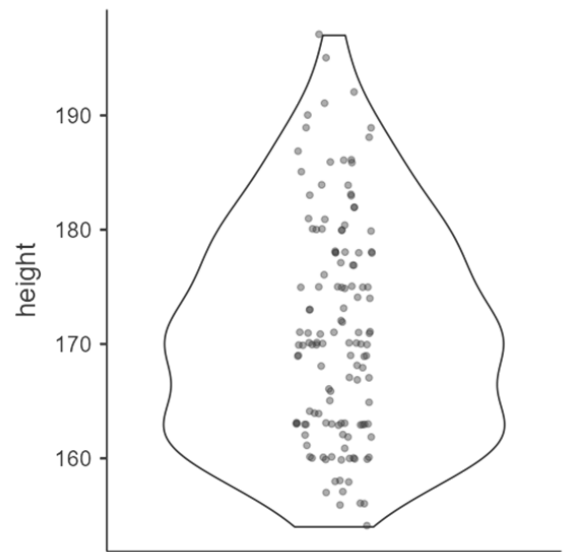
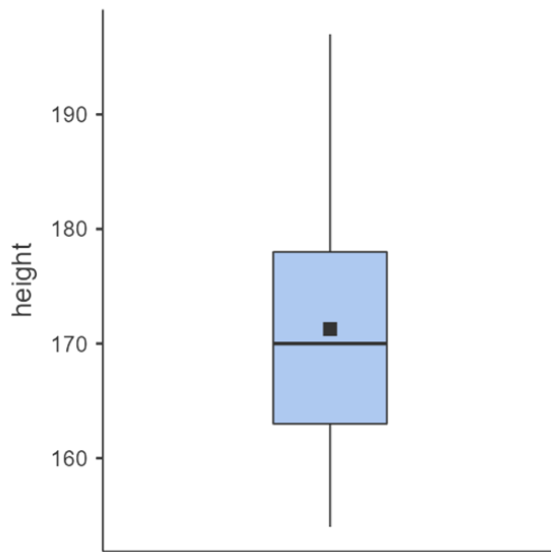


Analysen -> Exploration -> Deskriptivstatistik -> Pareto-Diagramm



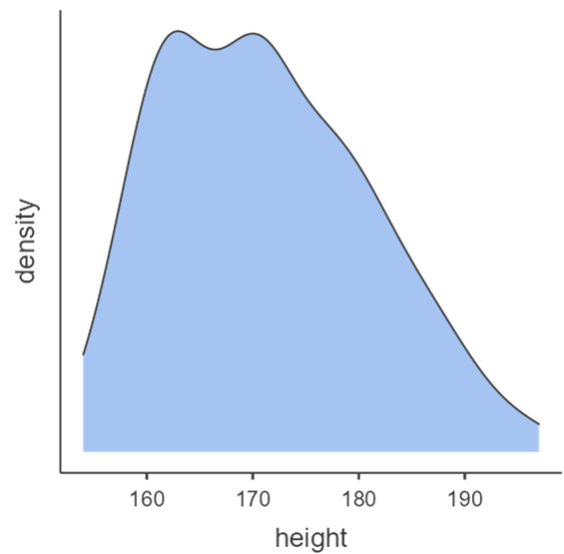
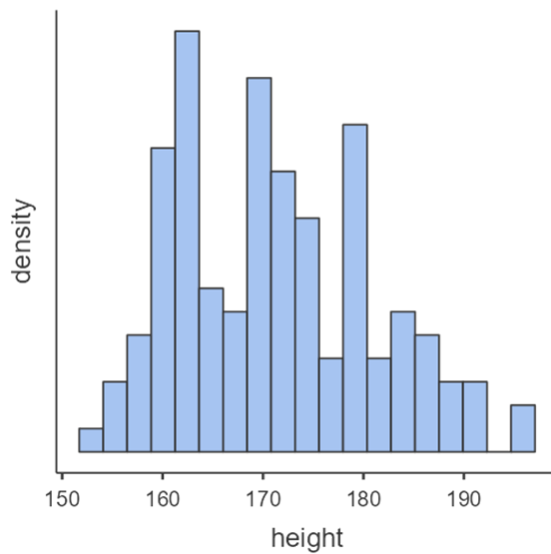
## Numerische Daten und Ordinale Daten: Boxplot und Violinplot

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



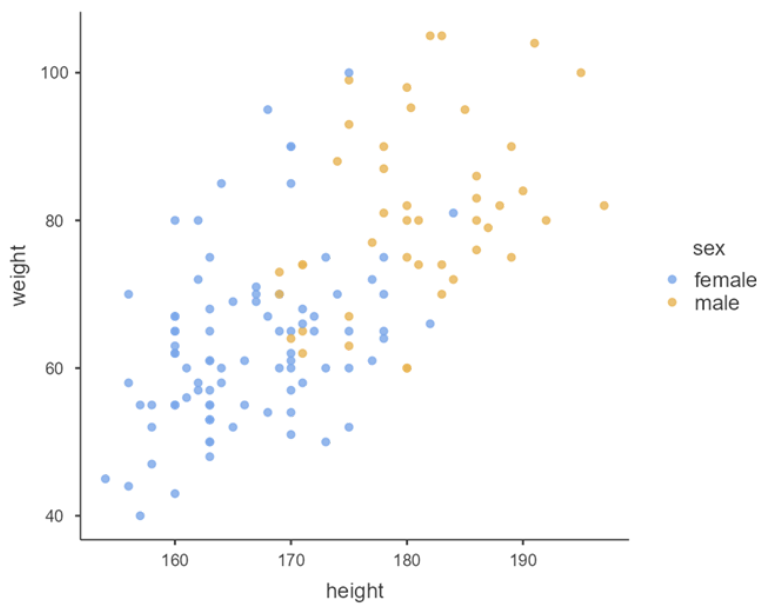
### Numerische Daten: Histogramm und Dichte

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	
			<i>Unabhängig</i>	<i>Abhängig</i>
Numerisch	Mittelwert	t-Test	t-Test (Varianzhomogenität) Welch-Test	Gepaarter t-Test
			Mann-Whitney-U-Test (*)	Wilcoxon-Vorzeichen-Rang-Test (*)
Ordinal	Median	Vorzeichen-Test (Wilcoxon W) (*)		

(\*) Nichtparametrische Tests

### Schritt 1: Problemstellung und Hypothesen formulieren

$$H_0 : \text{DurchschnittsgrößeMann} \leq \text{DurchschnittsgrößeFrau}$$

$$H_1 : \text{DurchschnittsgrößeMann} > \text{DurchschnittsgrößeFrau}$$

### 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

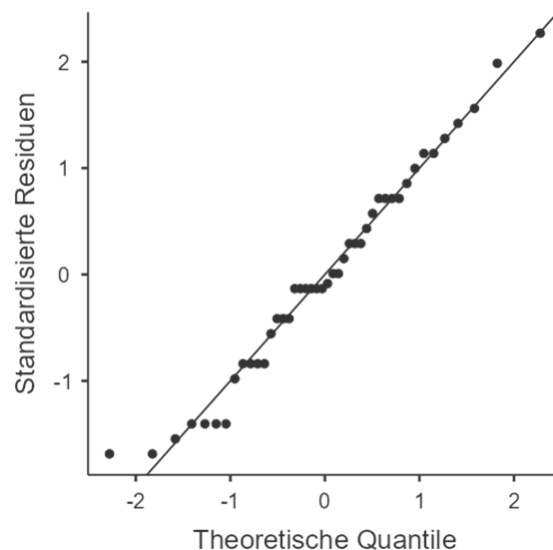
#### T1.

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:

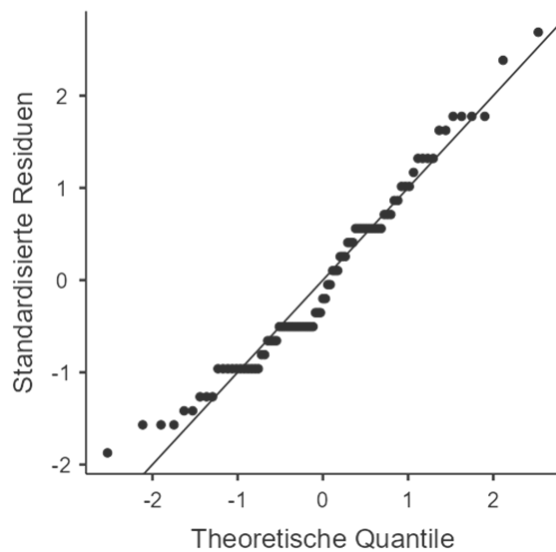


#### Deskriptivstatistik

	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 **X**

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

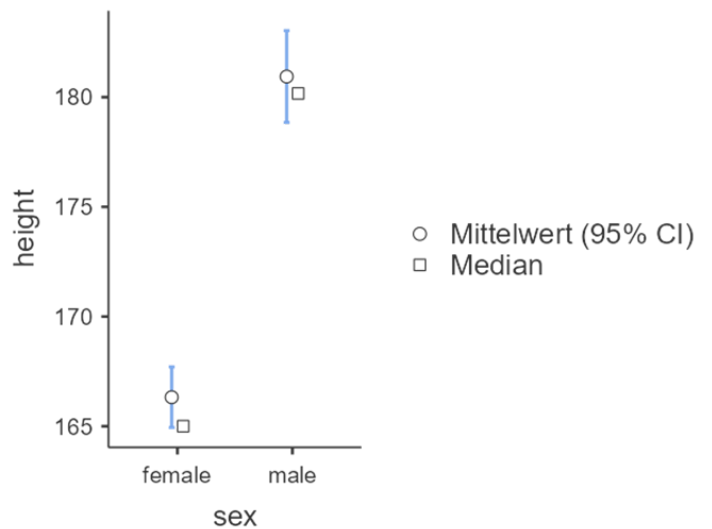
</ul>

#### 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

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

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

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

Effektstärke:

$$\text{Cohens } d = \frac{\text{Mittelwert}_1 - \text{Mittelwert}_2}{\text{gepoolte Standardabweichung}}$$

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

$$BF_{10} = \frac{P(\text{Beobachtete Daten} \mid H_1 \text{ wahr})}{P(\text{Beobachtete Daten} \mid H_0 \text{ wahr})}$$

## 2.3 Testen von Zusammenhängen

Übersicht:

Variable 2	Variable 1	Numerisch	Ordinal	Nominal	
				nicht-binär	binär
Numerisch		Pearson Korrelation	Spearman $\rho$ Kendall $\tau$	Eta Quadrat	t-Test Punkt-Biserial Korrelation
					Mann-Whitney-U Test Cramers V
Ordinal	nicht-binär		Spearman $\rho$ Kendall $\tau$	Chi2- Test	Chi2- Test
Nominal	binär				Chi2- Test Exakter Test nach Fisher

### 2.3.1 Pearson Korrelationskoeffizient

Die Kovarianz misst die lineare Beziehung zwischen zwei Variablen X und Y:

$$\text{cov}(X, Y) = \frac{(x_1 - \text{Mittelwert}_x)(y_1 - \text{Mittelwert}_y) + \dots + (x_n - \text{Mittelwert}_x)(y_n - \text{Mittelwert}_y)}{n - 1}$$

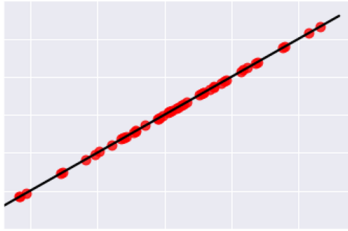
Da die Kovarianz von der Einheit der Messungen abhängt wird in der Praxis eine normierte Variante der Kovarianz verwendet - der *Pearson Korrelationskoeffizient*  $r$ :

$$r = \frac{\text{cov}(X, Y)}{\text{Standardabweichung}_X \cdot \text{Standardabweichung}_Y}$$

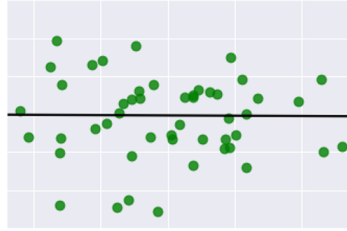
Der Pearson Korrelationskoeffizient kann nur Werte zwischen -1 und +1 annehmen, wobei -1 eine perfekte negative lineare Beziehung und +1 eine perfekte positive lineare Beziehung anzeigt:



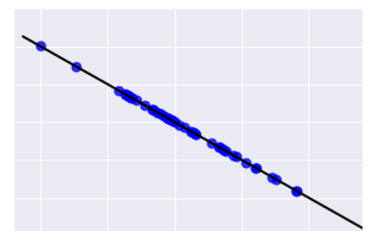
$r_{x,y} = 1$   
perfectly positive linear



$r_{x,y} = 0$   
not linear



$r_{x,y} = -1$   
perfectly negative linear



### Beispiel: Korrelationstest mit Pearson Korrelationskoeffizient (Pearsons r)

#### Schritt 1: Problemstellung und Hypothesen formulieren

Wir wollen wissen, ob Größe und Gewicht positiv korreliert sind. Wir müssen also testen, ob der Korrelationskoeffizient signifikant positiv ist.

$$H_0 : r_{GrößeGewicht} \leq 0$$

$$H_1 : r_{GrößeGewicht} > 0$$

#### Schritt 2: Passenden Test auswählen

Da Größe und Gewicht beide numerisch → Pearson Korrelationskoeffizient

#### Schritt 3: Voraussetzungen des Tests überprüfen

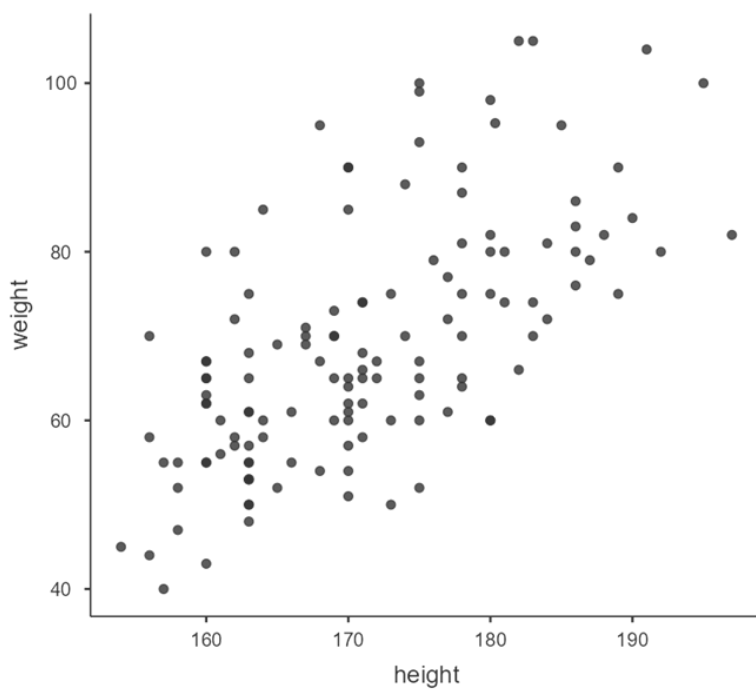
##### Voraussetzungen Pearson Korrelationstest

- **PK1. Beide Variablen sind numerisch.**
- **PK2. Normalität.** Die Variablen sind normalverteilt.
- **PK3. Unabhängigkeit.** Die Messungen sind unabhängig.

Besprechung der Voraussetzungen → Vorlesung.

#### Schritt 4: Voranalyse

Analysen -> Exploration -> Streudiagramm



### Schritt 5: Test durchführen und interpretieren

Korrelationsmatrix

		height	weight
height	Pearson's r	—	
	df	—	
	p-Wert	—	
	N	—	
weight	Pearson's r	0.640***	—
	df	130	—
	p-Wert	< .001	—
	N	132	—

Anmerkung.  $H_a$  ist eine positive Korrelation

Anmerkung. \*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$ , einseitig

Analysen -> Regression -> Korrelationsmatrix

Interpretation	Correlation value
Small correlation	0.10 to 0.29
Medium correlation	0.30 to 0.49
Large correlation	0.50 to 1.0

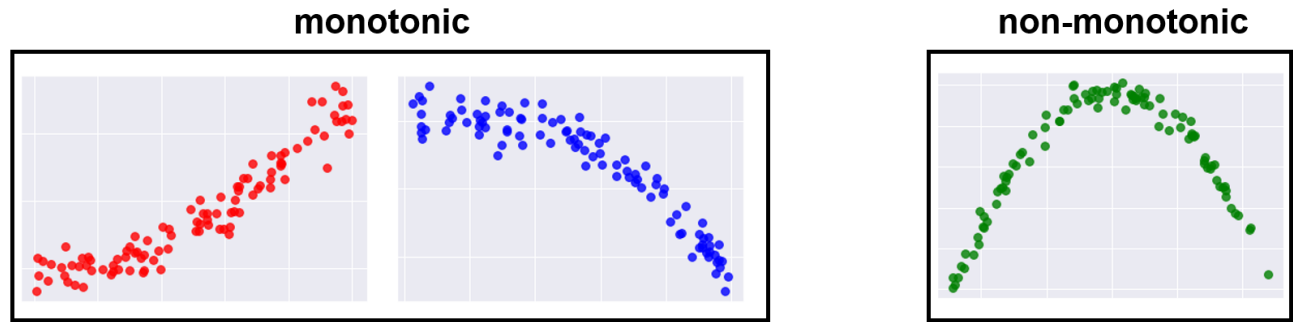
Außerdem ist  $r^2 = 0,64^2 = 40,96\%$  der Anteil der Varianz, den die Variablen teilen und der somit erklärt wird.

### 2.3.2 Spearman Rho und Kendall Tau Korrelationskoeffizient

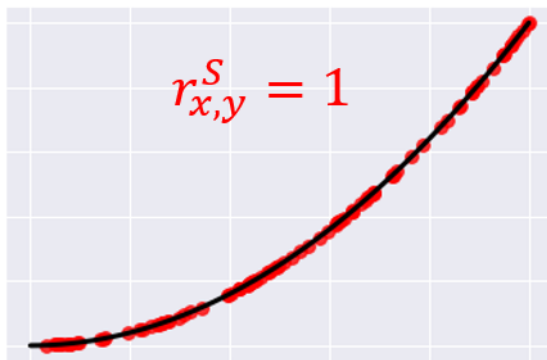
Im Falle ordinaler Variablen können wir die Kovarianz nicht berechnen, da wir keinen Mittelwert berechnen

können. Korrelationskoeffizienten werden mithilfe von *rangbasierten* Ansätzen bestimmt, die die Daten der Größe nach ordnen und jeder Beobachtung entsprechend ihrer Position einen Rang zuweisen. Beliebte rangbasierte Korrelationskoeffizienten sind: Spearmans Rho und Kendalls Tau.

Rangbasierte Ansätze identifizieren allgemeinere monotone Zusammenhänge:



Ein Koeffizient von 1 zeigt eine perfekt positive monotone Beziehung an:



#### Voraussetzungen rangbasierte Korrelationstests

- **RK1. Ordinal.** Beide variablen sind mindestens ordinal.
- **RK2. Unabhängigkeit.** Die Messungen sind unabhängig.

#### Spearmans Rho:

Dieser Koeffizient funktioniert genauso wie Pearsons r, mit dem Unterschied, dass er die Kovarianz und die Standardabweichungen in Bezug auf die Ränge anstelle der Werte der Variablen berechnet.

$$r^S = \frac{\text{cov}(\text{Rang}(X), \text{Rang}(Y))}{\text{Standardabweichung}_{\text{Rang}(X)} \cdot \text{Standardabweichung}_{\text{Rang}(Y)}}$$

#### Kendall Tau:

"Diese Koeffizienten basieren auf der Anzahl der konkordanten und diskordanten Paare in einem Datensatz. Gegeben zwei Variablen X und Y, sind zwei Beobachtungspaare  $(x_i, y_i)$  und  $(x_j, y_j)$

- *konkordant* wenn  $x_i > x_j$  and  $y_i > y_j$  oder if  $x_i < x_j$  and  $y_i < y_j$  Beispiele:
  - (1,3) und (6,9) sind konkordant
  - (3,1) und (6,9) sind diskordant

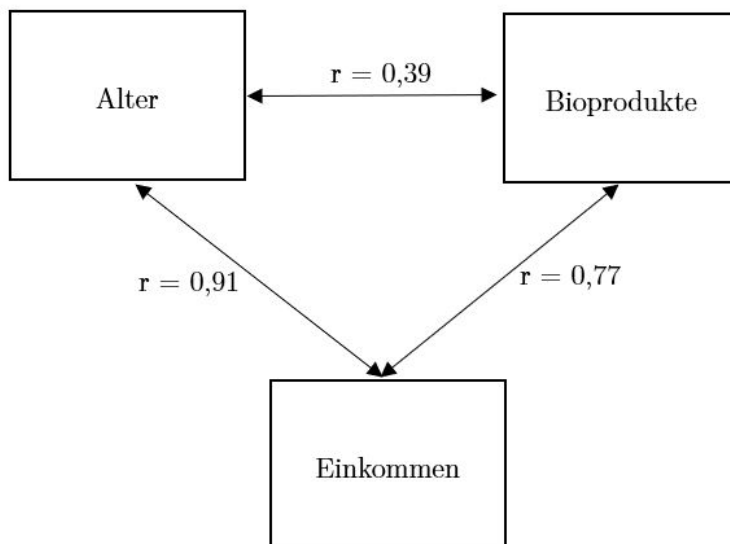
Beziehung zwischen dem Spearman und Kendall Koeffizient:

$$\text{Kendall} \approx 0.7 \cdot \text{Spearman}$$

Ausführlichere Beispiele zu Spearman Rho und Kendall Tau → Vorlesung.

### 2.3.3 Partielle Korrelation

Beim Testen von Korrelationen müssen wir potenzielle Störvariablen berücksichtigen. Angenommen, wir möchten testen, ob das Alter mit dem Kauf von Bio-Produkten korreliert. Dann müssen wir auch berücksichtigen, dass das Alter mit dem Einkommen korreliert, das wiederum mit dem Kauf der (teureren) Bio-Produkte korreliert sein könnte.



Die Korrelation von 0,39 könnte zum Teil auf die positive Korrelation zwischen Alter und Einkommen zurückzuführen sein. Daher müssen wir den Effekt des Einkommens eliminieren. Die partielle Korrelationsanalyse bietet eine Möglichkeit, dies zu tun.

Partieller Korrelationskoeffizient:

Seien X, Y und Z drei Variablen. Angenommen, wir möchten die Korrelation zwischen X und Y untersuchen, während wir für Z kontrollieren. Der angepasste Korrelationskoeffizient ist dann:

$$r_{XY,Z} = \frac{r_{XY} - r_{XZ} \cdot r_{YZ}}{\sqrt{1 - r_{XZ}^2} \cdot \sqrt{1 - r_{YZ}^2}}$$

### 2.3.4 Unabhängigkeitstests

Um zu testen, ob eine nominale und eine nominale oder ordinale Variable miteinander assoziiert sind, können wir einen Unabhängigkeitstest verwenden:

Chi2 ( $\chi^2$ ) Test oder exakter Test nach Fisher (bei  $2 \times 2$  Kontingenztafeln)

Hypothesen:

$H_0$  : Die Variablen X und Y sind unabhängig und  $H_1$  : Die Variablen X und Y sind abhängig.

Beispiel: Unabhängigkeitstest

**Schritt 1:** Problemstellung und Hypothesen formulieren

Wir wollen wissen, ob Haarfarbe und Augenfarbe voneinander abhängen.

$H_0$  : Haarfarbe und Augenfarbe sind unabhängig

$H_1$  : Haarfarbe und Augenfarbe sind abhängig

## Schritt 2: Passenden Test auswählen

Da Haarfarbe (black, blonde, brown, red) und Augenfarbe (blue, brown, green) zu einer  $4 \times 3$ -Tafel führen → Chi2-Test

## Schritt 3: Voraussetzungen des Tests überprüfen

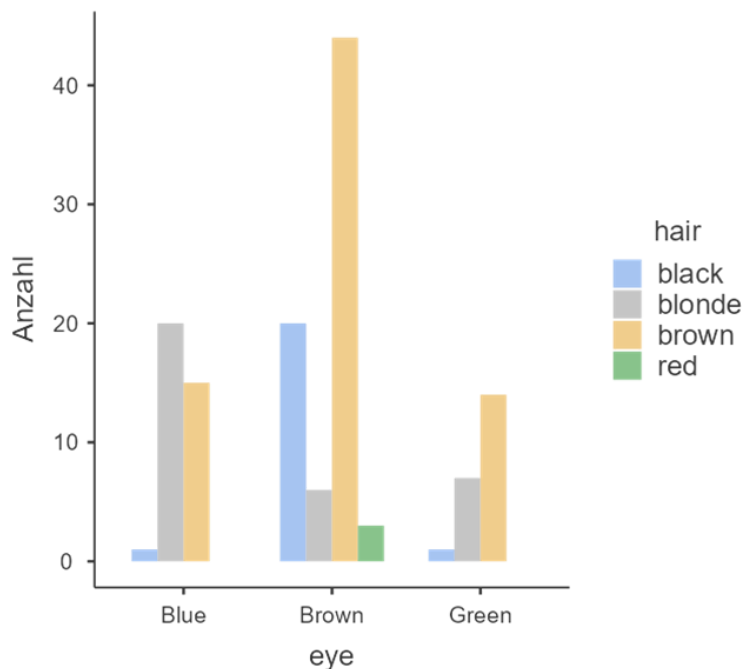
### Vorraussetzungen Chi2-Unabhängigkeitstest

- **C1. Beide Variablen sind kategorial.**
- **C2. Große Stichprobe.** Daumenregel:  $n > 50$ .
- **C3. Hinreichend große erwartete Häufigkeiten.** Alle erwarteten Häufigkeiten  $> 5$ .
- **C4. Unabhängigkeit.** Messungen sind unabhängig.

Besprechung der Voraussetzungen → Vorlesung.

## Schritt 4: Voranalyse

Analysen -> Häufigkeiten -> Unabhängige Stichproben -> Balkendiagramm



Analysen -> Häufigkeiten -> Unabhängige Stichproben -> Anzahl Beobachtet / Erwartet

Kreuztabellen					
eye	hair				Insgesamt
	black	blonde	brown	red	
Blue	1	20	15	0	36
Brown	20	6	44	3	73
Green	1	7	14	0	22
Insgesamt	22	33	73	3	131

Kreuztabellen						
eye		hair				Insgesamt
		black	blonde	brown	red	
Blue	Erwartet	6.05	9.07	20.1	0.824	36.0
Brown	Erwartet	12.26	18.39	40.7	1.672	73.0
Green	Erwartet	3.69	5.54	12.3	0.504	22.0
Insgesamt	Erwartet	22.00	33.00	73.0	3.000	131.0

Wie berechnet man die erwarteten Häufigkeiten E?

$$E_{\text{Zeile } i, \text{Spalte } j} = \frac{(\text{Beobachtet Zeile } i) \times (\text{Beobachtet Spalte } j)}{\text{Beobachtungen Gesamt}}$$

Beispiel:

$$E_{\text{eye blue, hair black}} = \frac{36 \times 22}{131} \approx 6,05$$

Wie berechnet man die Chi2-Teststatistik?

Allgemein für eine  $n \times m$  Kontingenztafel mit B = beobachtete Häufigkeit und E = erwartete Häufigkeit:

$$\chi^2 = \frac{(B_{1,1} - E_{1,1})^2}{E_{1,1}} + \dots + \frac{(B_{n,m} - E_{n,m})^2}{E_{n,m}}$$

Im Beispiel:

$$\chi^2 = \frac{(1 - 6,05)^2}{6,05} + \dots + \frac{(0 - 0,504)^2}{0,504} \approx 37,1$$

Wie bestimmt man die Freiheitsgrade eines Chi2-Tests?

$$\text{Freiheitsgrade} = (\text{AnzahlZeilen} - 1) \times (\text{AnzahlSpalten} - 1)$$

Im Beispiel:

$$\text{Freiheitsgrade} = 2 \times 3 = 6$$

**Schritt 5:** Test durchführen und interpretieren

Analysen -> Häufigkeiten -> Unabhängige Stichproben -> Tests Chi2

$\chi^2$ -Tests

	Wert	df	p
$\chi^2$	37.1	6	< .001
N	131		

Effektstärke:

Analysen -> Häufigkeiten -> Unabhängige Stichproben -> Phi und Cramers V

Nominal

	Wert
Phi-Koeffizient	NaN
Cramer's V	0.377

$$\text{Cramers } V = \sqrt{\frac{\chi^2/n}{\min(Z-1, S-1)}},$$

wobei n = Stichprobengröße, Z = Anzahl Zeilen, S = Anzahl Spalten

<u>Value of <math>\phi</math> or Cramer's V</u>	<u>Description</u>
.00 and under .10	Negligible association
.10 and under .20	Weak association
.20 and under .40	Moderate association
.40 and under .60	Relatively strong association
.60 and under .80	Strong association
.80 to 1.00	Very strong association

Quelle: Rea, L. M., and Parker, R. A. (1992). Designing and conducting survey research. San Francisco: Jossey-Boss.

## 2 × 2-Kontingenztafel

Wenn jede Variable 2 Kategorien hat, kann man entweder einen Chi2-Test mit Kontinuitätskorrektur oder einen exakten Test nach Fisher (Voraussetzungen siehe unten) durchführen:

Analysen -> Häufigkeiten -> Unabhängige Stichproben -> Tests

Voraussetzungen exakter Test nach Fisher

- **E1. Binäre kategoriale Variablen.**
- **E2. Unabhängigkeit.**

Effektstärken bei 2 × 2-Kontingenztafeln

Analysen -> Häufigkeiten -> Unabhängige Stichproben -> Vergleichene Maße

Gegeben folgende Kontingenztafel mit beobachteten Häufigkeiten a, b, c, d:

a	b
c	d

$$Odds - Ratio(OR) = \frac{a/b}{c/d}$$

$$Relatives Risiko(RR) = \frac{a/(a+b)}{c/(c+d)}$$

Interpretation → Vorlesung.

## 2.4 Multiples Testen

Problem: Multiple Tests führen zu einer Alphafehler-Inflation

Mögliche Korrekturen (p-Wert-Anpassungen) bei multiplen Tests:

**Bonferroni-Korrektur:**

$$p_{bonf} = p_{unangepasst} \times (\text{Anzahl Tests})$$

**Sidak-Korrektur:**

$$p_{sid} = 1 - (1 - p_{unangepasst})^{(\text{Anzahl Tests})}$$

### Holm-Bonferroni-Korrektur:

- Sortiere die unangepassten p-Werte von niedrig nach hoch:  $p(1) < \dots$  Passe den  $i$ -ten p-Wert wie folgt an:

$$p_{hbonf} = (\text{Anzahl Tests} - i + 1) \cdot p_{unangepasst}$$

### Benjamini Hochberg:

- Sortiere die unangepassten p-Werte von niedrig nach hoch:  $p(1) < \dots$  Multipliziere jeden p-Wert mit der Anzahl der Tests und dividiere ihn durch seinen Rang.  $(\frac{p \cdot t}{i})$
- Die resultierende Sequenz sollte nicht abnehmen. Falls sie abnimmt, setze den vorherigen p-Wert gleich dem nachfolgenden. Wiederhole diesen Schritt, bis die Sequenz nicht mehr abnimmt.

Beispiele → Vorlesung.

Die Anpassungen können von konservativ (lehnen die Nullhypothese seltener) ab nach liberal geordnet werden:

konservativ -- Bonferroni -- Sidak -- Holm-Bonferroni -- Benjamini-Hochberg