Vertiefte Forschungsmethodik der Psychologie und Psychotherapie I

I. Messtheorie

Dr. Leonard Pleschberger

Heinrich-Heine-Universität Düsseldorf

WS 25/26

Übersicht

I.1. Messtheorie

- Die **Statistik** wertet Daten mit mathematischen Modellen aus. Damit wir die Modelle rechnen können, müssen die Daten in Form von Zahlen vorliegen.
- Die **Messtheorie** befasst sich mit der korrekten Zuordnung von Zahlen zu Objekten der Realität, sodass deren interessanten empirischen Relationen durch die zugewiesenen numerischen Daten in ihrer Struktur erhalten bleiben. Ist das der Fall, sprechen wir von einem Homomorphismus.

$$\text{Realität} \xrightarrow{\text{Messtheorie}} \text{Daten} \xrightarrow{\text{Statistik}} \text{Inferenz}$$

Beispiel: Valenz von Emotionen

Folgende empirischen Relationen sind **unmittelbar einsichtig**:
„$\prec$“ : „weniger angenehm“, „$\asymp$“ : „genauso angenehm“

Wir ordnen den Valenzen folgende Daten von -5 bis $+5$ zu:

-5	$<$	**0**	$<$	**+1**	$=$	**+1**	$<$	**+4**	✔
-2	$>$	**-3**	$<$	**+1**	$<$	**+2**	$<$	**+4**	✘

1. Zeile: Homomorphismus.
2. Zeile: Kein Homomorphismus: Die numerischen Relationen stimmen nicht mit den empirischen Relationen überein.

Psychologische Merkmale

- Eine beobachtbare Verhaltensweise, i.e. ein **manifestes Merkmal** oder **Indikator**, kann direkt gemessen, also operationalisiert, werden.
- Im Gegensatz dazu sind psychologische Merkmale wie Eigenschaften nicht unmittelbar und sinnlich wahrnehmbar, i.e. nicht phänomenal gegeben. Diese **latenten Merkmale** müssen vielmehr aus manifesten Merkmalen konstruiert werden.
- Problematisch sind daher sowohl die theoretische Rechtfertigung der Operationalisierung als auch die Festlegung der Gewichtung der einzelnen Indikatoren.

Manifeste Merkmale (beobachtbar)

Latente Merkmale (Konstrukte)

Beispiel Konstrukte: Stressbelastung

Manifeste Merkmale

(direkt beobachtbare Indikatoren)

- **Erhöhter Blutdruck**
- **Erhöhter Cortisolspiegel** (Laborwert)
- **Muskelverspannungen** (z.B. Nacken)
- **Magen-Darm-Beschwerden**
- **Häufige Arztbesuche oder Krankschreibungen**

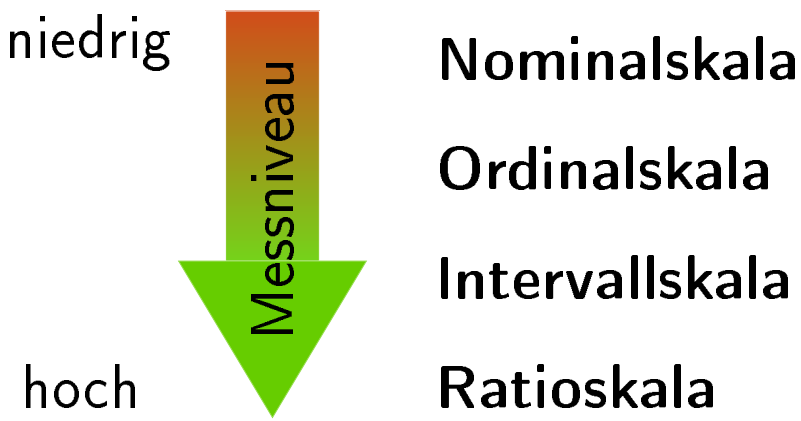
Latentes Merkmal

(nicht direkt beobachtbar)

- **Stressbelastung**
  - psychologisches Konstrukt
  - nicht direkt messbar
  - muss über Indikatoren erschlossen werden

I.2. Messniveaus und Skalentypen

- Eine **Skala** ist eine Regel für die Zuordnung von Zahlen zu Beobachtungen. Auf diese Weise werden Daten generiert.
- Die Skalen bestimmen, welche mathematischen Operationen und statistischen Verfahren auf den erzeugten Daten zulässig sind.
- Je nach Merkmal können Messungen unterschiedlich exakt vorgenommen werden. Es werden meist vier **Messniveaus** unterschieden, die die folgenden Skalen ergeben.



Nominalskala: Beispiel Blutgruppen

- **Merkmal**: Blutgruppe.
- **Ausprägungen**: 0, A, B, AB.
- **Warum Nominalskala?**
  - Blutgruppen sind Kategorien ohne Rangordnung.
  - Zwei Datenpunkte sind entweder „gleich“ oder „verschieden“.
- **Statistische Verfahren**: Häufigkeiten zählen, Modus bestimmen.

Blutgruppe	n=200	Anteil in %
0	82	41%
A	86	43%
B	22	11%
AB	10	5%

$\Rightarrow$ **Modus = A**

(häufigste Ausprägung)

Nominalskala: Analyse mit R

Exportiere ein Balkendiagramm mit gekennzeichneten Modus als .png:

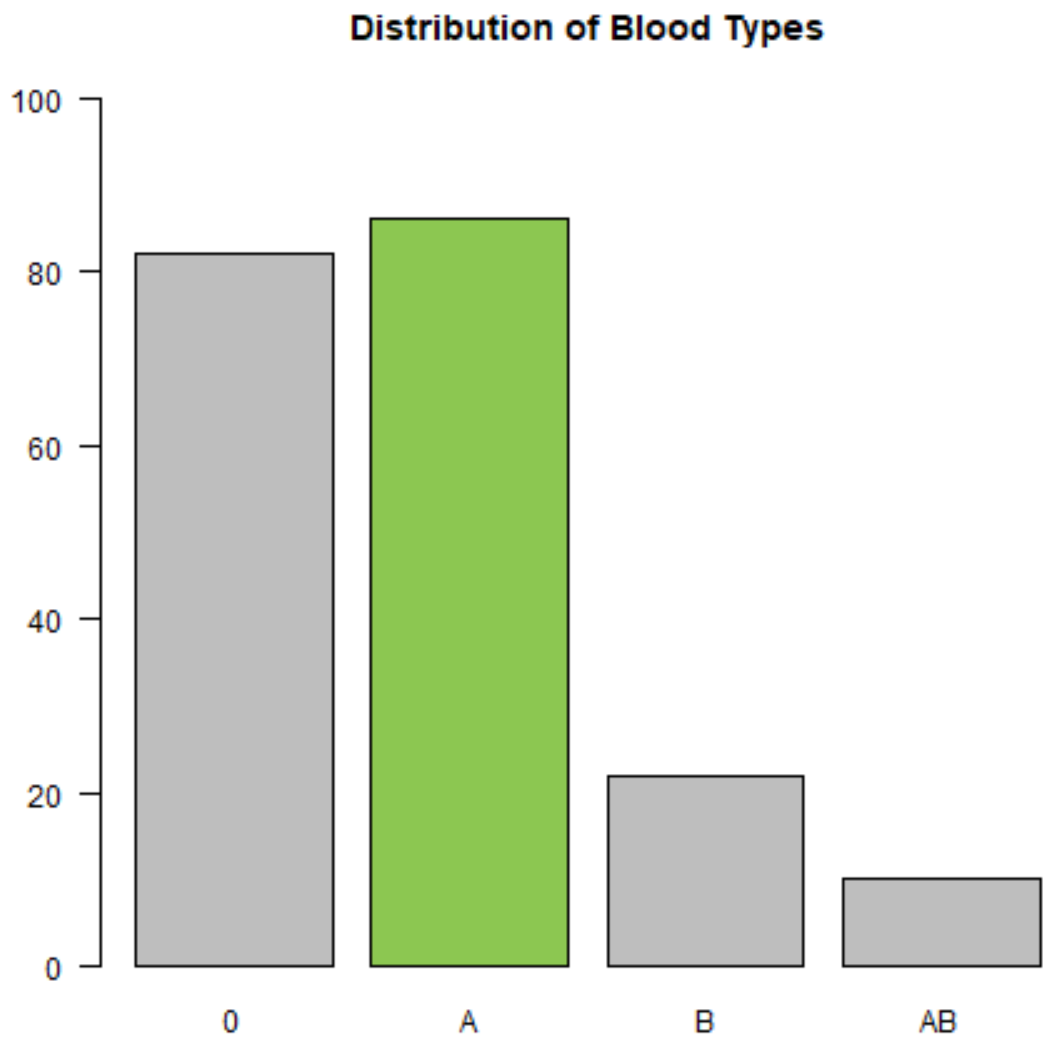
```r
# Create a virtual graphic device
png("Bloodtypes.png")

# Enter the data and color the mode bar
blood_types <- c("O", "A", "B", "AB")
counts <- c(82, 86, 22, 10)
colors[which.max(counts)] <- "#8cc751" # HTC Green

# Draw the bar chart; las: orientation of labels
barplot(counts, names.arg = blood_types,
main = "Distribution of Blood Types",
ylim=c(0,100), las=1, col=colors)

# Close the device and write to the file
dev.off()
```

Nominalskala: Fertiges Balkendiagramm



Ordinalskala: Beispiel Subjektive Schmerzintensität & Stress

- **Merkmal**: Subjektive Schmerzintensität & Stressbelastung.
- **Ausprägungen**:
  0 : Kein Schmerz / Stress.
  1 : Leichter Schmerz / Stress.
  2 : Mäßiger Schmerz / Stress.
  3 : Starker Schmerz / Stress.
  4 : Unerträglicher Schmerz / Stress.
- **Warum Ordinalskala?**
  - Es gibt eine klare Rangfolge: $0 < 1 < 2 < 3 < 4$.
  - Die Abstände sind nicht eindeutig interpretierbar:
    Der Unterschied von $1 \to 2$ könnte kleiner sein als der von $3 \to 4$.
- **Statistische Verfahren**: Deskriptive Statistik ohne Mittelwert und Varianz, Abhängigkeitsmaße.

Ordinalskala: Deskriptive Statistik (Modus)

Wir betrachten die folgenden Daten:

Schmerzintensität (n=10)	2	1	3	2	4	2	1	3	2	1

Zunächst ermitteln wir die Anzahl der einzelnen Ausprägungen. Daraus ergibt sich der Modus.

Skalenwert	0	1	**2**	3	4
Häufigkeit	0	3	**4**	2	1

$\Rightarrow$ **Modus** $= 2$

(häufigste Ausprägung)

Dies lässt sich wie bei den Blutgruppen gut als Balkendiagramm visualisieren. (Übung!)

Ordinalskala: Deskriptive Statistik (Median, Perzentile)

Wir können zwei
die Daten zusätzlich noch der Größe nach sortieren. Die Position der Werte heißt **Rang**.

Rang	1	2	3	4	5	6	7	8	9	10
Daten (sortiert)	1	1	1	2	2	2	2	3	3	4

Wir können das 25., 50. & 75. **Perzentil** - oder: Q1, Q2 & Q3 - wie folgt berechnen:

$$25.\text{ Perzentil: } Q1 = 0.25 \cdot (n+1) = 2.75.$$

Der Wert 2.75 liegt zwischen Rang 2 und Rang 3. Daher gilt $Q1 = 1$.
Analog folgen: $Q2 =$ **Median** $= 2$ und $Q3 = 3$. (Übung!)

Ordinalskala: Abhängigkeitsmaße mit R

Wir können den monotonen Zusammenhang zweier ordinalskalierter Merkmale mit einem Wert in $[-1, 1]$ bestimmen.

Schmerzintensität	1	1	1	2	2	2	2	3	3	4
Stressbelastung	3	2	3	2	4	3	1	3	1	1

Übliche Kennzahlen sind **Spearman's** ϱ und **Kendall's** τ. Wir erhalten:

```
pain <- c(2, 1, 3, 2, 4, 2, 1, 3, 2, 1)
stress <- c(3, 2, 3, 2, 4, 3, 1, 3, 1, 1)

# Spearman's rank correlation
spearman <- cor(pain, stress, method = "spearman")

# Kendall's Tau
kendall <- cor(pain, stress, method = "kendall")
```

Ordinalskala: Abhängigkeitsmaße interpretiert

Diese Rangkorrelationskoeffizienten bewerten den Zusammenhang zwischen Schmerzintensität und Stressbelastung wie folgt:

Maß	**Wert**	**Interpretation**
Spearman's ρ	0.84	Starke positive monotone Beziehung.
Kendall's τ	0.63	Moderat positive monotone Beziehung.

Spearman's ϱ arbeitet mit der Differenz der Ränge; Kendall's τ hingegen lediglich mit sich unterscheidenden Rängen. Bei kleinen Stichproben ($n < 30$) und vielen Bindungen, i.e. eine Merkmalsausprägung kommt öfter vor, ist Kendall's τ präziser und zu bevorzugen.

τ	0.00 - 0.09	0.10 - 0.49	0.50 - 0.69	0.70 - 1.00
Monotonie	Keine	(Sehr) schwach	Moderat	(Sehr) stark

Intervallskala: Beispiel Depressionsschwere

Viele psychologische Konstrukte wie *Intelligenz*, *Depression* oder *Stress* sind latente Merkmale. Man erhebt sie indirekt über ordinalskalierte Items, e.g. Tests mit Likert-Skalen. Oft werden diese Ordinaldaten addiert oder gemittelt. Damit behandelt man sie wie (quasi-)intervallskalierte Werte. Zudem setzt man oft theoretisch begründete Annahmen über kausale oder korrelative Beziehungen zu anderen Konstrukten voraus.

- **Merkmal**: Depressionsschwere (via Beck-Depressions-Inventar II).
- **Ausprägungen**: 0–63 Punkte.
- **Warum Intervallskala?**
  - Die Abstände sind interpretierbar:
    Unterschied von 10 $\rightarrow$ 15 Punkten ist genauso groß wie 25 $\rightarrow$ 30.
  - Es gibt keinen echten Nullpunkt: 0 Punkte bedeutet „keine Symptome“, aber nicht „Depression existiert nicht“.
- **Statistische Verfahren**: Sämtliche statistische Verfahren außer Verhältnisse zwischen Datenpunkten wie „doppelt so groß“.

Ratioskala: Beispiel Cortisolspiegel

Das höchste Messniveau weisen ratioskalierte Merkmale auf. Sie sind meist physiologischer Natur und besitzen einen eindeutigen Nullpunkt.

- **Merkmal**: Cortisolspiegel im Blut ($\mu g/dl$).
- **Ausprägungen**: I.d.R. 0-25 $\mu g/dl$.
- **Warum Ratioskala?**
  - Die Abstände sind interpretierbar:
    Unterschied von 5 $\rightarrow$ 10 $\mu g/dl$ ist genauso groß wie 20 $\rightarrow$ 25 $\mu g/dl$.
  - Es gibt einen echten Nullpunkt: 0 $\mu g/dl$ bedeutet wirklich „kein Cortisol im Blut“.
- **Statistische Verfahren**: Sämtliche statistische Verfahren und Verhältnisaussagen wie „doppelt so hoch“.

I.3. Messmodelle: Klassische Testtheorie (KTT)

Theoretische Annahme: Das beobachtete Merkmal setzt sich zusammen aus einem „echten“ Merkmal T (*true score*) und einem zufälligen Messfehler ε (*error*) zusammen. Addition liefert den Testwert

$$X = T + \varepsilon.$$

Über den Fehlerterm ε werden starke Annahmen gemacht:

- $\mathbb{E}[\varepsilon] = 0$. Fehler gleichen sich im Schnitt aus.
- $\mathrm{Cov}(T, \varepsilon) = 0$. Fehler und wahre Werte sind unkorreliert.
- Die Fehler ε sind paarweise unabhängig.

Messmodelle: Item-Response-Theorie (IRT)

- **Theoretische Annahme**: Die Wahrscheinlichkeit, dass eine Person ein Item X mit „Ja“ beantwortet (1), hängt von der latenten Fähigkeit ϑ und der Itemschwierigkeit σ ab.
- Die bedingte Wahrscheinlichkeit für eine positive Antwort lautet

$$\mathbb{P}(X = 1 \mid \vartheta) = \frac{e^{\vartheta - \sigma}}{1 + e^{\vartheta - \sigma}}$$

- Diese Gleichung beschreibt den Zusammenhang der beiden Parameter: Je größer ϑ im Verhältnis zu σ, desto höher ist die Lösungswahrscheinlichkeit.
- Wir möchten $\hat{\theta}$ schätzen und als Testwert annehmen - vermöge Maximum-Likelihood Estimation (MLE) oder Bayesscher Statistik.

Messmodelle: Gütekriterien psychologischer Tests

Ein psychologischer Test sollte sowohl in der KTT, als auch in der IRT mindestens den folgenden Kriterien genügen:

- **Objektivität**: Ein Test ist objektiv, wenn das Ergebnis unabhängig von der Person ist, die den Test durchführt, auswertet oder interpretiert. In der KTT ist **Cohen's** κ ein geeignetes Maß, welches Zufallstreffer berücksichtigt.
- **Reliabilität**: Ein Test ist reliabel, wenn er das zu messende Merkmal zuverlässig und frei von Zufallsfehlern misst. In der KTT ist **Cronbach's** α eine geeignete Kennzahl für interne Konsistenz.
- **Validität**: Ein Test ist valide, wenn er tatsächlich das misst, was er messen soll.

I.4. Fallstudie Depressionsschwere: Studiendesign

Fragestellung: Hilft eine kognitive Verhaltenstherapie (KVT), die Depressionsschwere zu reduzieren?

- **Instrument:** Beck-Depressions-Inventar II, Intervallskala 0–63.
- **Stichprobe:** Gruppe A (Therapie) vs. Gruppe B (Kontrolle), je $n = 6$.
- **Messzeitpunkte:** Prä (vor Therapie) und post (nach Therapie).
- **Bemerkung**: Um Berechnungen zur Korrelation durchführen zu können, betrachten wir weiterhin das Merkmal *Stresslevel*. Die Ausprägung wird durch die Summation von 10 ordinalen Items à 0-4 Punkten bestimmt und gilt damit praktisch als (quasi-) intervallskaliert.

Fallstudie Depressionsschwere: Daten

In dem Selbstbeurteilungsbogen BDI-II werden 21 Items mit je 0-3 Punkten aufsummiert, um die Schwere einer Depression einzustufen.

Gruppe A (Therapie), prä	28	24	26	20	22	23
Gruppe A (Therapie), post	18	17	15	20	14	15
Gruppe B (Kontrolle), prä	15	11	14	12	10	14
Gruppe B (Kontrolle), post	13	11	15	13	12	13

Das Stresslevel aller Teilnehmer wird prä-therapeutisch durch Summation von Items bestimmt.

Stresslevel	32	28	30	26	25	27	20	18	22	21	19	23

Fallstudie Depressionsschwere: Deskriptive Statistik mit R

Wir laden die Daten in R und verschaffen uns mittels der Funktion `describe()` einen Überblick über die Daten.

```
# Daten eingeben
A_pre  <- c(28, 24, 26, 20, 22, 23)
A_post <- c(18, 17, 15, 20, 14, 15)

B_pre  <- c(15, 11, 14, 12, 10, 14)
B_post <- c(13, 11, 15, 13, 12, 13)

# Deskriptive Statistik
library(psych)
describe(data.frame(A_pre, A_post, B_pre, B_post))
```

Fallstudie Depressionsschwere: Deskriptive Statistik mit R

Wichtig sind der **Mittelwert** $\bar{x}$, die **empirische Standardabweichung** s, bzw. **empirische Varianz** s^2, gegeben durch

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

	$\bar{x}$	s	s^2
Gruppe A (Therapie), prä	23.83	2.72	7.40
Gruppe A (Therapie), post	16.50	2.15	4.62
Gruppe B (Kontrolle), prä	12.67	1.87	3.50
Gruppe B (Kontrolle), post	12.83	1.27	1.61

Interpretation: Gruppe A startet mit höherer Depressionsschwere und zeigt bereits im Schnitt einen Rückgang (-7,3 Punkte).

Fallstudie Depressionsschwere: Unabhängiger t-Test in R

Fragestellung: Waren die beiden prä-Gruppen gleich stark betroffen?

In R kann man mit einer Zeile Code einen unabhängigen *t*-Test durchführen:

```
# Independent t-Test: A_pre vs. B_pre
t_indep <- t.test(A_pre, B_pre, var.equal = TRUE)
```

Dieser führt zu folgendem Ergebnis:

```
Two Sample t-test

data:  A_pre and B_pre
t = 7.8851, df = 10, p-value = 1.336e-05
alternative hypothesis: true difference in means is
    not equal to 0
```

Fallstudie Depressionsschwere: Unabhängiger t-Test

Wir testen also, ob die Erwartungswerte μ_A und μ_B der beiden Gruppen signifikant zum Niveau $\alpha = 0.05$ voneinander abweichen. Es ergeben sich die Nullhypothese H_0 und Alternative H_1 wie folgt:

$$H_0 : \mu_A = \mu_B \quad \text{vs.} \quad H_1 : \mu_A \neq \mu_B.$$

Wir gehen davon aus, dass die beiden Gruppen jeweils normalverteilt sowie unabhängig voneinander sind und die gleichen Varianzen besitzen. Um die Hypothesen zu testen eignet sich also einen zweiseitiger unabhängiger t-Test. Es gilt die Testvorschrift

$$\varphi(x_1, ..., x_n, y_1, ..., y_m) = \begin{cases} 0, & |t| < t_{1-\alpha/2, n+m-2}, \\ 1, & |t| > t_{1-\alpha/2, n+m-2}. \end{cases}$$

Fallstudie Depressionsschwere: Unabhängiger t-Test

Wir müssen die Teststatistik t aus den Beobachtungen berechnen. Aus der **gepoolten Standardabweichung**

$$s_p = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}}$$

und den **Standardfehler**

$$SE = s_p \cdot \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$$

berechnen wir die **empirische t-Statistik**

$$t = \frac{\bar{x}_A - \bar{x}_B}{SE}.$$

Diese ergibt gerade $t \approx 7.89$, wie zuvor von R berechnet.

Fallstudie Depressionsschwere: Unabhängiger t-Test

Zuletzt bestimmen wir noch numerisch das 0.975-Quantil der t-Verteilung mit $n+m-2=10$ Freiheitsgraden (df). Das ergibt $t_{0.975,10} \approx 2.2$. In die Testvorschrift eingesetzt erhalten wir

$$\varphi = 1,$$

da $|t| = 7.9 > t_{0.975,10} = 2.2$. Demnach können wir die Nullhypothese signifikant zum Niveau $\alpha = 5\%$ verwerfen. Also ist eine der beiden Gruppen signifikant stärker betroffen als die andere. Unsere deskriptive Statistik ergibt, dass dies die Gruppe A ist. Wie signifikant die Testaussage wirklich ist, berichtet uns der p-**Wert**

$$p = 2 \cdot \mathbb{P}(T > |t|) = 2 \cdot (1 - F(7.9)) \approx 0.00001 << 0.001$$

mit der Zufallsvariablen $T \sim t(df = 10)$ und der Verteilungsfunktion F der $t(df = 10)$-Verteilung $\Rightarrow$ Die Nullhypothese wird sehr klar verworfen.

Fallstudie Depressionsschwere: Gepaarter t-Test in R

Fragestellung: Wirkt die Therapie in Gruppe A?

R liefert in einer Zeile Code den folgenden *t*-Test:

```
# Right-tailed paired t-test: A_pre vs. A_post
t_paired <- t.test(A_pre, A_post, paired = TRUE,
    alternative = "greater")
```

Dieser führt zu folgendem Ergebnis:

```
Paired t-test

data:  A_pre and A_post
t = 4.6277, df = 5, p-value = 0.002848
alternative hypothesis: true mean difference is
    greater than 0
```

Fallstudie Depressionsschwere: Gepaarter t-Test

Wir überprüfen also, ob sich der Erwartungswert μ_{post} der post-Therapiegruppe signifikant zum Niveau $\alpha = 0.05$ geringer ist, als der Erwartungswert $\mu_{\text{prä}}$ der prä-Therapiegruppe. Es ergeben sich die Nullhypothese H_0 und Alternative H_1 wie folgt:

$$H_0 : \mu_{\text{prä}} \leq \mu_{\text{post}} \quad \text{vs.} \quad \mu_{\text{prä}} > \mu_{\text{post}}$$

Klarerweise bilden die prä- und post-Therapiegruppe eine paarweise verbundene Stichprobe. Wir nehmen an, dass die Grundgesamtheit normalverteilt ist. Wegen des kleinen Stichprobenumfangs setzen wir weiter voraus, dass die Differenzen in der Grundgesamtheit normalverteilt sind.

Fallstudie Depressionsschwere: Gepaarter t-Test

Um die Hypothesen zu testen eignet sich also einen rechtsseitiger gepaarter t-Test. Es gilt die Testvorschrift

$$\varphi(x_1, ..., x_n, y_1, ..., y_n) = \begin{cases} 0, & t < t_{1-\alpha,n-1}, \\ 1, & t \geq t_{1-\alpha,n-1}. \end{cases}$$

Wir müssen die Teststatistik t aus den Beobachtungen berechnen. Betrachte das **arithmetische Mittel der Differenzen**

$$\bar{D} = \frac{\sum_{i=1}^{n}(x_i - y_i)}{n}$$

und die **Standardabweichung der Differenzen**

$$s_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}((x_i - y_i) - \bar{d})^2}.$$

Fallstudie Depressionsschwere: Gepaarter t-Test

Wir berechnen die **empirische t-Statistik**

$$t = \sqrt{n}\frac{\bar{D}}{s_D}.$$

Diese ergibt sich zu $t \approx 4.63$ - der erste Ausgabewert von R. Zuletzt bestimmen wir noch das 0.95-Quantil der t-Verteilung mit $n - 1 = 5$ Freiheitsgraden. Das ergibt $t_{0.95,5} \approx 2.0$. In die Testvorschrift eingesetzt folgt

$$\varphi = 1,$$

da $t = 4.63 \geq t_{0.95,5} = 2.0$. Demnach können wir die Nullhypothese signifikant zum Niveau $\alpha = 5\%$ verwerfen. Also hat sich die Therapiegruppe nach der Behandlung signifikant verbessert.

Fallstudie Depressionsschwere: Gepaarter t-Test

Wie signifikant die Testaussage wirklich ist, berichtet uns der p-**Wert**

$$p = \mathbb{P}(T \geq t) = 1 - F(t), \;\; T \sim t(df = 5)$$

mit der Verteilungsfunktion F der t-Verteilung mit 5 Freiheitsgraden. Die Verteilungsfunktion erhalten wir über Integration der Dichtefunktion, für allgemeine ν Freiheitsgrade gegeben durch

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi} \cdot \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

ergibt sich für $\nu = 5$ und $t = 4.63$ gerade

$$p = \frac{8}{3\sqrt{5}\pi} \int_{4.63}^{\infty} \left(1 + \frac{x^2}{5}\right)^{-3} \mathrm{d}x.$$

Fallstudie Depressionsschwere: Gepaarter t-Test

Leider gibt es für den Integranden keine einfache Stammfunktion. Aber wir können in R den Wert des Integrals numerisch ausrechnen:

```r
# Funktion definieren
f <- function(x) (1 + (x^2)/5)^(-3)

# Integral von 4.63 bis Inf
integral <- integrate(f, lower = 4.63, upper = Inf)

# p-Wert berechnen
p <- 8 / (3 * sqrt(5) * pi) * integral$value
```

Dies ergibt gerade $p \approx 0.0028 < 0.01$. Also wird die Nullhypothese klar verworfen.

Interpretation: Stark signifikanter Rückgang der Depressionswerte in der Therapiegruppe A nach der Behandlung.

Fallstudie Depressionsschwere: ANOVA in R

Fragestellung: Gibt es einen generellen Unterschied von prä zu post über beide Gruppen hinweg (Haupteffekt Zeit)? Verläuft die Veränderung über die Zeit in beiden Gruppen gleich (Interaktion)? In R:

```
# 2x2-ANOVA (group x time): Factor all 24 data points
    in long format (one row = one observation)

score <- c(A_pre, A_post, B_pre, B_post)
group <- factor(rep(c("A","B"), each=12))
time  <- factor(rep(rep(c("pre","post"), each=6), 2))

anova_data <- data.frame(score, group, time)

# Calculate the model
model <- aov(score ~ group * time, data = anova_data)
summary(model)
```

Fallstudie Depressionsschwere: ANOVA

Hypothesen: Es sollen also folgende Hypothesen über die Faktoren Gruppe (Therapie & Kontrolle) x Zeit (prä & post Therapie) und deren Interaktion mit einem F-Test bestätigt oder widerlegt werden.

Haupteffekt	**Nullhypothese H_0**	**Alternative H_1**
Gruppe	$\mu_{\text{A,prä}} = \mu_{\text{B,prä}}$ und $\mu_{\text{A,post}} = \mu_{\text{B,post}}$	$\mu_{\text{A,prä}} \neq \mu_{\text{B,prä}}$ oder $\mu_{\text{A,post}} \neq \mu_{\text{B,post}}$
Zeit	$\mu_{\text{prä}} = \mu_{\text{post}}$	$\mu_{\text{prä}} \neq \mu_{\text{post}}$
Interaktion	$(\mu_{\text{A,prä}} - \mu_{\text{A,post}}) = (\mu_{\text{B,prä}} - \mu_{\text{B,post}})$	$(\mu_{\text{A,prä}} - \mu_{\text{A,post}}) \neq (\mu_{\text{B,prä}} - \mu_{\text{B,post}})$

Fallstudie Depressionsschwere: ANOVA

Ergebnisse: Die folgenden F-Werte sind allesamt als F(1,20)-Werte mit Effekt-Freiheitsgrad 1 und Fehler-Freiheitsgraden 20 aufzufassen.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	1	330.0	330.0	69.85	5.89e-08
time	1	77.0	77.0	16.30	0.000644
group:time	1	84.4	84.4	17.86	0.000415
Residuals	20	94.5	4.7		

Im Folgenden werden die Sum of Squares je Effekt bezeichnet durch

$$SS_{\text{Effekt}}.$$

Fallstudie Depressionsschwere: ANOVA

Interpretation: Mit einem p-Wert von je < 0.001 sind alle nachfolgenden Hypothesen hochsignifikant nachgewiesen.

Haupteffekt	**Interpretation**
Gruppe	Die Therapiegruppe A hat insgesamt höhere Depressionswerte als die Kontrollgruppe B.
Zeit	Über beide Gruppen hinweg sinken die Depressionswerte von prä zu post.
Interaktion	Der Rückgang tritt fast ausschließlich in der Therapiegruppe A auf; in der Kontrollgruppe B bleiben die Werte praktisch stabil.

Fazit: Nur in der Therapiegruppe A sinkt der BDI-Wert signifikant – die Interaktion macht den Therapieeffekt deutlich.

Fallstudie Depressionsschwere: Effektstärke

Wir haben folgende Hypothesen statistisch signifikant nachgewiesen:

Verfahren	**Hypothese**
Unabhängiger t-Test	Prä-Gruppen nicht gleich stark betroffen.
Gepaarter t-Test	Die Therapiegruppe A hat sich verbessert.
ANOVA (Gruppe)	Gruppe A hat höhere Depressionswerte als Gruppe B.
ANOVA (Zeit)	In beiden Gruppen niedrigere Werte von prä zu post.
ANOVA (Interaktion)	Rückgang nur in Gruppe A; stabile Werte in Gruppe B.

Fallstudie Depressionsschwere: Effektstärke

Die Hypothesen besagen aber nur, *dass* Abweichungen vorliegen, jedoch nicht, *wie stark* diese Abweichungseffekte ausgeprägt sind. Maße für die Effektstärke sind **Cohen's d** für die t-Tests, bzw. η^2 für die ANOVA.

Mit den arithmetischen Mitteln der Stichproben $\bar{x}_A, \bar{x}_B$ und der gepoolten Standardabweichung s_p berechnen wir **Cohen's d für unabhängige Stichproben** mittels

$$d = \frac{\bar{x}_A - \bar{x}_B}{s_p}.$$

Das arithmetische Mittel der Differenzen $\bar{D}$ und die Standardabweichung der Differenzen s_D ergeben **Cohen's d für abhängige Stichproben** durch

$$d_z = \frac{\bar{D}}{s_D}.$$

Fallstudie Depressionsschwere: Effektstärke

Mit den Sum of Squares (SS) von Effekten und Fehler lässt sich das **partielle η^2** berechnen durch

$$\eta_p^2 = \frac{SS_{\text{Effekt}}}{SS_{\text{Effekt}} + SS_{\text{Fehler}}}.$$

Verfahren	**Effektmaß**	**Interpretation**
Unabhängiger t-Test	$d = 4.55$	Extrem großer Effekt.
Gepaarter t-Test	$d_z = 1.89$	Sehr großer Effekt.
ANOVA (Gruppe)	$\eta_p^2 = 0.78$	Sehr großer Unterschied.
ANOVA (Zeit)	$\eta_p^2 = 0.45$	Starker Rückgang.
ANOVA (Interaktion)	$\eta_p^2 = 0.47$	Starker Effekt.

Fallstudie Depressionsschwere: z-Transformation

Möchte man wissen, wie stark die Ausprägung X eines gemessenen Merkmals bei einer Person im Vergleich zum Durchschnitt der Gruppe ausfällt, wendet man die lineare z-Transformation an.

Dabei gilt mit dem arithmetischen Mittel $\bar{x}$ und der empirischen Standardabweichung s:

$$z = \frac{x - \bar{x}}{s}.$$

Beispiel: In unserer prä-Therapiegruppe ergibt ein BDI-Wert von 30 gerade

$$z = \frac{30 - 23.83}{2.72} \approx 2.27.$$

Interpretation: Der Wert liegt 2.27 Standardabweichungen über dem Mittelwert – die Person ist also deutlich stärker betroffen als der Durchschnitt der Gruppe.

Fallstudie Depressionsschwere: Korrelation & Regression

Wir vergleichen die erhobenen Stresslevel x_i mit den Depressionswerten y_i beider prä-Gruppen, also für $i = 1, ..., 12$, die folgenden Daten:

x	32	28	30	26	25	27	20	18	22	21	19	23
y	28	24	26	20	22	23	15	11	14	12	10	14

Als Maß für den linearen Zusammenhang berechnen wir aus der empirischen Kovarianz s_{xy} und den empirischen Standardabweichungen S_x und s_y die **Pearson-Korrelation**

$$r = \frac{s_{xy}}{s_x \cdot s_y} \approx 0.96.$$

Interpretation: Es besteht ein fast perfekt positiv linearer Zusammenhang zwischen Stresslevel und Depressionswerten.

Fallstudie Depressionsschwere: Korrelation & Regression

Es ergibt sich direkt das **Bestimmtheitsmaß**

$$R^2 = r^2 \approx 0.92.$$

Interpretation: 92% der Varianz der Depressionswerte wird durch das Stresslevel erklärt.

Für Prognosen modellieren wir den linearen Zusammenhang der beiden Merkmale mit der **Regressionsgeraden**

$$\hat{y} = a \cdot x + b = \frac{s_{xy}}{s_x^2} \cdot x + (\bar{y} - \frac{s_{xy}}{s_x^2} \cdot \bar{x}) \quad \Leftrightarrow \quad \hat{y} \approx 1.35x - 14.59.$$

Interpretation: Pro 1 Stresspunkt steigen die Depressionswerte im Mittel um circa 1.35 Punkte – Stressabbau könnte Depression deutlich senken.

Fallstudie Depressionsschwere: Gesamtinterpretation

Unsere bisherigen statistischen Analysen ergeben die folgenden Resultate:

- Die KVT zeigt hochsignifikant (p-Werte) eine starke Wirkung (Effektstärke).
- Die Depressionswerte sinken im Falle einer Therapie im Mittel um über 7 Punkte (deskriptive Statistik).
- Die Intervallskala ermöglichte den Einsatz parametrischer Verfahren (t-Tests, ANOVA, Korrelation, Regression).
- Pro Stresspunkt steigt der Depressionswert linear um 1.35 Punkte (Regression).