Regressionsmodelle

Teil 1: Einführung in die Lineare Regression

Thomas Zerjatke
thomas.zerjatke@tu-dresden.de

TU Dresden
Institut für Medizinische Informatik und Biometrie

WS 2023/24

Was ist eine Regression?

eine Variable durch eine andere erklären

Y auf X zurückführen

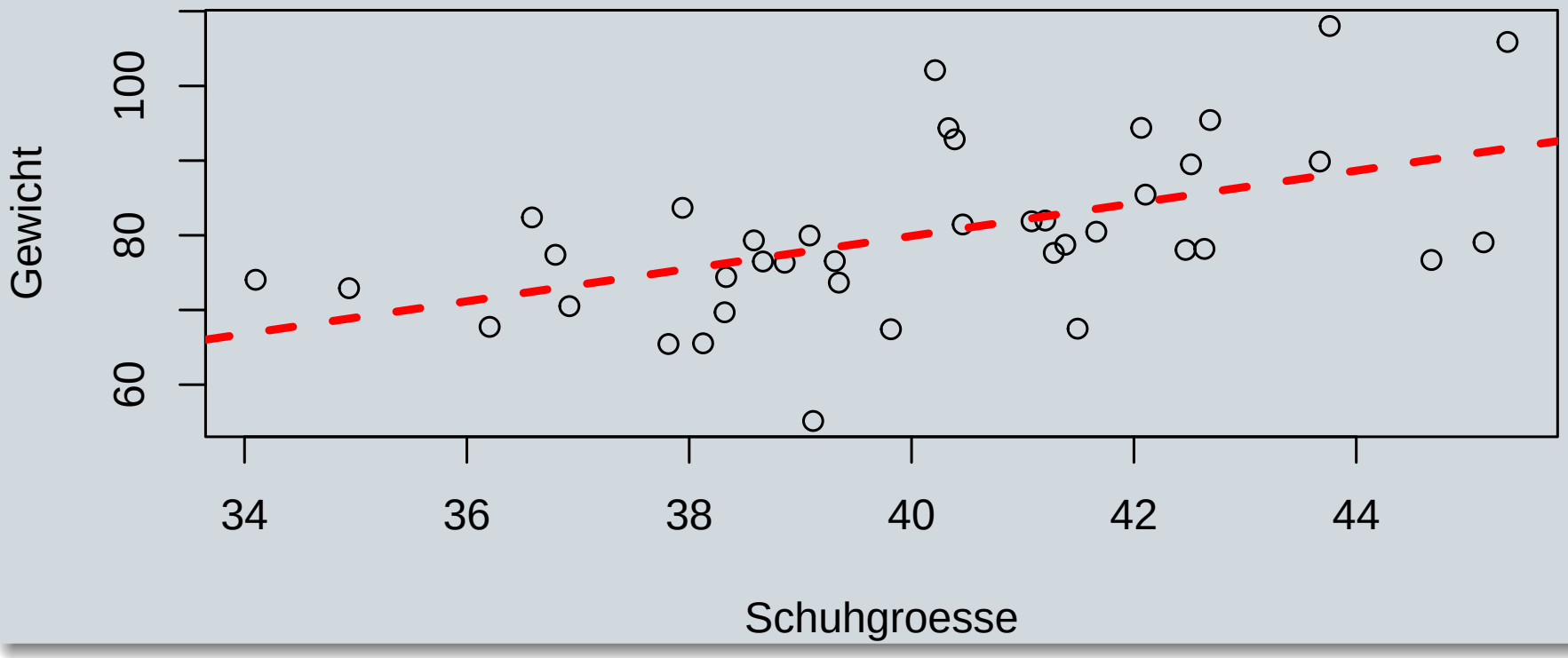
- Y: Zielvariable, abhängige Variable
- X: erklärende Variable, unabhängige Variable, Prädiktor, Feature

Beispiel 1

Zusammenhang zwischen Gewicht und Schuhgröße

- Prädiktor: Schuhgröße
- Zielvariable: Gewicht

wie hängt das Gewicht von der Schuhgröße ab?

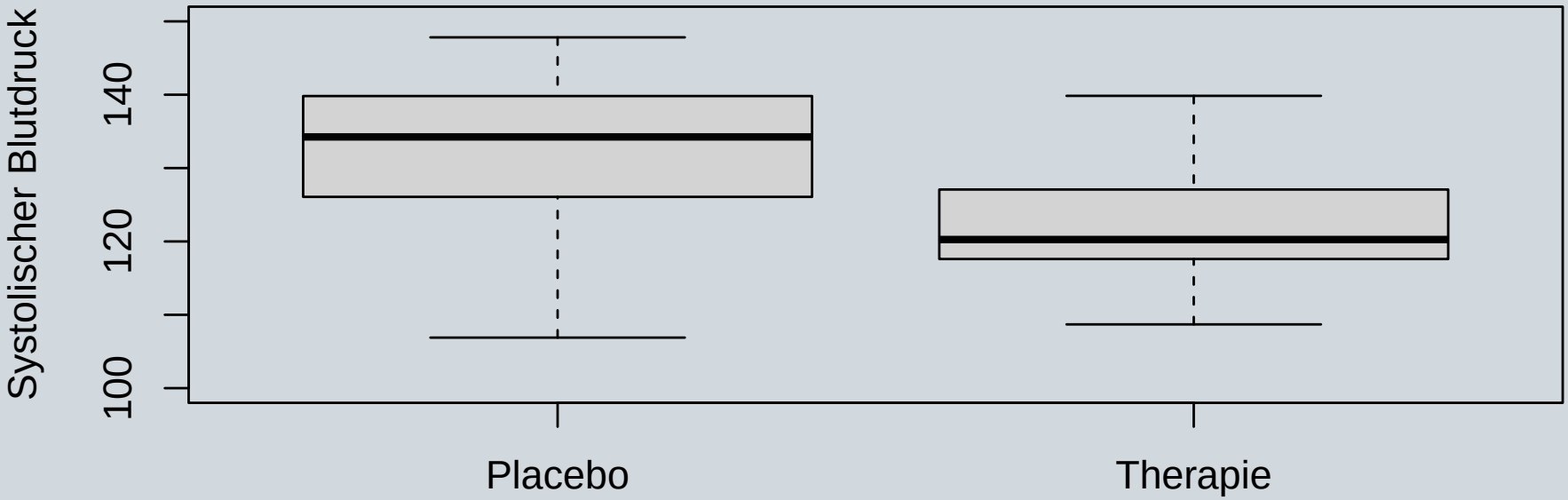


Beispiel 2

Wirkt Medikament besser als Placebo?

- Prädiktor: Gruppenzugehörigkeit - Therapiegruppe oder Placebogruppe
- Zielvariable: Blutdruck

Unterscheidet sich der Blutdruck zwischen den beiden Gruppen?



Beispiel 3

Risiko an Diabetes zu erkranken

Welche Größen haben Einfluss?

- Alter
- BMI
- Taillenumfang
- Körperliche Aktivität
- Ernährung
- Bluthochdruck
- Blutzuckerwerte
- Diabetes in der Familie
- ...

Zielvariable = Wahrscheinlichkeit

erklärende Variablen = Alter, BMI etc.

Und warum das Ganze?

Ziele einer Regressionsanalyse

- **Zusammenhänge** zwischen X und Y
  - **verstehen** und
  - **quantifizieren**
- **Vorhersagen** machen

→ es kann nicht auf Kausalität geschlossen werden!

Inhaltliche Gliederung

Lineares Modell

- metrische Zielgröße
- umfasst verschiedene Spezialfälle:
  - Lineare Regression
  - t-Test
  - Varianzanalyse (ANOVA)

Logistische Regression

- binäre Zielgröße (Ja/Nein, Misserfolg/Erfolg etc.)
- „verallgemeinertes lineares Modell“

Wahlfach *Computer und Medizin*

- Vertiefung zur Regressionsanalyse
  - Anwendung mit SPSS
  - Überlebens-/Ereigniszeitanalyse (Cox-Regression) in R

weiterführende Informationen

Bücher

- Fox: An R Companion to Applied Regression
- Faraway: Linear models using R

Variablentypen

Einteilung von Messgrößen

- metrische Variablen
  - kontinuierliche Größen
  - z.B. Körpergröße, Alter, Temperatur
- kategoriale Variablen
  - Einteilung in Klassen
  - z.B. Augenfarbe, Bewertungen, Altersklassen
- binäre/dichotome Variablen
  - kategoriale Variablen mit zwei Ausprägungen
  - z.B. ja/nein, Erfolg/Misserfolg, krank/gesund

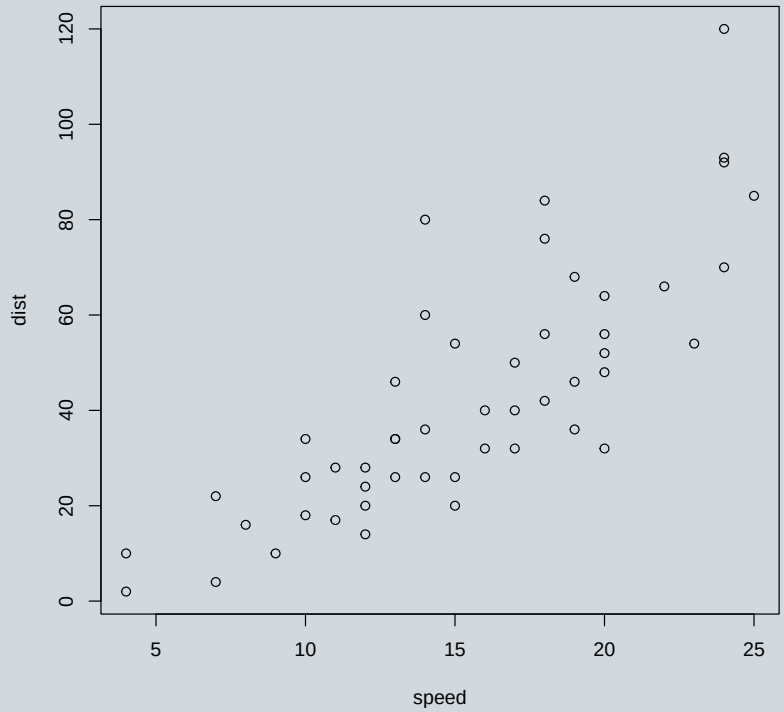
Grafische Darstellung von Zusammenhängen

	metrisch	kategorial
metrisch	- Streudiagramm $\hat{y}$: Zielgröße	- Boxplot - Beanplot - Dotplot (Punkt-diagramm)
kategorial	- Dotplot - konditionaler Dichteplot	- Balkendiagramm z.B. Mosaikplot

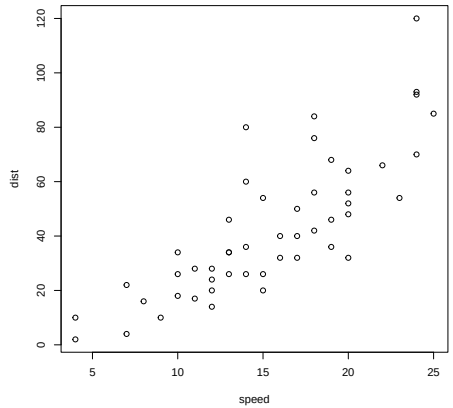
Grafische Darstellung von Zusammenhängen

metrisch vs metrisch: Streudiagramm / Scatterplot

```
plot(dist ~ speed, data=cars)
```



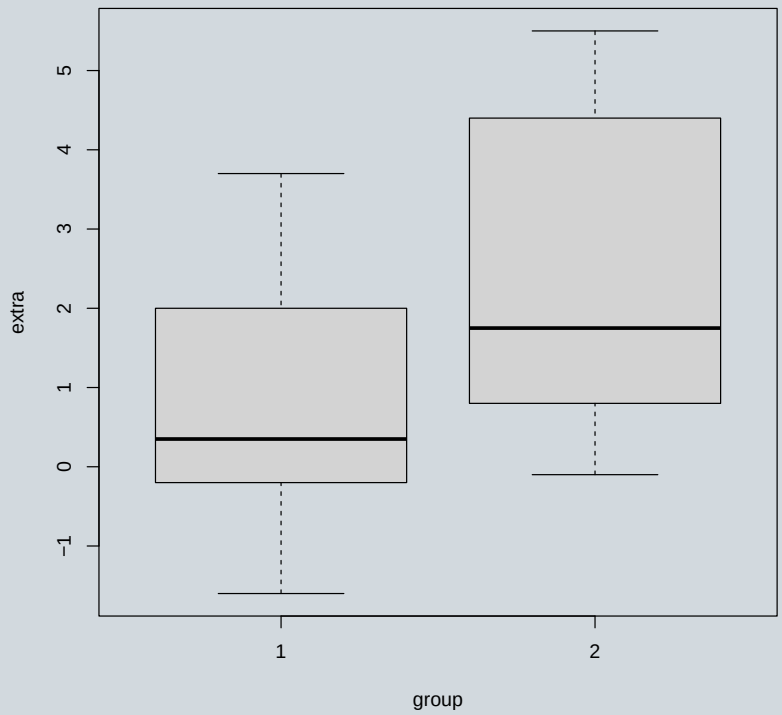
Grafische Darstellung von Zusammenhängen

	metrisch	kategorial
metrisch		
kategorial		

Grafische Darstellung von Zusammenhängen

metrisch vs kategorial: Boxplot

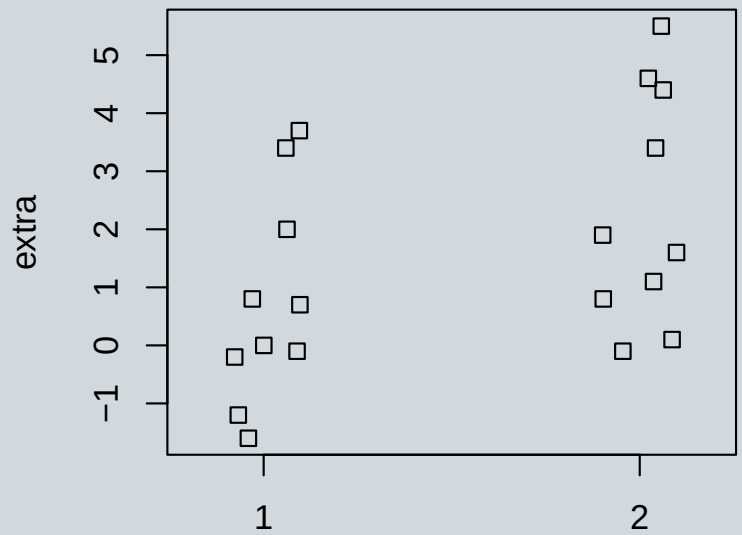
```
boxplot(extra ~ group, data=sleep)
```



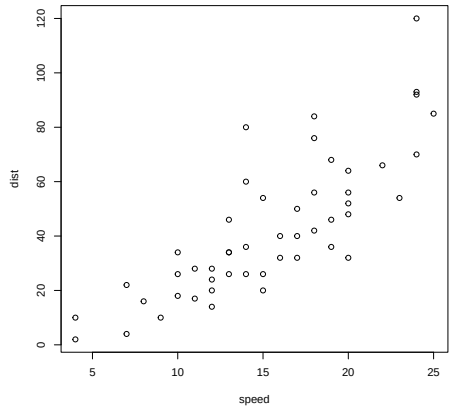
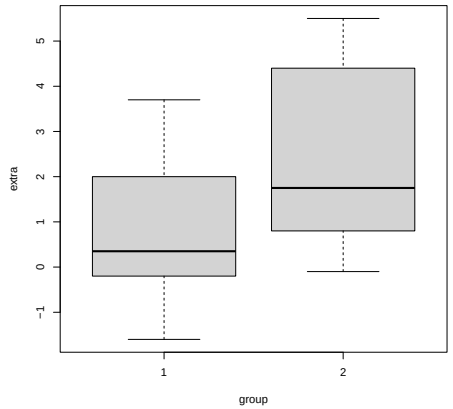
Grafische Darstellung von Zusammenhängen

metrisch vs kategorial: Dotplot

```
stripchart(extra ~ group, data=sleep, vertical=TRUE,
           method="jitter")
```



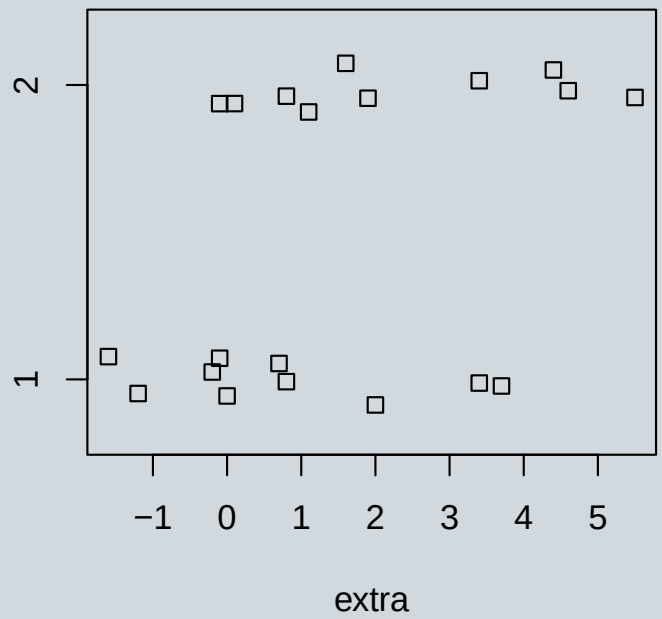
Grafische Darstellung von Zusammenhängen

	metrisch	kategorial
metrisch		
kategorial		

Grafische Darstellung von Zusammenhängen

kategorial vs metrisch: Dotplot

```
stripchart(extra ~ group, data=sleep, vertical=FALSE,
           method="jitter")
```

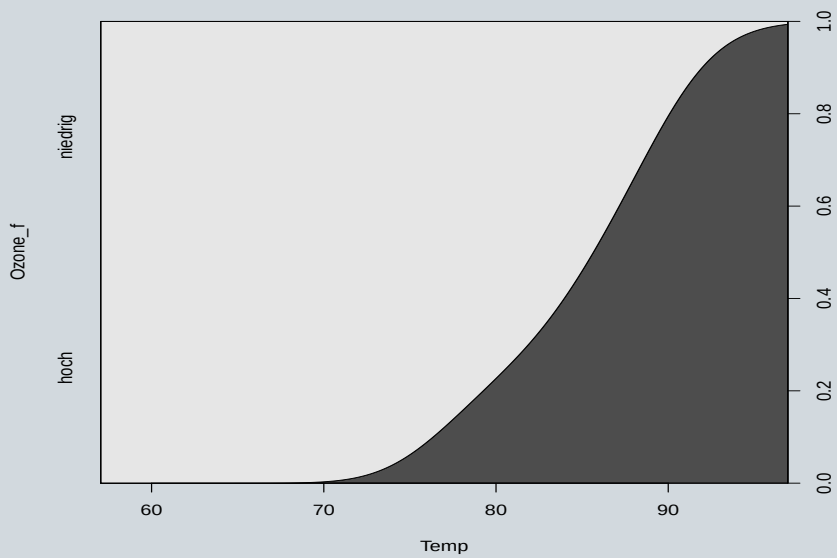


Grafische Darstellung von Zusammenhängen

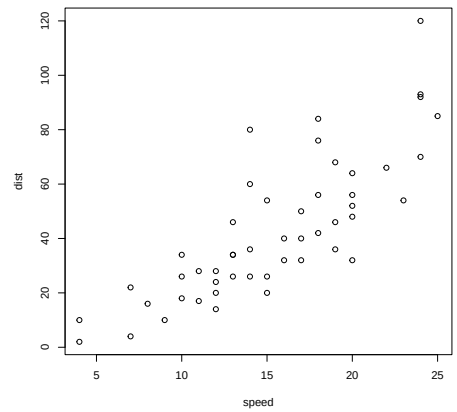
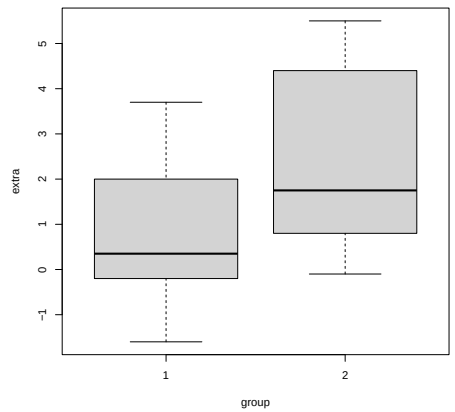
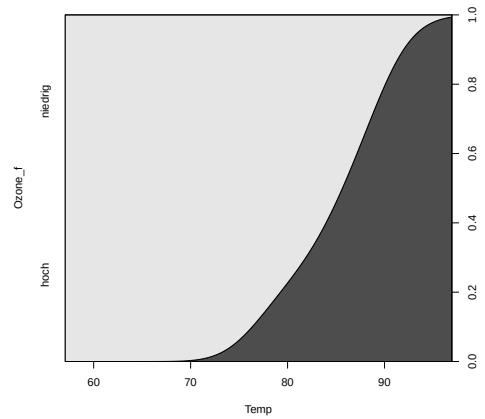
kategorial vs metrisch: konditionaler Dichteplot

```
cdplot(Ozone_f ~ Temp, data=airquality)
```

Darstellung zweier Gruppen in Abh. zur Einflussgröße



Grafische Darstellung von Zusammenhängen

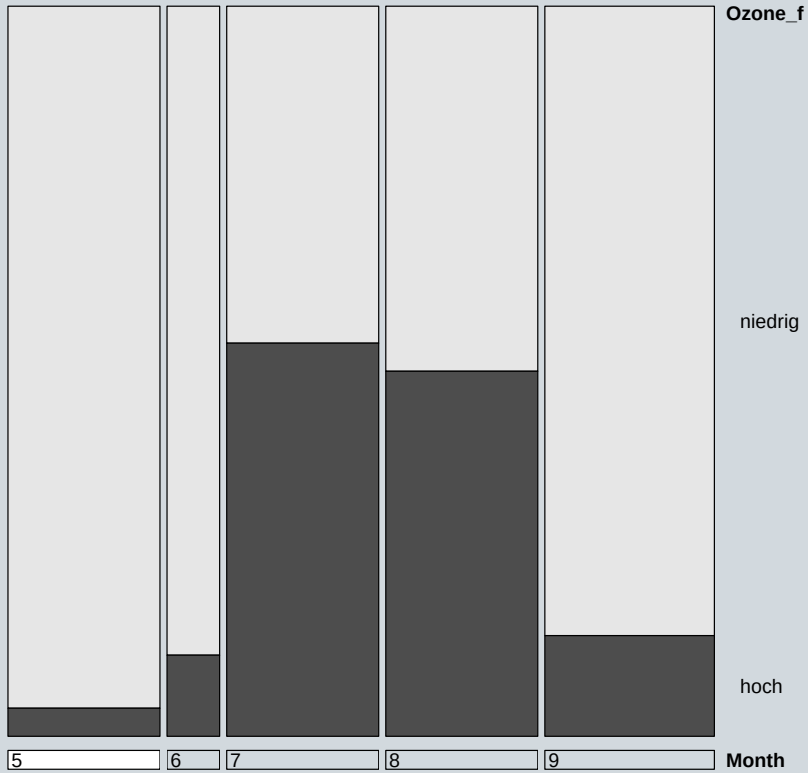
	metrisch	kategorial
metrisch		
kategorial		

Grafische Darstellung von Zusammenhängen

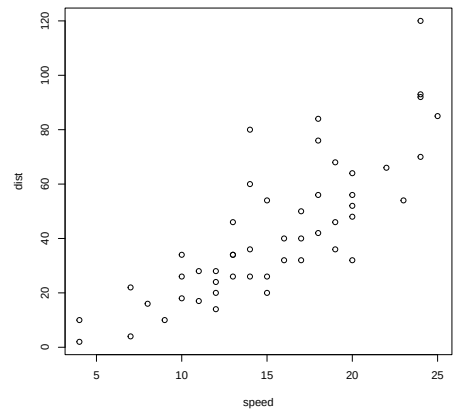
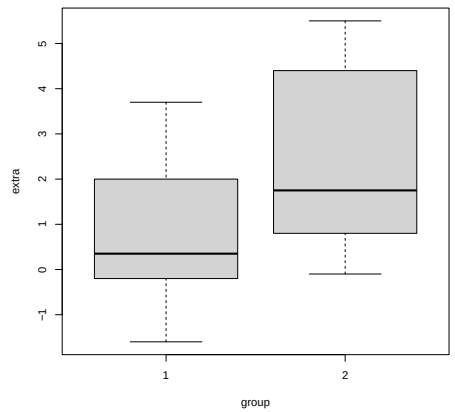
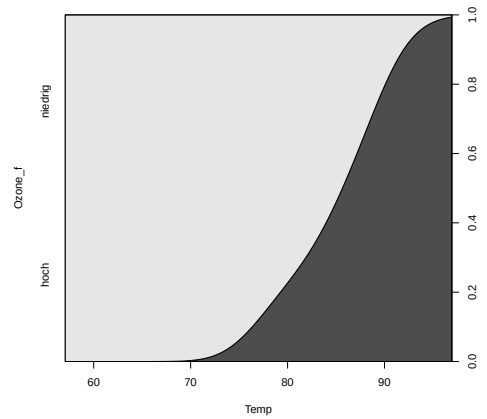
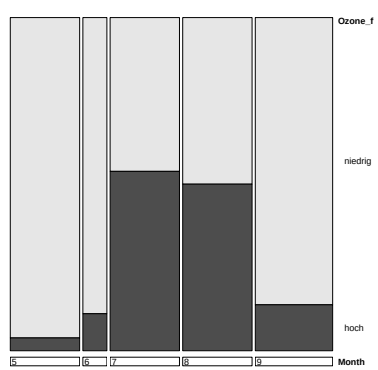
kategorial vs kategorial: Mosaik-Plot

```
library("vcd")
doubledecker(Ozone_f ~ Month, data=airquality)
```

Breite der Balken ≙ Anzahl der Werte



Grafische Darstellung von Zusammenhängen

	metrisch	kategorial
metrisch		
kategorial		

Regressionsmodelle

Zielgröße

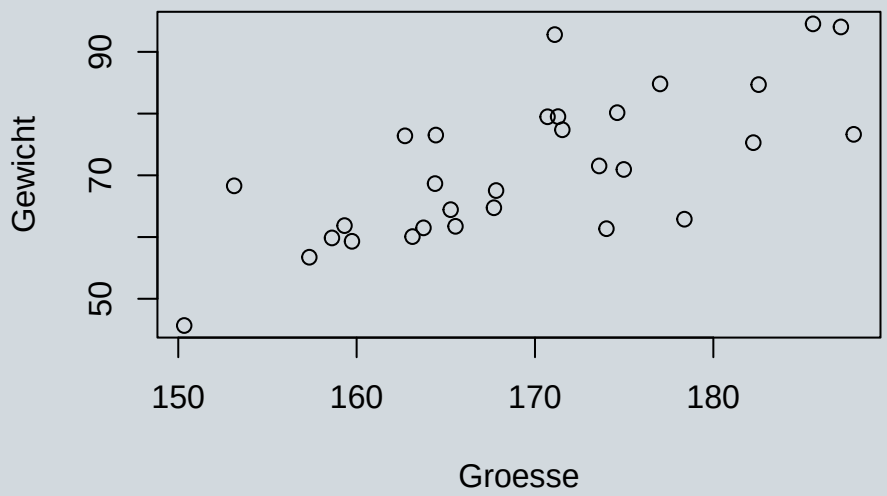
- metrische Zielgröße: lineare Regression
- binäre Zielgröße: logistische Regression

Einflussgrößen

- Einflussgrößen können sowohl metrisch als auch kategorial sein
- Anzahl:
  - nur eine Einflussgröße: einfache Regression
  - mehrere Einflussgrößen: multiple Regression

Einfache lineare Regression

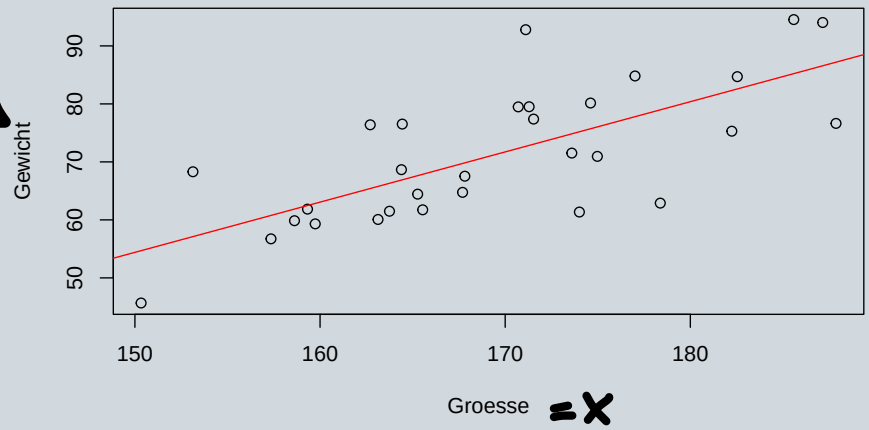
Zusammenhang zweier metrischer Größen



Einfache lineare Regression

Regressionsgerade

$(y = a \cdot x + b)$

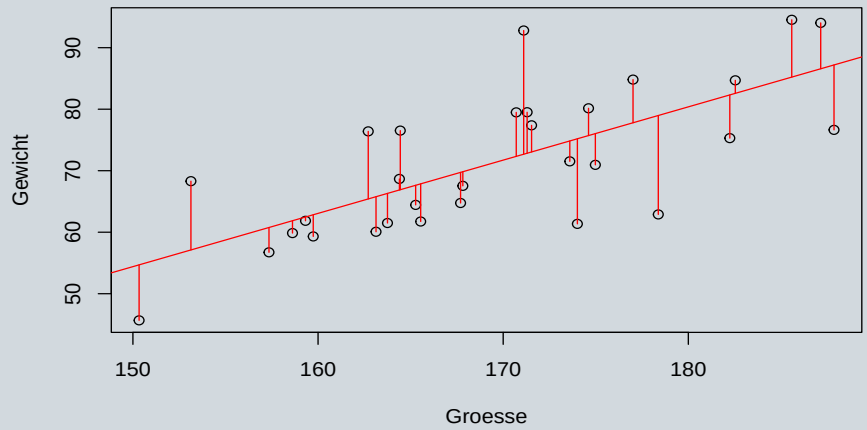


$$Gewicht = \beta_0 + \beta_1 \cdot Groesse$$

- β_0 Schnittpunkt mit y-Achse (*intercept*) (b)
- β_1 Anstieg (wenn Größe um 1 wächst) (*slope*) (a)

Einfache lineare Regression

Lineares Modell



Nicht durch Gerade erklärte Varianz

$$Gewicht_i = \beta_0 + \beta_1 \cdot Groesse_i + \epsilon_i$$

Residuum ϵ_i beschreibt Abweichung von der Geraden mit $E[\epsilon_i] = 0$

Statistische Modelle

Bestandteile

Eine Regression basiert auf einem statistischen Modell, das aus zwei Teilen besteht:

- einem systematischen Teil
- einem zufälligen Teil (Residuen)

Lineares Modell

$$Gewicht_i = \beta_0 + \beta_1 \cdot Groesse_i + \epsilon_i$$

Lineares Modell

Parameter gehen linear in systematischen Teil ein!

Beispiele für lineare Modelle

- $y = \beta_0 + \beta_1 \cdot x + \epsilon$
- $y = \beta_0 + \beta_1 \cdot x^2 + \epsilon$
- $y = \beta_0 + \beta_1 \cdot \log(x) + \epsilon$

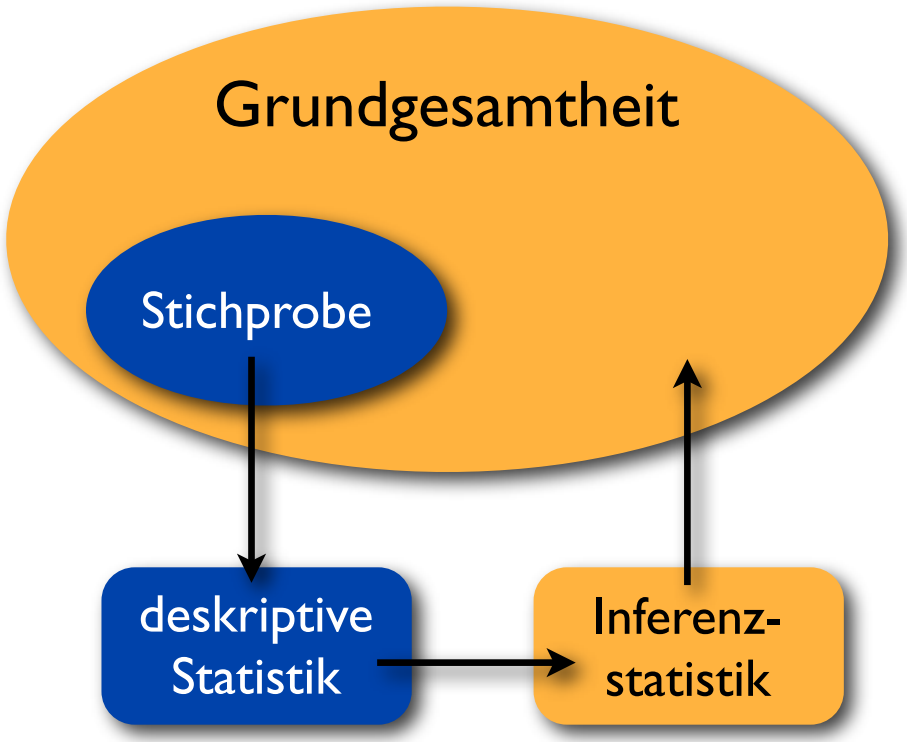
Voraussetzung: Parameter müssen linear sein!

da die Parameter β_0 und β_1 linear eingehen

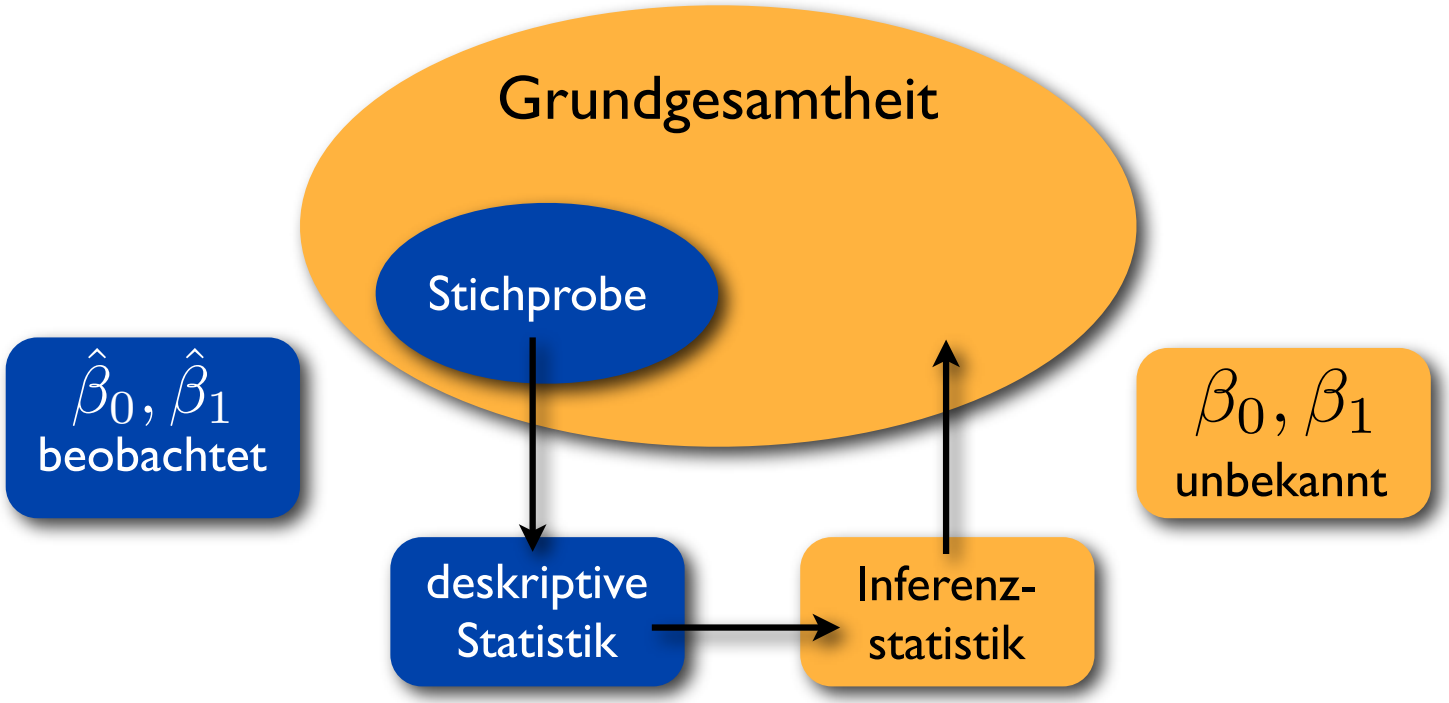
Beispiele für **nicht**-lineare Modelle

- $y = \frac{\beta_1 \cdot x}{\beta_2 + x} + \epsilon$ (Michaelis-Menten Modell)
- $y = \beta_0 + \beta_1 \cdot x^{\beta_2} + \epsilon$

Parameterschätzung



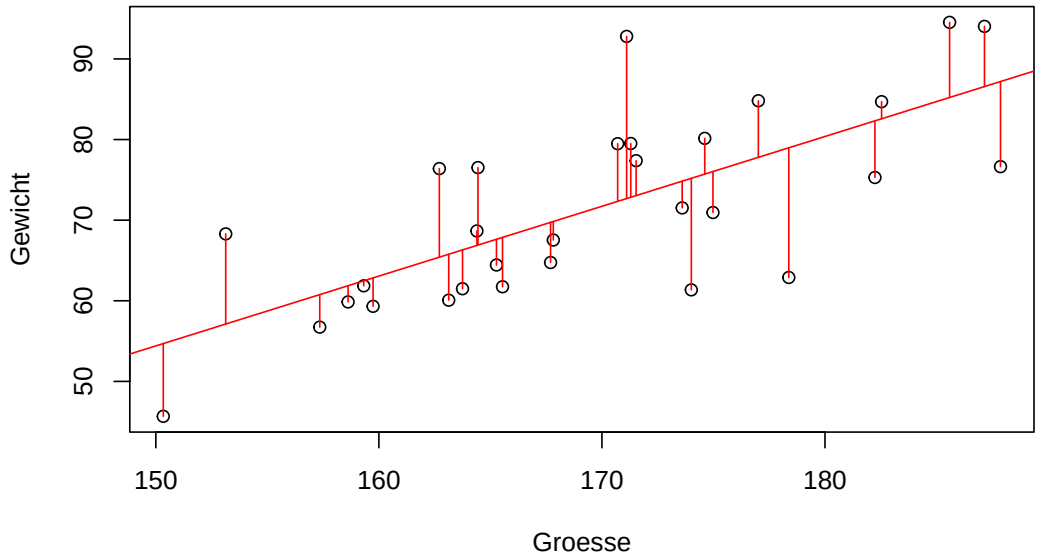
Parameterschätzung



Terminologie

- β_0 und β_1 sind nicht bekannt
- $\hat{\beta}_0$ und $\hat{\beta}_1$ werden aus Daten geschätzt

Parameterschätzung



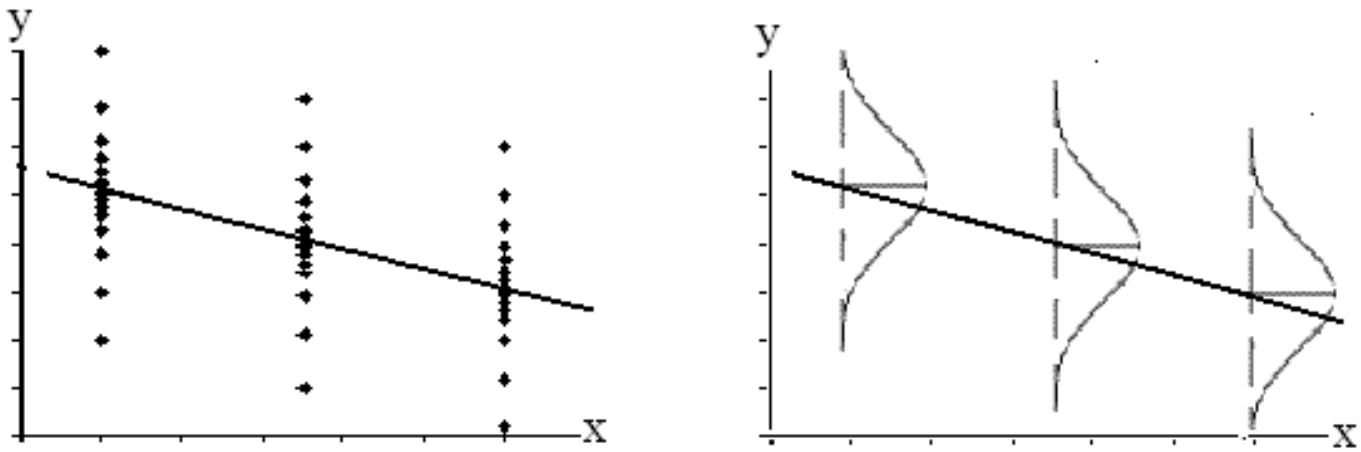
Quadrate um Abweichungen zuordnen

Methode der kleinsten Quadrate

Wähle $\hat{\beta}_0$ und $\hat{\beta}_1$ so, dass die Summe der quadratischen Abweichungen minimal ist:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 \cdot x_i) \Big)^2$$

Annahmen zu Residuen



$$y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Verteilung der Residuen

- Normalverteilung
- konstante Varianz
- Unabhängigkeit

} Voraussetzungen

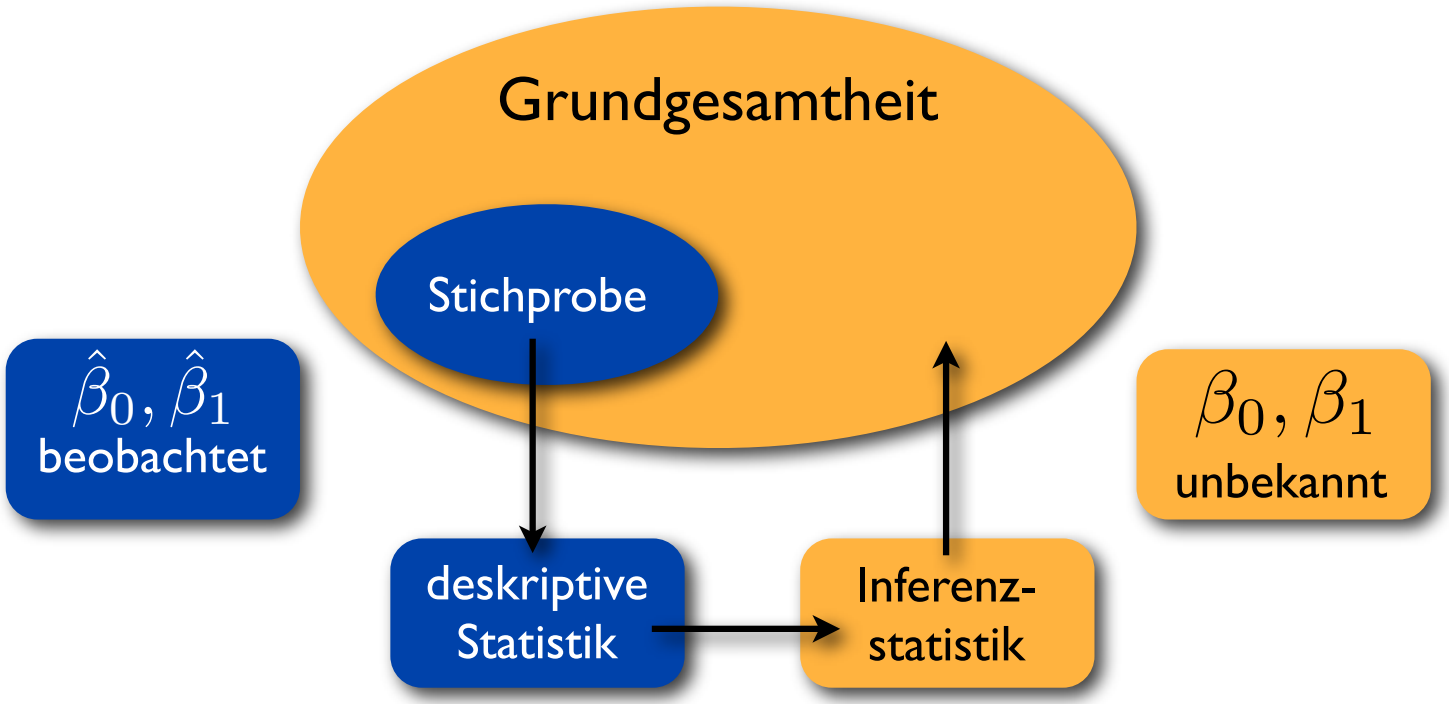
Signifikanztests für Parameter

H0 = Es gibt keinen Anstieg $\beta_1 = 0$ (kein Zusammenhang)
H1 = Es gibt einen Anstieg $\beta_1 \neq 0$ (Zusammenhang)

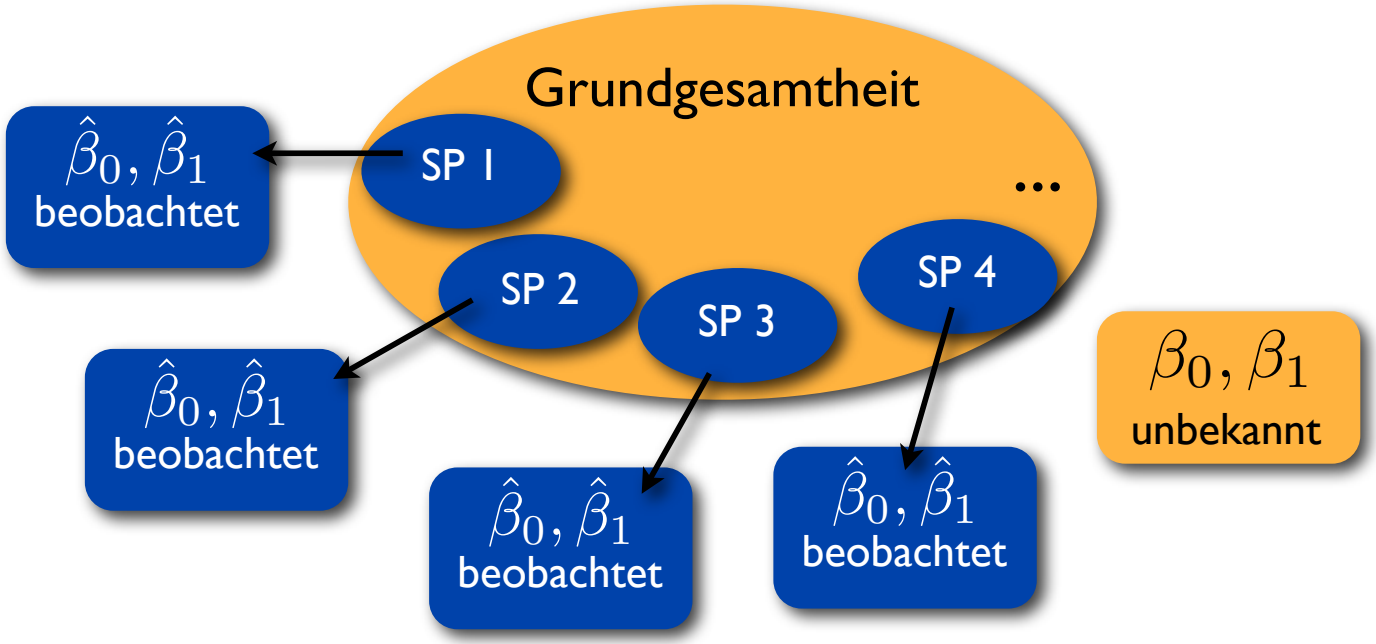
Hat X einen linearen Einfluss auf Y?

- $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$
- Test-Statistik $\frac{\hat{\beta}_1}{se_{\hat{\beta}_1}} \overset{H_0}{\sim} t_{n-p}$
- Wald Signifikanz-Test

Konfidenzintervalle



Konfidenzintervalle



Konfidenzintervall (KI)

Mittels jeder Stichprobe lässt sich ein 95%-KI für β_i berechnen. Im Mittel überdecken 95% der KI den wahren Wert β_i (wenn Modellannahmen stimmen).

Umsetzung mit R

Daten

```
head(lmExample)

##   Groesse Gewicht
## 1   164.4   68.66
## 2   167.7   64.75
## 3   185.6   94.54
## 4   170.7   79.49
## 5   171.3   79.51
## 6   187.2   94.04
```

Umsetzung mit R

lm-Funktion

```
(model1 <- lm(Gewicht ~ Groesse, data = lmExample))

##
## Call:
## lm(formula = Gewicht ~ Groesse, data = lmExample)
##
## Coefficients:
## (Intercept)      Groesse
##     -75.494        0.866
```

abhängige Variable — in Abhängigkeit von

β_1 → Anstieg

β_0 → Schnittpunkt mit y-Achse

Wenn Größe um 1cm erhöht wird, erhöht sich das Gewicht um 0,866 Kg im Mittel

Übersicht der wichtigsten Informationen

Summary-Befehl

```
summary(model1)


##
## Call:
## lm(formula = Gewicht ~ Groesse, data = lmExample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.09  -5.06  -2.15   6.93  20.12
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -75.494     26.977   -2.80   0.0092 **
## Groesse        0.866      0.159    5.45  8.1e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.39 on 28 degrees of freedom
## Multiple R-squared:  0.515,Adjusted R-squared:  0.497
## F-statistic: 29.7 on 1 and 28 DF,  p-value: 8.1e-06
```

(p-wert)

Extraktion einzelner Größen

Koeffizienten (Parameter)

```
coef(model1)
```

```
## (Intercept)     Groesse
##      -75.494       0.866
```

angepasste Werte

```
fitted(model1)
```

(Werte, die genau auf der Geraden liegen)

```
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## 66.87 69.73 85.22 72.33 72.84 86.57 75.71 60.77 65.77 67.86 82.32 74.84 75.19
##    14    15    16    17    18    19    20    21    22    23    24    25    26
## 72.68 66.91 87.19 76.03 54.69 77.79 67.63 62.47 69.83 62.84 65.41 66.31 57.11
##    27    28    29    30
## 78.98 73.05 61.86 82.58
```

Extraktion einzelner Größen

Residuen (Abstände zwischen beobachteten und angepassten Werten)

```
residuals(model1)
```

```
##       1       2       3       4       5       6       7       8
##  1.7932 -4.9795  9.3202  7.1556  6.6689  7.4649  4.4367 -4.0350
##       9      10      11      12      13      14      15      16
## -5.7005 -6.1223 -7.0266 -3.3171 -13.8370 20.1177  9.6144 -10.5563
##      17      18      19      20      21      22      23      24
## -5.0818 -9.0227  7.0195 -3.1876 -0.6183 -2.2979 -3.5242 10.9888
##      25      26      27      28      29      30
## -4.8157 11.1837 -16.0849  4.3315 -2.0072  2.1197
```

Summe quadratischer Abstände (RSS, deviance)

```
deviance(model1)
```

```
## [1] 1972
```

Konfidenzintervalle

mit R

```
confint(model1)
```

(Standard: 95% KI)

```
##                 2.5 %  97.5 %
## (Intercept) -130.7528 -20.235
## Groesse        0.5405   1.191
```

signifikant

KI [0,5405; 1.191]

Konfidenzintervalle

Die Breite des Konfidenzintervalls hängt ab von:

- der Varianz der Daten
- der Stichprobengröße

Regressionsmodelle

Teil 1: Einführung in die Lineare Regression

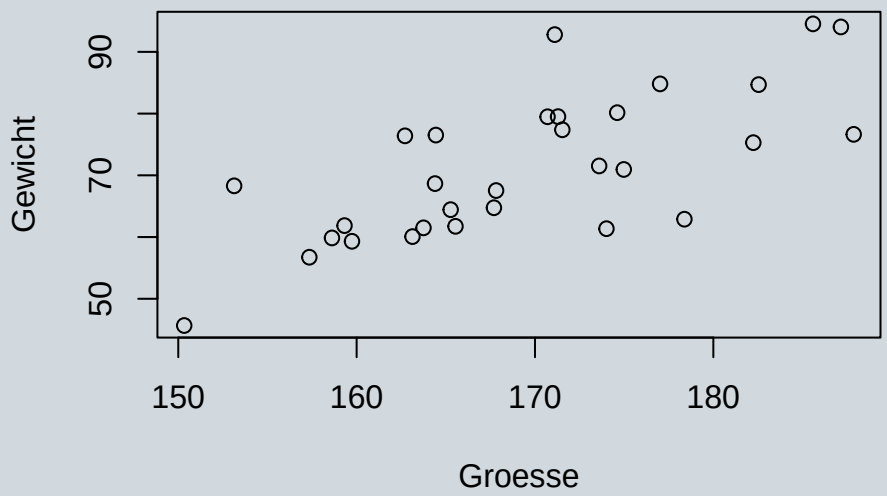
Thomas Zerjatke
thomas.zerjatke@tu-dresden.de

TU Dresden
Institut für Medizinische Informatik und Biometrie

WS 2023/24

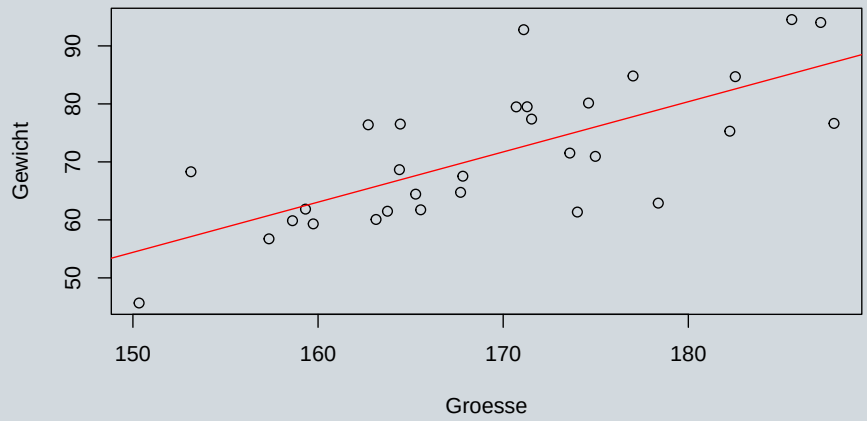
Einfache lineare Regression

Zusammenhang zweier metrischer Größen



Einfache lineare Regression

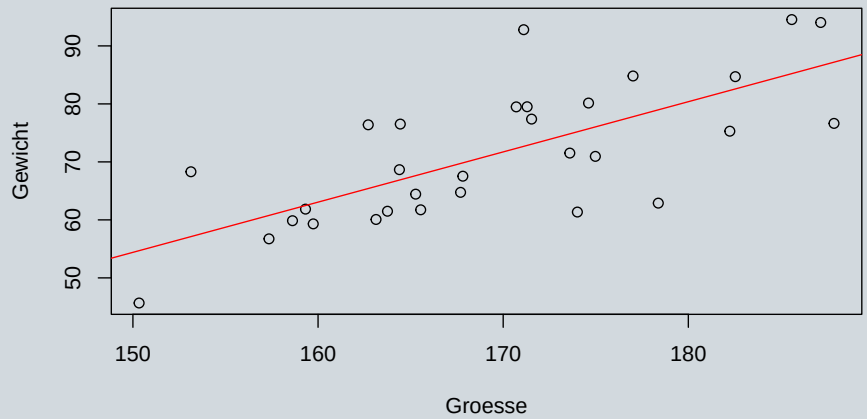
Regressionsgerade



$$Gewicht = \beta_0 + \beta_1 \cdot Groesse$$

Einfache lineare Regression

Regressionsgerade



$$Gewicht = \beta_0 + \beta_1 \cdot Groesse$$

- β_0 Schnittpunkt mit y-Achse (*intercept*)
- β_1 Anstieg (wenn Größe um 1 wächst) (*slope*)

Vorhersage

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i,$$

$$\hat{\beta}_0 = -75.5, \hat{\beta}_1 = 0.87$$

Welches Gewicht wird man im Mittel für eine Größe von 175 cm erwarten?

Vorhersage

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i,$$

$$\hat{\beta}_0 = -75.5, \hat{\beta}_1 = 0.87$$

Welches Gewicht wird man im Mittel für eine Größe von 175 cm erwarten?

$$\hat{y} = -75.5 + 0.87 \cdot 175cm = 76.1kg$$

Vorhersage

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i,$$

$$\hat{\beta}_0 = -75.5, \hat{\beta}_1 = 0.87$$

Welches Gewicht wird man im Mittel für eine Größe von 80 cm erwarten?

Vorhersage

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i,$$

$$\hat{\beta}_0 = -75.5, \hat{\beta}_1 = 0.87$$

Welches Gewicht wird man im Mittel für eine Größe von 80 cm erwarten?

$$\hat{y} = -75.5 + 0.87 \cdot 80cm = -6.2kg$$

Vorhersage

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i,$$

$$\hat{\beta}_0 = -75.5, \hat{\beta}_1 = 0.87$$

Welches Gewicht wird man im Mittel für eine Größe von 80 cm erwarten?

$$\hat{y} = -75.5 + 0.87 \cdot 80cm = -6.2kg$$

Vorsicht bei Extrapolation

Vorhersagen für Werte außerhalb der Beobachtungen sind oft nicht sinnvoll und sollten deshalb nicht gemacht werden!

Vorhersage mit R

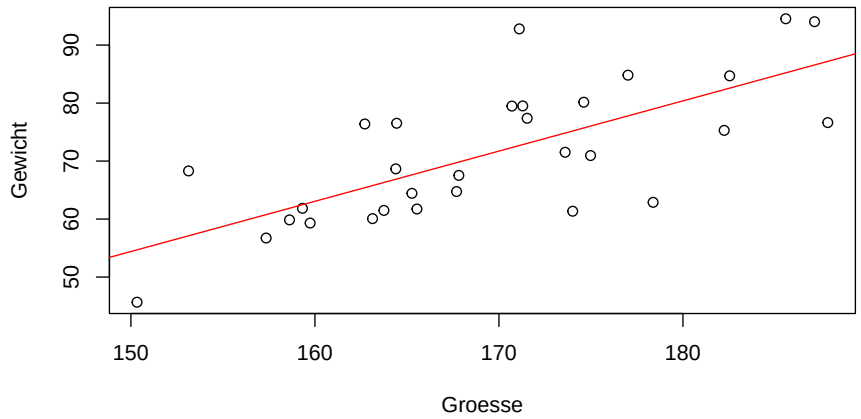
Funktion predict()

```
neueGroessen <- data.frame(
  Groesse = c(160,175,190))

predict(model1, neueGroessen)

##      1      2      3
## 63.06 76.05 89.04
```

Konfidenzintervalle für Vorhersagen

2 Fälle

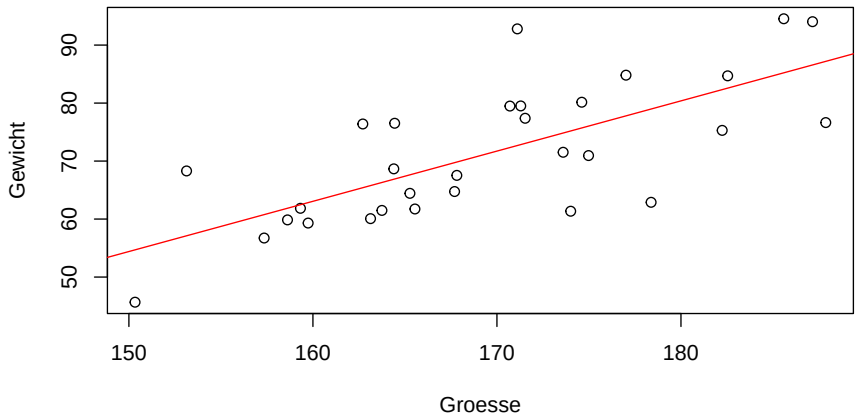
- Vorhersage für neuen Fall
- Vorhersage für Mittelwert von Fällen mit gleichem Prädiktorwert



Konfidenzintervalle für Vorhersagen

2 Fälle

- Vorhersage für neuen Fall
  - In welchem Bereich erwarten wir neue Fälle?
- Vorhersage für Mittelwert von Fällen mit gleichem Prädiktorwert
  - Wie genau ist die Schätzung des Mittelwertes?



Konfidenzintervalle für Vorhersagen

Vorhersage für neuen Fall

```
predict(model1, neueGroessen, interval = "prediction")
```

```
##      fit   lwr    upr
## 1 63.06 45.31  80.81
## 2 76.05 58.48  93.62
## 3 89.04 70.34 107.74
```

Vorhersage für Mittelwert

```
predict(model1, neueGroessen, interval = "confidence")
```

```
##      fit   lwr   upr
## 1 63.06 58.65 67.47
## 2 76.05 72.44 79.66
## 3 89.04 81.68 96.40
```

Varianzzerlegung

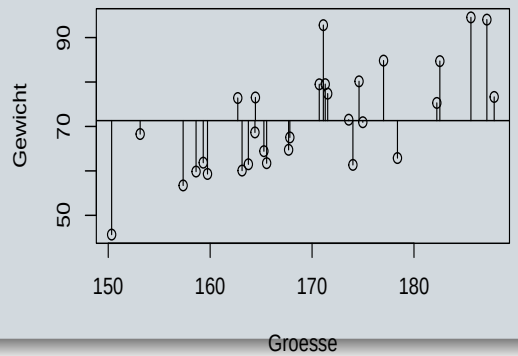
Teile der Varianz von Y kann durch die Einflussgröße X erklärt werden:

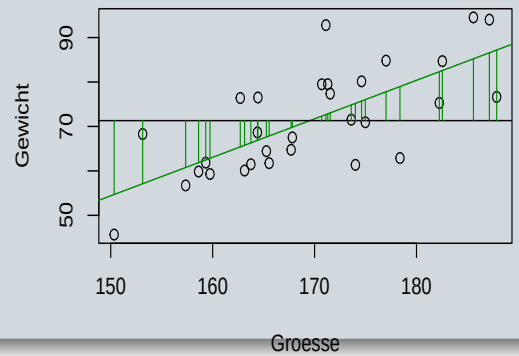
$$\text{Gesamtvarianz} = \text{erklärte Varianz} + \text{Residualvarianz}$$

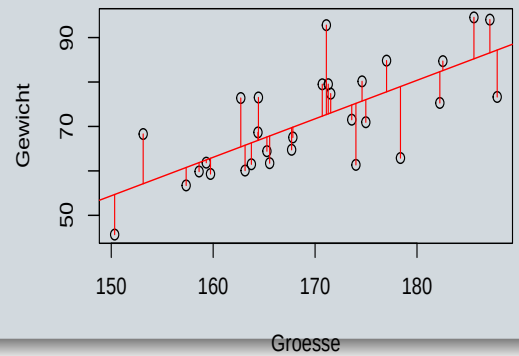
Varianzzerlegung

Teile der Varianz von Y kann durch die Einflussgröße X erklärt werden:

$$\text{Gesamtvarianz} = \text{erklärte Varianz} + \text{Residualvarianz}$$







Bestimmtheitsmaß B (oder R^2)

Anpassungsgüte des linearen Modells

Anteil der erklärten Varianz an der Gesamtvarianz:

$$B = R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{erklärte Varianz}}{\text{Gesamtvarianz}}$$

Bestimmtheitsmaß B (oder R^2)

Anpassungsgüte des linearen Modells

Anteil der erklärten Varianz an der Gesamtvarianz:

$$B = R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{erklärte Varianz}}{\text{Gesamtvarianz}}$$

Bei der einfachen linearen Regression (d.h. eine Einflussgröße) entspricht das Bestimmtheitsmaß B dem Quadrat des Pearsonschen Korrelationskoeffizienten r.

Modelldiagnostik

Die drei Säulen

- Annahmen über die Residuen
- Ungewöhnliche Beobachtungen
- Modellstruktur

Modelldiagnostik

Die drei Säulen

- Annahmen über die Residuen
- Ungewöhnliche Beobachtungen
- Modellstruktur

Sinnvolle Aussagen können nur dann getroffen werden, wenn die Adäquatheit des statistischen Modells überprüft wurde!

Modelldiagnostik

Annahmen über die Residuen:

- Normalverteilung
- konstante Varianz
- Unabhängigkeit

Ungewöhnliche Beobachtungen

Gibt es ungewöhnliche Beobachtungen, so sollten diese besonders auf Plausibilität überprüft werden.

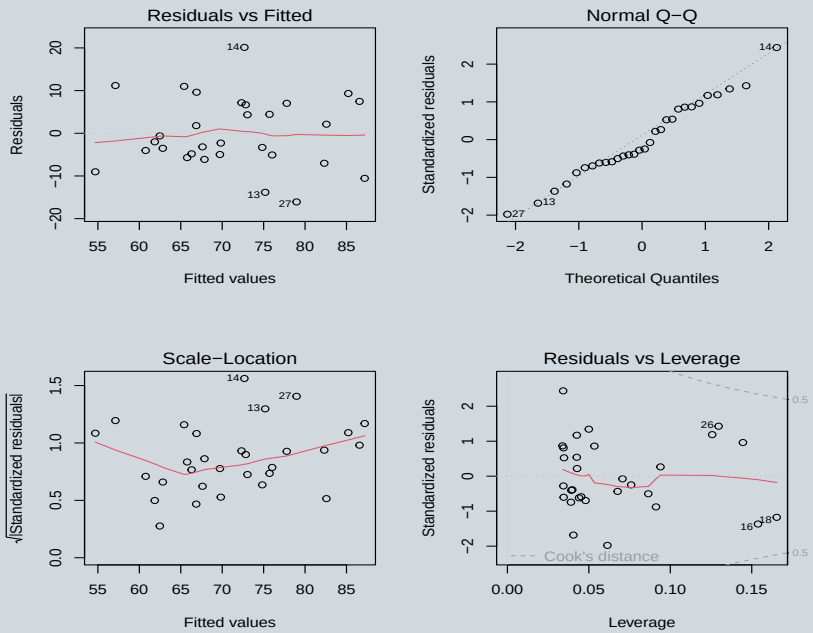
Modellstruktur

Ist die Annahme der Linearität gerechtfertigt?

Modelldiagnostik: Residuen

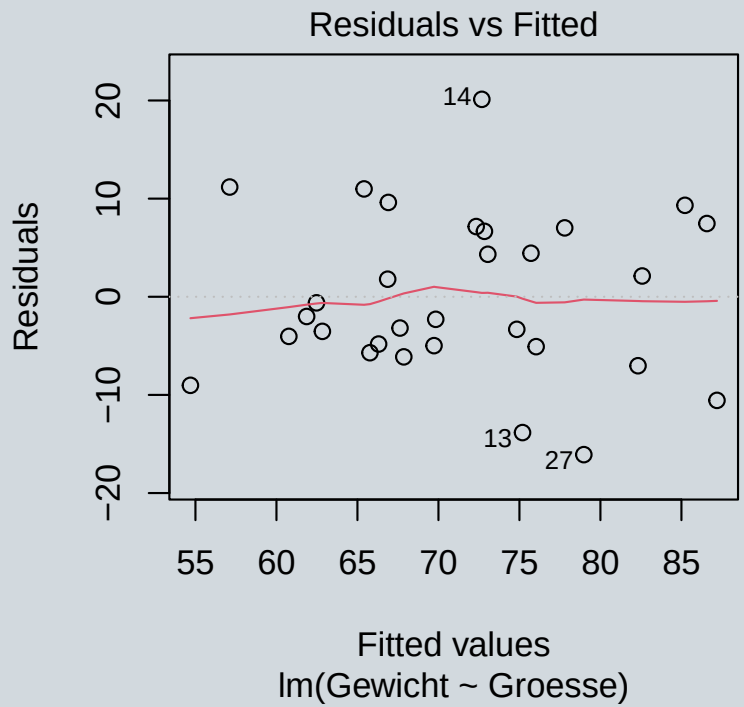
Diagnostische Residuen-Plots

```
par(mfrow = c(2,2))
plot(model1)
```



Residuen-Plot I

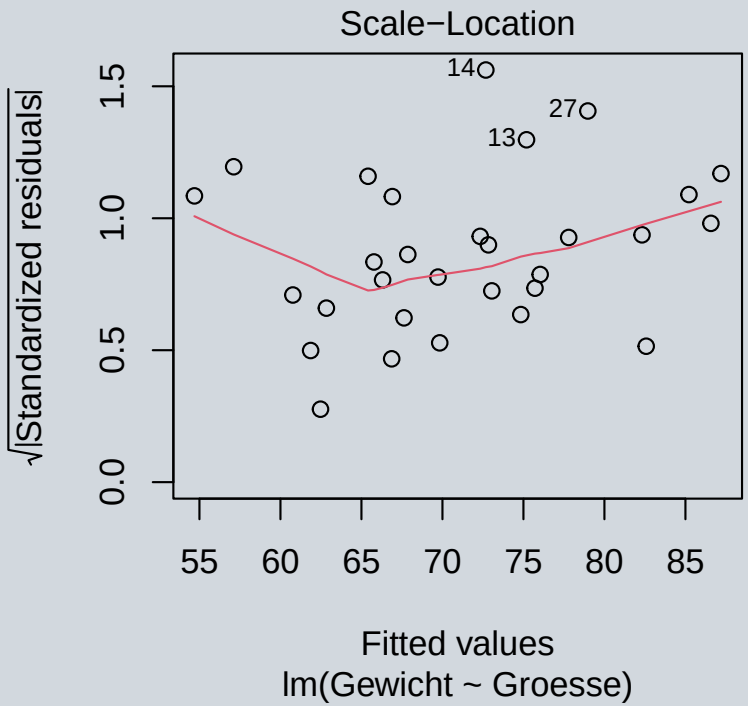
Residuen vs Angepasste Werte



Residuen sollten um die Nulllinie streuen,
anderenfalls könnte die Linearitätsannahme verletzt sein.

Residuen-Plot II

transformierte Residuen vs Angepasste Werte

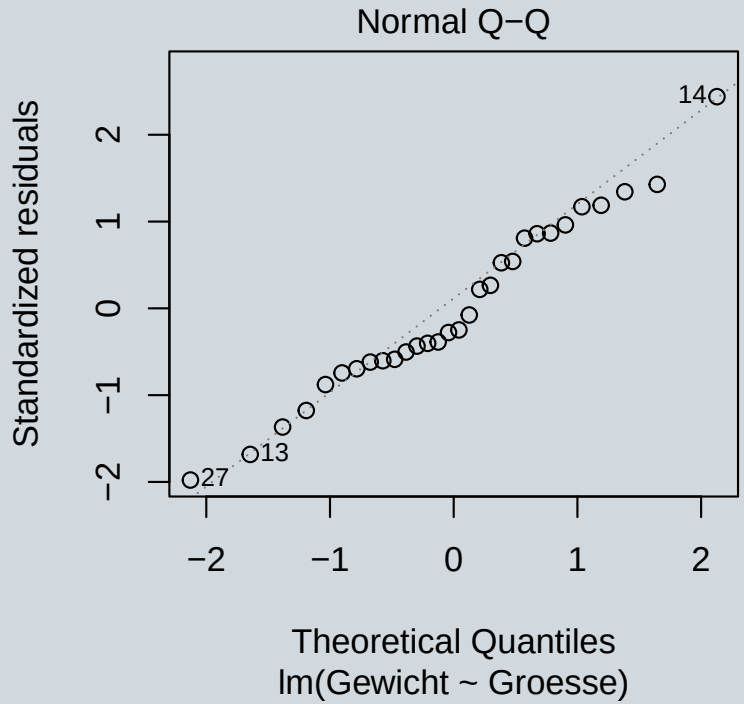


Überprüfung, ob Residuenvarianz konstant ist:
Rote Linie sollte näherungsweise bei 0.8 verlaufen.

immer bei 0.8, wegen Standardisierung

Residuen-Plot III

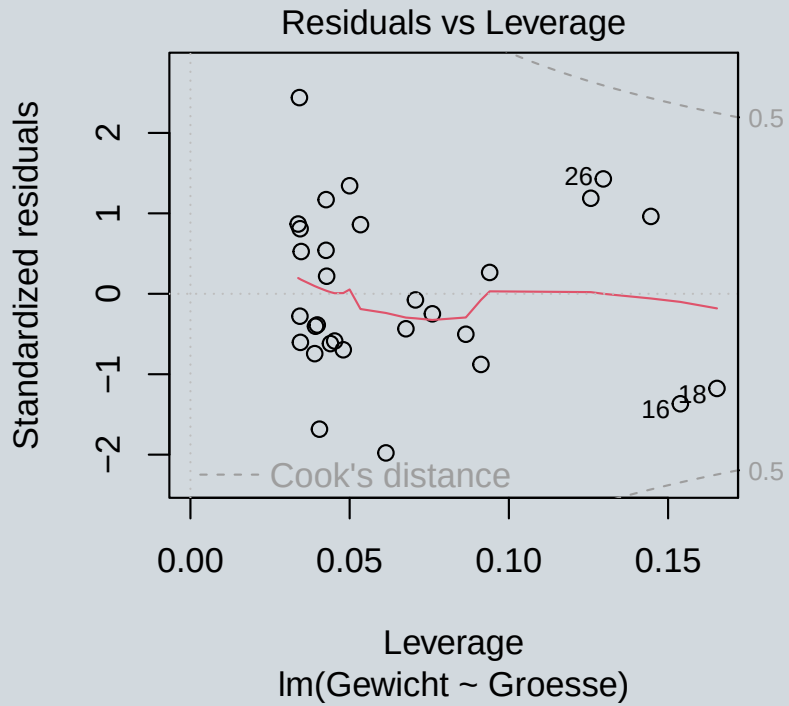
Q-Q-Plot: Normalverteilungsannahme



Bei Normalverteilung der Residuen bilden Punkte eine Gerade.

Ungewöhnliche Beobachtungen

Ausreißer und einflussreiche Beobachtungen

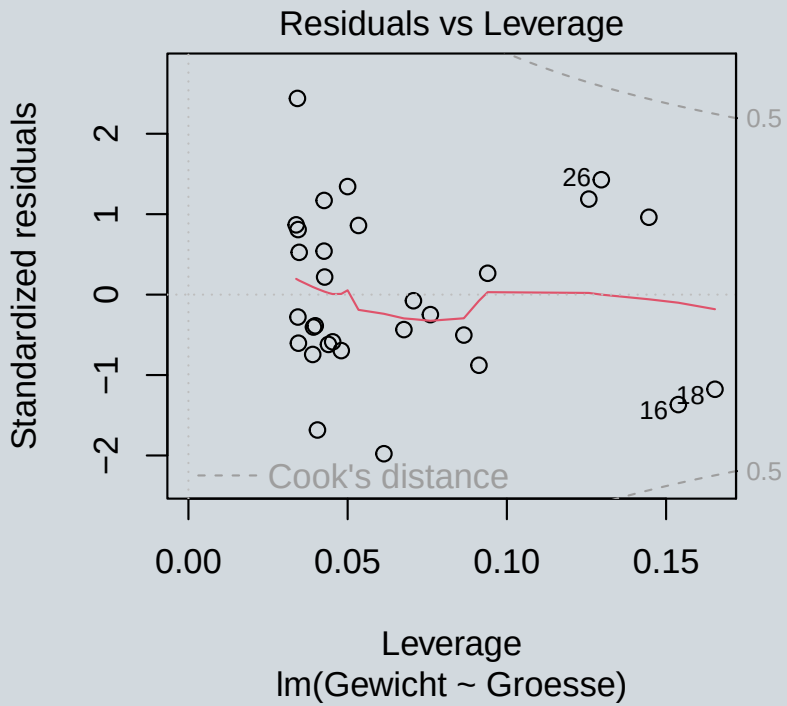


Leverage = Einfluss eines Datenpunkts auf den Fit

Cook's distance = Maß für Ausreißer *und* Leverage

Ungewöhnliche Beobachtungen

Ausreißer und einflussreiche Beobachtungen



Ausreißer (d.h. große Residuen) und Punkte mit großer Leverage oder Cook-Distanz sollten besonders auf Plausibilität überprüft werden.

Regressionsmodelle

Teil 1: Einführung in die Lineare Regression

III. Multiple Lineare Regression

Thomas Zerjatke
thomas.zerjatke@tu-dresden.de

TU Dresden
Institut für Medizinische Informatik und Biometrie

WS 2023/24

Erweiterungen des einfachen linearen Modells

Lineare Modelle

- einfache lineare Regression: ein metrischer Prädiktor
- multiple lineare Regression: mehrere metrische Prädiktoren

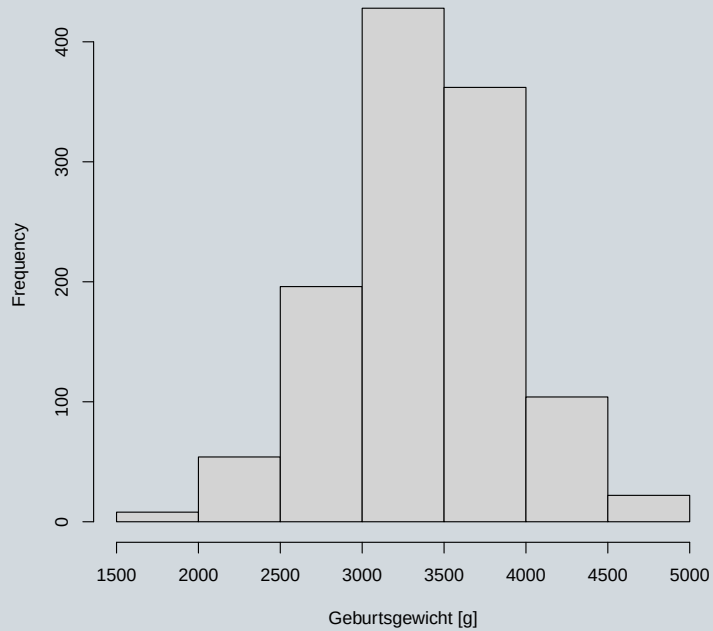
- Varianzanalyse: kategoriale Prädiktoren
- Kovarianzanalyse: metrische und kategoriale Prädiktoren

Umsetzung in R

Alle Formen des linearen Modells können einheitlich mithilfe der `lm`-Funktion umgesetzt werden.

Multiple lineare Regression

Beispiel: Geburtsgewicht bei Säuglingen



Zielgröße: Geburtsgewicht

Prädiktoren: Alter, Gewicht und Größe der Mutter

Modell

$$Geburtsgewicht = \beta_0 + \beta_1 \cdot Alter + \beta_2 \cdot Gewicht + \beta_3 \cdot Groesse + \epsilon$$

Multiple lineare Regression

Modell

$$Geburtsgewicht = \beta_0 + \beta_1 \cdot Alter + \beta_2 \cdot Gewicht + \beta_3 \cdot Groesse + \epsilon$$

Modellanpassung

Analog zur einfachen Regression:

Schätzung der Parameter über Methode der kleinsten Quadrate

Umsetzung in R

Prädiktoren werden mit + in der Formel verknüpft:

```
lm(Geburtsgewicht ~ Alter + Gewicht + Groesse, data=babies)
```

Multiple lineare Regression

summary

```
##
## Call:
## lm(formula = bwt ~ age + weight + height, data = babies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1894.9  -296.8    19.4   313.9  1587.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  863.048    394.153    2.19    0.029 *
## age            1.455      2.585    0.56    0.574
## weight         2.004      0.806    2.49    0.013 *
## height        34.764      6.540    5.32  1.3e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 508 on 1170 degrees of freedom
## Multiple R-squared:  0.0473,Adjusted R-squared:  0.0449
## F-statistic: 19.4 on 3 and 1170 DF,  p-value: 2.89e-12
```

Intercept = Schnittpunkt

Gewicht & Größe haben signifikanten Einfluss

→ Wenn Größe um 1 Einheit steigt, steigt die Größe des Babys um 34,764 Einheiten

Multiple lineare Regression

Interpretation der Koeffizienten

Koeffizienten β_0, β_1, β_2, usw. dürfen nicht isoliert betrachtet werden, sondern sind nur im Kontext des konkreten angepassten Modells zu interpretieren.

→ Korrelation ist nicht zu sehen!

Je nachdem, welche weiteren Prädiktoren betrachtet werden, ändert sich der Koeffizient eines Prädiktors.

Es sollte deshalb immer klar angegeben werden, welche weiteren Parameter ins Modell aufgenommen wurden.

Beispiel: Bei einem Gewichtsunterschied der Mutter von einer Einheit bei **gleichem** Alter und **gleicher** Größe ändert sich das Geburtsgewicht des Kindes um β_2 Einheiten.

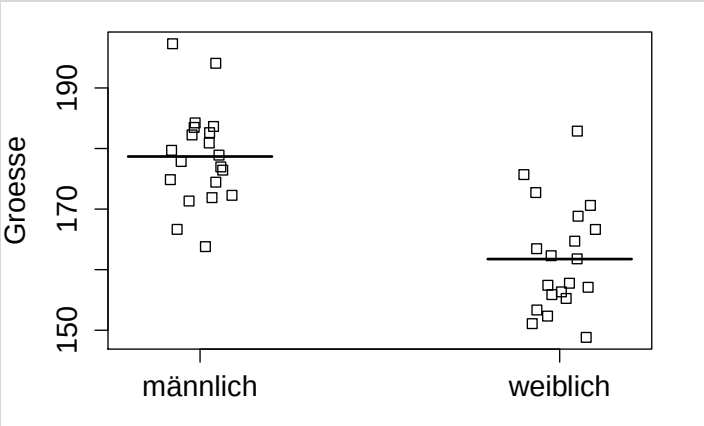
Multiple lineare Regression

Inferenz

Konfidenzintervalle, Vorhersagen und Modelldiagnostik werden analog zur einfachen Regression durchgeführt.

Varianzanalyse: kategoriale Prädiktoren

Beispiel 1: Prädiktor mit zwei Kategorien



Zielgröße: Körpergröße

Prädiktor: Geschlecht

Modell

$$Groesse = \beta_0 + \beta_1 \cdot Geschlecht + \epsilon$$

Kodierung der kategorialen Variable:

$$Geschlecht = \begin{cases} 0, \text{wenn männlich}, \\ 1, \text{wenn weiblich} \end{cases}$$

Varianzanalyse: kategoriale Prädiktoren

Modell

$$Groesse = \beta_0 + \beta_1 \cdot Geschlecht + \epsilon$$

Kodierung der kategorialen Variable:

$$Geschlecht = \begin{cases} 0, \text{wenn männlich}, \\ 1, \text{wenn weiblich} \end{cases}$$

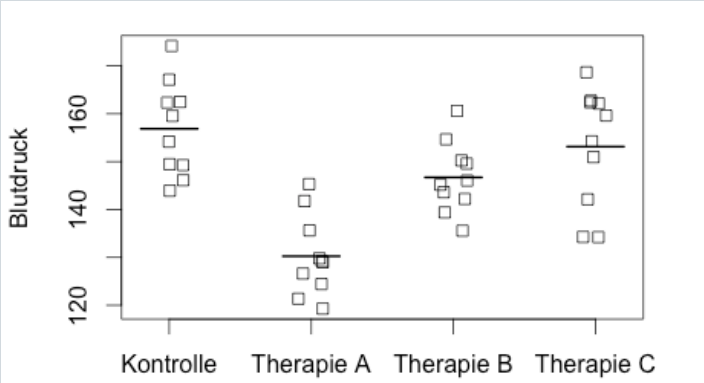
Interpretation der Parameter

β_0: Schätzer für Mittelwert bei Männern
β_1: Schätzer für Mittelwertunterschied zwischen Geschlechtern
$\rightarrow \beta_0 + \beta_1$: Schätzer für Mittelwert der Frauen

Varianzanalyse: kategoriale Prädiktoren

Beispiel 2: Prädiktor mit vier Kategorien

3 Therapiegruppen
1 Kontrollgruppe



Zielgröße: Blutdruck
Prädiktor: Behandlung

β_0 = MW der Basis-Kategorie

wenn Kontrolle: Blutdruck $= \beta_0 + \epsilon$
wenn Therapie A: Blutdruck $= \beta_0 + \beta_1 + \epsilon$
β_1 = Unterschied von TA und Kontrolle (Effekt von TA)

Modell

$$Blutdruck = \beta_0 + \beta_1 \cdot Therapie_A + \beta_2 \cdot Therapie_B + \beta_3 \cdot Therapie_C + \epsilon$$

Kodierung der kategorialen Variable:

$$Therapie_A = \begin{cases} 1, \text{wenn Therapie A,} \\ 0, \text{sonst} \end{cases}$$

andere Variablen analog

wenn Therapie B: Blutdruck $= \beta_0 + \beta_1 + \epsilon$
β_2 = Unterschied von TB und Kontrolle (Effekt von TB)

Varianzanalyse: kategoriale Prädiktoren

Modell

$$Blutdruck = \beta_0 + \beta_1 \cdot Therapie_A + \beta_2 \cdot Therapie_B + \beta_3 \cdot Therapie_C + \epsilon$$

Kodierung der kategorialen Variable:

$$Therapie_A = \begin{cases} 1, \text{wenn Therapie A,} & \text{andere Variablen analog} \\ 0, \text{sonst} \end{cases}$$

Interpretation der Parameter

β_0: Schätzer für Mittelwert in Kontrollgruppe
β_1: Schätzer für Unterschied zwischen Therapie A und Kontrolle
β_2: Schätzer für Unterschied zwischen Therapie B und Kontrolle
β_3: Schätzer für Unterschied zwischen Therapie C und Kontrolle

Varianzanalyse: kategoriale Prädiktoren

Modellbildung

Kodierung von kategorialen Variablen durch Einführung von zusätzlichen Variablen, sogenannter Dummy-Variablen

bei n Kategorien müssen n-1 Parameter geschätzt werden

Art der Kodierung ist abhängig von Fragestellung, z. B.:

- Vergleich zu Kontrollgruppe (Standard in R)
- Vergleich jeweils benachbarter Kategorien (bei ordinalen Variablen)

Varianzanalyse: kategoriale Prädiktoren

Umsetzung in R

```
summary(lm(Blutdruck ~ Behandlung))
```

→ abhängige Variable in Abhängigkeit der unabhängigen Variable

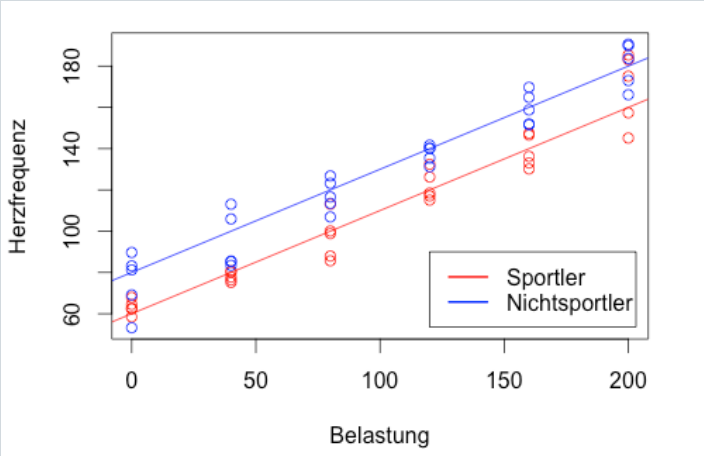
```
##
## Call:
## lm(formula = Blutdruck ~ Behandlung)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.47  -4.60  -1.34    3.80  20.77
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            159.82       2.27   70.41  < 2e-16 ***
## BehandlungTherapie A   -27.48       3.21   -8.56  3.3e-10 ***
## BehandlungTherapie B   -16.11       3.21   -5.02  1.4e-05 ***
## BehandlungTherapie C    -1.68       3.21   -0.52      0.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.18 on 36 degrees of freedom
## Multiple R-squared:  0.731,Adjusted R-squared:  0.709
## F-statistic: 32.7 on 3 and 36 DF,  p-value: 2.22e-10
```

159,82 ⇒ B0

TA und TB sind signifikant unterschiedlich im Vgl. zur Kontrollgruppe

Kovarianzanalyse: metrische und kategoriale Prädiktoren

Beispiel 1: Ergometrie



Zielgröße: Herzfrequenz
Prädiktoren: Belastung, Trainingszustand

$HF = \beta_0 + \beta_2 \times Training + \beta_1 \times Belastung + \epsilon$

Modell

$$Herzfrequenz = \beta_0 + \beta_1 \cdot Belastung + \beta_2 \cdot Training + \epsilon$$

β_0: mittlere HF ohne Belastung für Nichtsportler

Kodierung der kategorialen Variable:

$$Training = \begin{cases} 1, \text{wenn Sportler} \\ 0, \text{wenn Nichtsportler} \end{cases}$$

β_2: Unterschied zw. Sportlern und Nichtsportlern bei gleicher Belastung

β_1: Anstieg der HF, wenn Belastung um eine Einheit steigt

Kovarianzanalyse: metrische und kategoriale Prädiktoren

Modell

$$\textit{Herzfrequenz} = \beta_0 + \beta_1 \cdot \textit{Belastung} + \beta_2 \cdot \textit{Training} + \epsilon$$

Kodierung der kategorialen Variable:

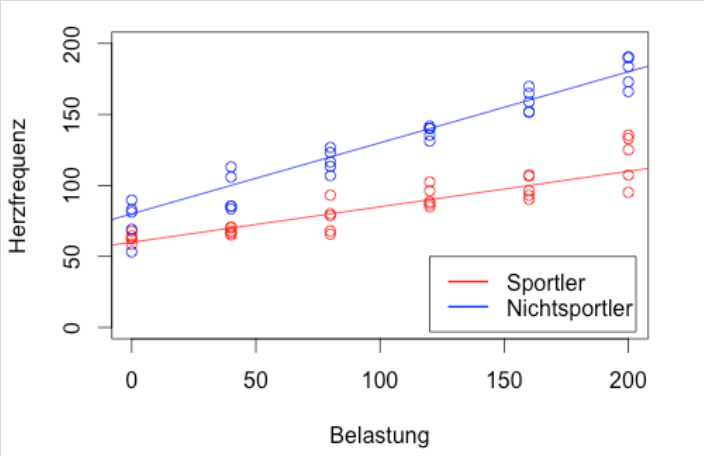
$$\textit{Training} = \begin{cases} 1, \text{wenn Sportler} \\ 0, \text{wenn Nichtsportler} \end{cases}$$

Interpretation der Parameter

β_0: Schnittpunkt mit y-Achse bei Nichtsportlern
β_1: Anstieg der Geraden bei steigender Belastung
β_2: Abstand der Geraden: Unterschied Sportler/Nichtsportler

Kovarianzanalyse: metrische und kategoriale Prädiktoren

Beispiel 2: Ergometrie **mit Interaktion**



für Sportler:
$HF = \beta_0 + \beta_2 + (\beta_1 + \beta_3) \times \text{Belastung} + \epsilon$

Zielgröße: Herzfrequenz
Prädiktoren: Belastung, Trainingszustand

für Nichtsportler:
$HF = \beta_0 + \beta_1 \times \text{Belastung} + \epsilon$

Modell

$$\textit{Herzfrequenz} = \beta_0 + \beta_1 \cdot \textit{Bel.} + \beta_2 \cdot \textit{Training} + \beta_3 \cdot \textit{Bel.} \cdot \textit{Training} + \epsilon$$

$\beta_3 \cdot \textit{Bel.} \cdot \textit{Training}$ = Interaktion

Kodierung der kategorialen Variable:

$$\textit{Training} = \begin{cases} 1, \text{wenn Sportler} \\ 0, \text{wenn Nichtsportler} \end{cases}$$

Kovarianzanalyse: metrische und kategoriale Prädiktoren

Modell

$$\textit{Herzfrequenz} = \beta_0 + \beta_1 \cdot \textit{Bel.} + \beta_2 \cdot \textit{Training} + \beta_3 \cdot \textit{Bel.} \cdot \textit{Training} + \epsilon$$

Kodierung der kategorialen Variable:

$$\textit{Training} = \begin{cases} 1, \text{wenn Sportler} \\ 0, \text{wenn Nichtsportler} \end{cases}$$

Interpretation der Parameter

β_0: Schnittpunkt mit y-Achse bei Nichtsportlern
β_1: Anstieg der Geraden bei steigender Belastung bei Nichtsp.
β_2: Abstand der Geraden: Unterschied Sportler/Nichtsp. ohne Belastung
β_3: Unterschied des Anstiegs zwischen Sp./N.sp. mit steigender Belastung

Kovarianzanalyse: metrische und kategoriale Prädiktoren

Interaktionen zwischen Prädiktoren

- sind nicht-additive Effekte: Wirkung eines Prädiktors hängt von anderem Prädiktor ab
- werden im linearen Modell als Produkt modelliert
- einfaches Beispiel: Rosenkohl schmeckt, Eis schmeckt, aber Rosenkohl mit Eis schmeckt nicht.
- Interaktion zwischen Prädiktoren ist **nicht** Korrelation!

Kovarianzanalyse: metrische und kategoriale Prädiktoren

Umsetzung in R

- ohne Interaktion der Prädiktoren

```
lm(Herzfrequenz ~ Belastung + Training)
```

- mit Interaktion der Prädiktoren

```
lm(Herzfrequenz ~ Belastung + Training + Belastung:Training)
```

oder äquivalent

```
lm(Herzfrequenz ~ Belastung*Training)
```

Kovarianzanalyse: metrische und kategoriale Prädiktoren

Umsetzung in R

```
summary(lm(Herzfrequenz ~ Belastung*Training))


##
## Call:
## lm(formula = Herzfrequenz ~ Belastung * Training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.336  -6.833  -0.289   6.360  21.391
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 74.5552     3.1408   23.74  < 2e-16 ***
## Belastung                    0.5295     0.0259   20.42  < 2e-16 ***
## TrainingSportler           -15.8416     4.4417   -3.57  0.00075 ***
## Belastung:TrainingSportler  -0.2532     0.0367   -6.90    5e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.7 on 56 degrees of freedom
## Multiple R-squared:  0.935,Adjusted R-squared:  0.931
## F-statistic:  267 on 3 and 56 DF,  p-value: <2e-16
```

Modellvielfalt

Die Zahl denkbarer Modelle ist praktisch unbeschränkt

Neue Prädiktoren können generiert werden

- Prädiktoren können miteinander verknüft werden
  z.B. Produkte $\rightarrow$ Interaktionen von Prädiktoren
- Prädiktoren können transformiert werden, z.B. Logarithmus, Wurzeln oder Polynome
  $\rightarrow$ bleibt trotzdem ein lineares Modell in Bezug auf die Parameter!

Wie wählt man das Modell, das man verwenden möchte, aus dieser riesigen Menge aus?

Modellauswahl

Modellauswahl

- Subjektiver Prozess
  - Analyse eines angepassten Modells: eher Handwerk
  - Auswahl eines Modells: eher Kunst
- Vorwissen über Zusammenhänge kann einfließen

mögliche Verfahren

- Backward selection
- Auswahl des Modells nach Akaikes Informationskriterium
- *Purposeful Selection*
- ...

Backward selection

Algorithmus

- Starte mit allen Prädiktoren im Modell (Maximales Modell)
- Entferne den Prädiktor mit dem größten p-Wert größer als α_{crit}
- Passe das kleinere Modell an und gehe zu Schritt 2
- Stoppe, wenn alle p-Werte kleiner als α_{crit} sind

Per Konvention setzt man α_{crit} meist auf 0.05. Wenn es mehr um Vorhersage geht als um Verstehen, kann man α_{crit} auch auf höhere Werte setzen, z.B. 0.1, 0.2.

Akaikes Informationskriterium

Berechnung

nicht über p-Werte, sondern über Restvarianz (bzw. Likelihood):

$$AIC = 2k + n \cdot log(\sigma^2)$$

k: Zahl der Parameter des Modells

Bedeutung

Das AIC für ein Modell ist ein relatives Maß dafür, wie sehr die Daten dieses Modell stützen. Je kleiner der Wert, umso besser.

Umsetzung in R

```
selectedModel <- step(maximalModel)
```

Strategie „purposeful selection“

- Bei vielen Prädiktoren: Vorauswahl von Prädiktorkandidaten
  - nach univariater Signifkanz (etwa $P < 0.25$)
  - oder nach inhaltlicher Relevanz
- Start mit vollem additiven Modell nach Vorauswahl
- Schrittweises Entfernen von Prädiktoren
  - nach Signifikanz (etwa $P > 0.05$ oder > 0.1)
  - und falls kein Confounder (keine große Änderung der βs)
- Überprüfe Hinzunahme nicht vorausgewählter Prädiktoren
- Überprüfe Linearität und Kategorien der Prädiktoren
- Interaktion zwischen den Prädiktoren im Modell
  - falls Interaktion plausibel
  - einzeln nach Signifikanz (etwa $P < 0.05$ oder < 0.01)
  - volles Interaktionsmodell vereinfachen (nach P-Wert)
- Überprüfe Modell-Fit: Passt das Modell zu den Daten?

Quelle: Hosmer, Lemeshow, Applied Logistic Regression, 2000

Regressionsmodelle

Teil 2: Einführung in die Logistische Regression

Thomas Zerjatke
thomas.zerjatke@tu-dresden.de

TU Dresden
Institut für Medizinische Informatik und Biometrie

WS 2023/24

Regression

Lineares Modell

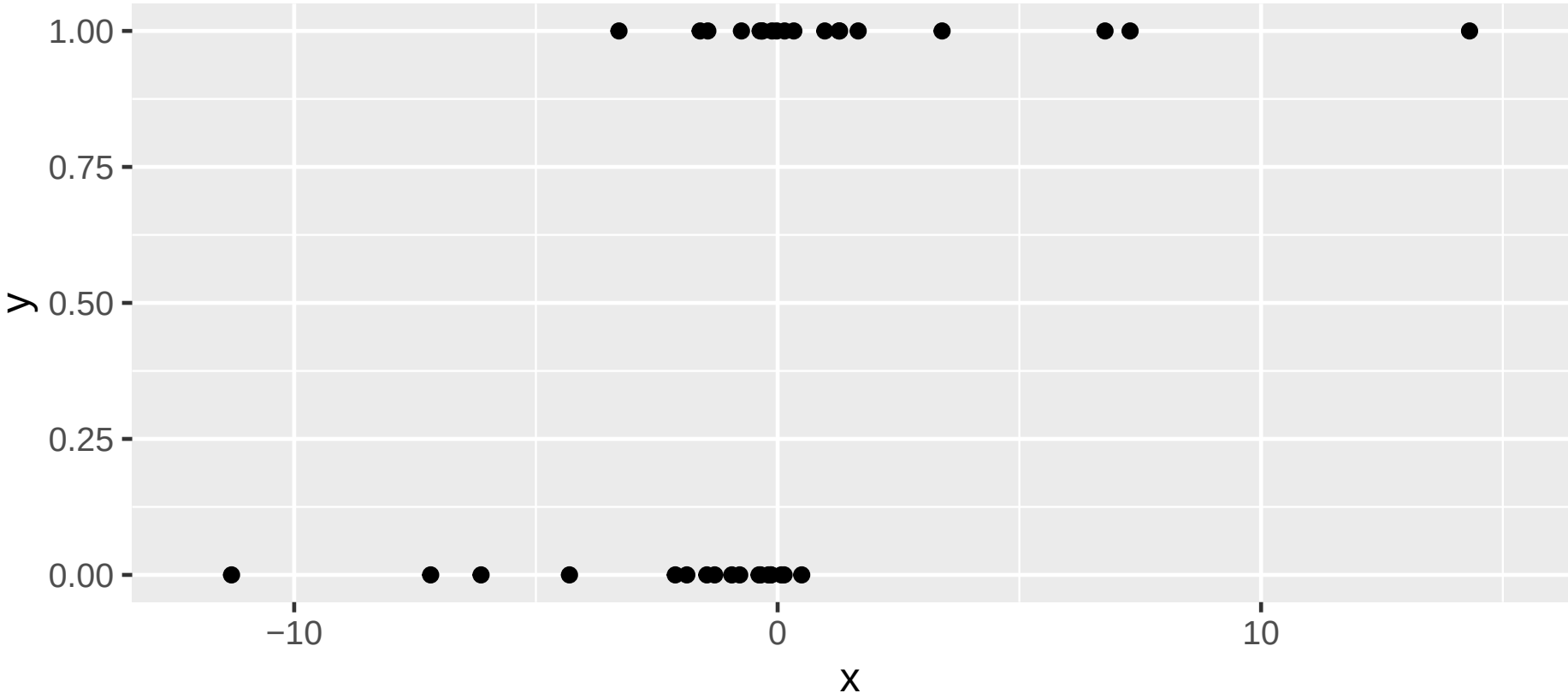
- metrische Zielgröße
- umfasst verschiedene Spezialfälle:
  - Lineare Regression
  - t-Test
  - Varianzanalyse (ANOVA)

Logistische Regression

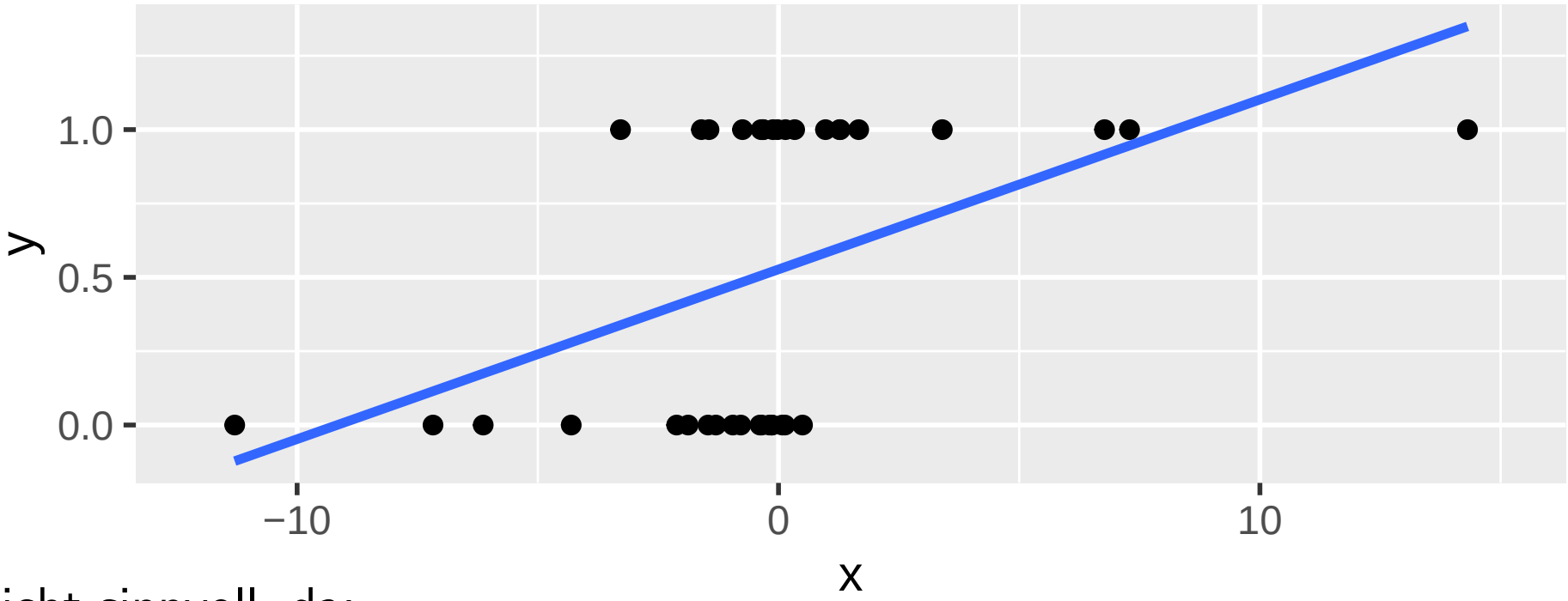
- binäre Zielgröße
- „verallgemeinertes lineares Modell“

Regressionsansatz

- Binäre Zielgröße Y (z. B. 0/1, ja/nein, krank/gesund)
- Einflussgröße X metrisch oder kategorial
- Ziel: Beschreibung der Wahrscheinlichkeit für $P(Y = 1)$ in Abhängigkeit von X



Lineare Regression?



nicht sinnvoll, da:

- Wahrscheinlichkeiten nur zwischen 0 und 1,
  metrische Zielgröße bei linearer Regression aber in ganz $\mathbb{R}$, d.h. zw. $-\infty$ and $+\infty$
- Normalverteilungsannahme und Varianzgleichkeit nicht erfüllt

$\rightarrow$ stattdessen Transformation der Wahrscheinlichkeiten

Wahrscheinlichkeit und Odds

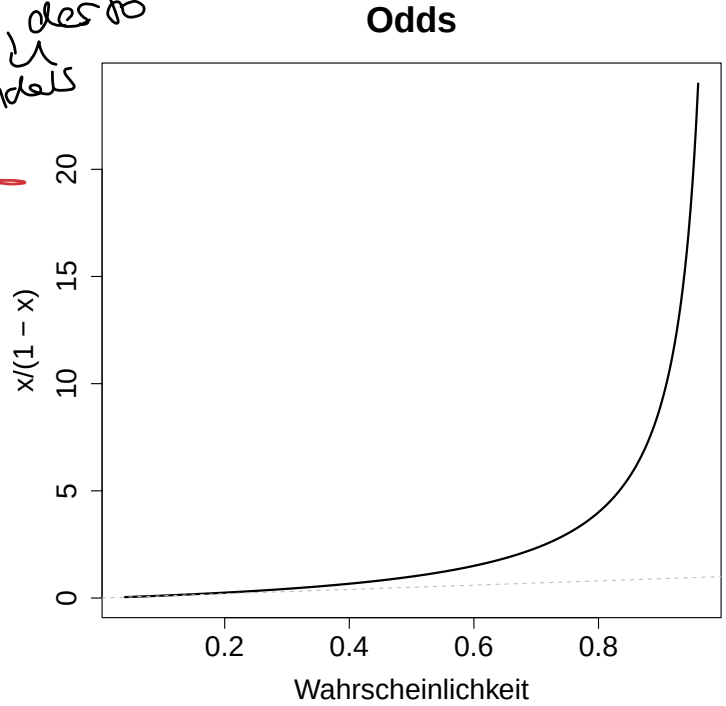
odds	3:1	1:1	1:4	1:99	∞
p	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{5}$	$\frac{1}{100}$	1

Je größer der p, desto mehr geht odds Richtung ∞

Je kleiner die odds, desto kleiner auch p (oder gleich Null)

- Wahrscheinlichkeit p als

$$\frac{\text{Anzahl der Günstigen}}{\text{Anzahl der Möglichen}} \in [0; 1]$$



- Odds (Chance) o als

$$\frac{\text{Anzahl der Günstigen}}{\text{Anzahl der Ungünstigen}} \in [0; \infty) = \mathbb{R}^+$$

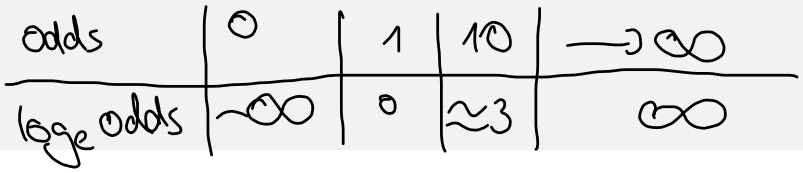
W'keit	0.03	0.10	0.25	0.40	0.50	0.60	0.75	0.90	0.97
Odds	0.03	0.11	0.33	0.67	1.00	1.50	3.00	9.00	32.33

- 1-zu-1-Umwandlung zwischen **Odds** und **Wahrscheinlichkeit**

W'keit zu Odds: $p \mapsto o = \frac{p}{1-p}$

Odds zu W'keit: $o \mapsto p = \frac{o}{1+o}$

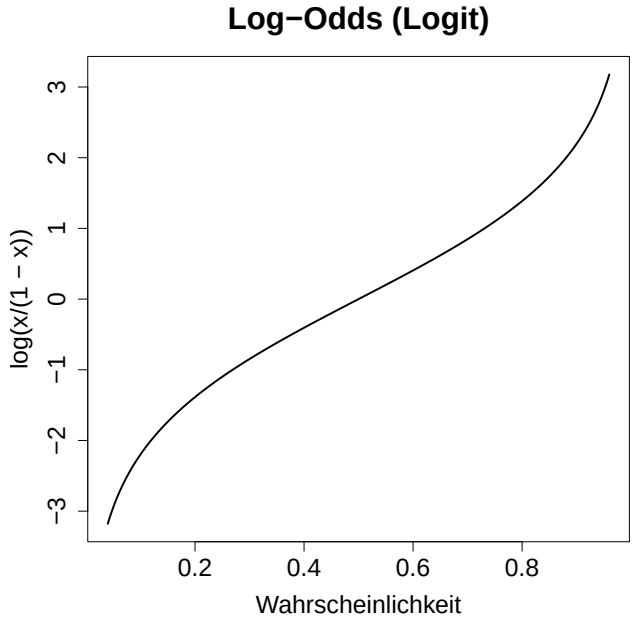
Logits – Log-odds



alle pos. reellen Zahlen

- Odds nur auf $\mathbb{R}^+$, nicht auf ganz $\mathbb{R}$
- Transformation mittels Logarithmus (ln zur Basis e) → natürlicher Logarithmus
- Log-Odds („Logit“) ist

$$\text{logit} = \ln(o) = \ln\left(\frac{p}{1-p}\right) \in \mathbb{R}$$



W'keit	0.03	0.10	0.25	0.40	0.50	0.60	0.75	0.90	0.97
Odds	0.03	0.11	0.33	0.67	1.00	1.50	3.00	9.00	32.33
Logit	-3.48	-2.20	-1.10	-0.41	0.00	0.41	1.10	2.20	3.48

Ansatz Logistische Regression

- Modelliere Logit statt Wahrscheinlichkeit
- Logit hängt ab von Wert des Prädiktors X:
  Lineare Abhängigkeit der Logits wie bei linearer Regression

$$\text{logit}_x = \beta_0 + \beta_1 \cdot x$$

- Interpretation der Koeffizienten wie bei linearer Regression:
  Erhöht sich x um 1, dann ändert sich das Logit um β_1

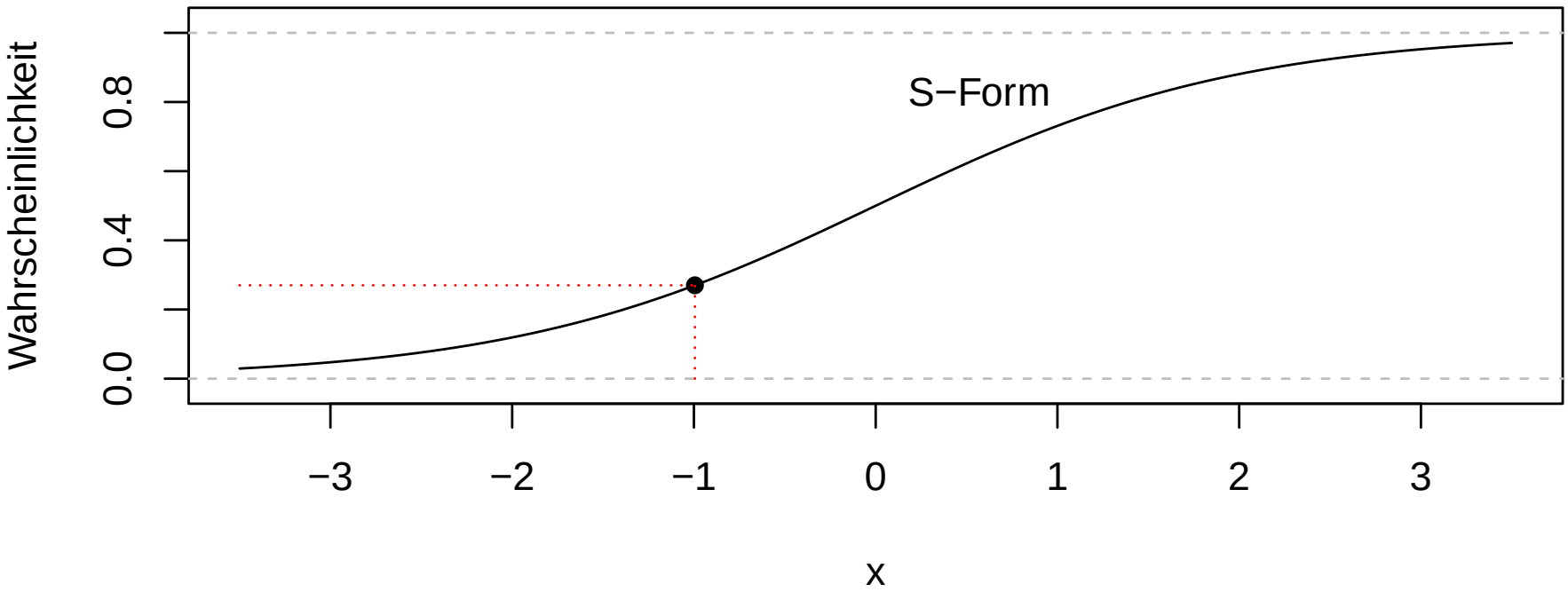
Umwandlung Logit in Wahrscheinlichkeit

- $\text{logit}_x = \ln(o_x) = \beta_0 + \beta_1 \cdot x$
- Odds: $o_x = e^{\beta_0 + \beta_1 \cdot x}$
- Odds o_x in Wahrscheinlichkeit p_x umwandeln via $p_x = \frac{o_x}{1+o_x}$

$$p_x = \frac{o_x}{1+o_x} = \frac{e^{\beta_0+\beta_1 \cdot x}}{1+e^{\beta_0+\beta_1 \cdot x}}$$

$p_x = p$ in Abhängigkeit von x

Logistische Kurve



Lineare Funktion auf der Logit-Skala ergibt eine sigmoide Funktion auf der Skala der Wahrscheinlichkeiten.

$$\underset{\text{Prädiktor}}{x} \quad \mapsto \quad \underset{\text{Logit}}{\beta_0 + \beta_1 \cdot x} \quad \mapsto \quad \underset{\text{Wahrscheinlichkeit}}{\frac{e^{\beta_0+\beta_1 \cdot x}}{1+e^{\beta_0+\beta_1 \cdot x}}}$$

Interpretation des Koeffizienten bei kategorialem Prädiktor

- 2-stufiger Faktor x 0/1-kodiert

$$\text{logit}_x = \ln(odds_x) = \beta_0 + \beta_1 \cdot x$$

$$\begin{aligned} x = 0: \quad & \text{logit}_{x=0} = \beta_0 \\ x = 1: \quad & \text{logit}_{x=1} = \beta_0 + \beta_1 \end{aligned}$$

$$\begin{aligned} x = 0: \quad & \text{odds}_{x=0} = e^{\beta_0} \\ x = 1: \quad & \text{odds}_{x=1} = e^{\beta_0+\beta_1} = e^{\beta_0} \cdot e^{\beta_1} \end{aligned}$$

x	logit	odds
0	β_0	e^{β_0}
1	$\beta_0+\beta_1$	$e^{\beta_0+\beta_1} = e^{\beta_0} \times e^{\beta_1}$
2	$\beta_0+2\times\beta_1$	$e^{\beta_0+2\times\beta_1} = e^{\beta_0} \times e^{\beta_1} \times e^{\beta_1}$

β_0 = Schnittpunkt
β_1 = Anstieg
e^{β_1} = OR wenn x um eine Einheit steigt

- $\text{OR} = \frac{\text{odds}_{x=1}}{\text{odds}_{x=0}} = \frac{e^{\beta_0} \cdot e^{\beta_1}}{e^{\beta_0}} = e^{\beta_1}$

  ist das Odds Ratio (= Chancenverhältnis) zwischen den beiden Stufen der kategorialen Variable.

Interpretation des Koeffizienten bei metrischem Prädiktor

- Einfache logistische Regression mit metrischer Einflussgröße x

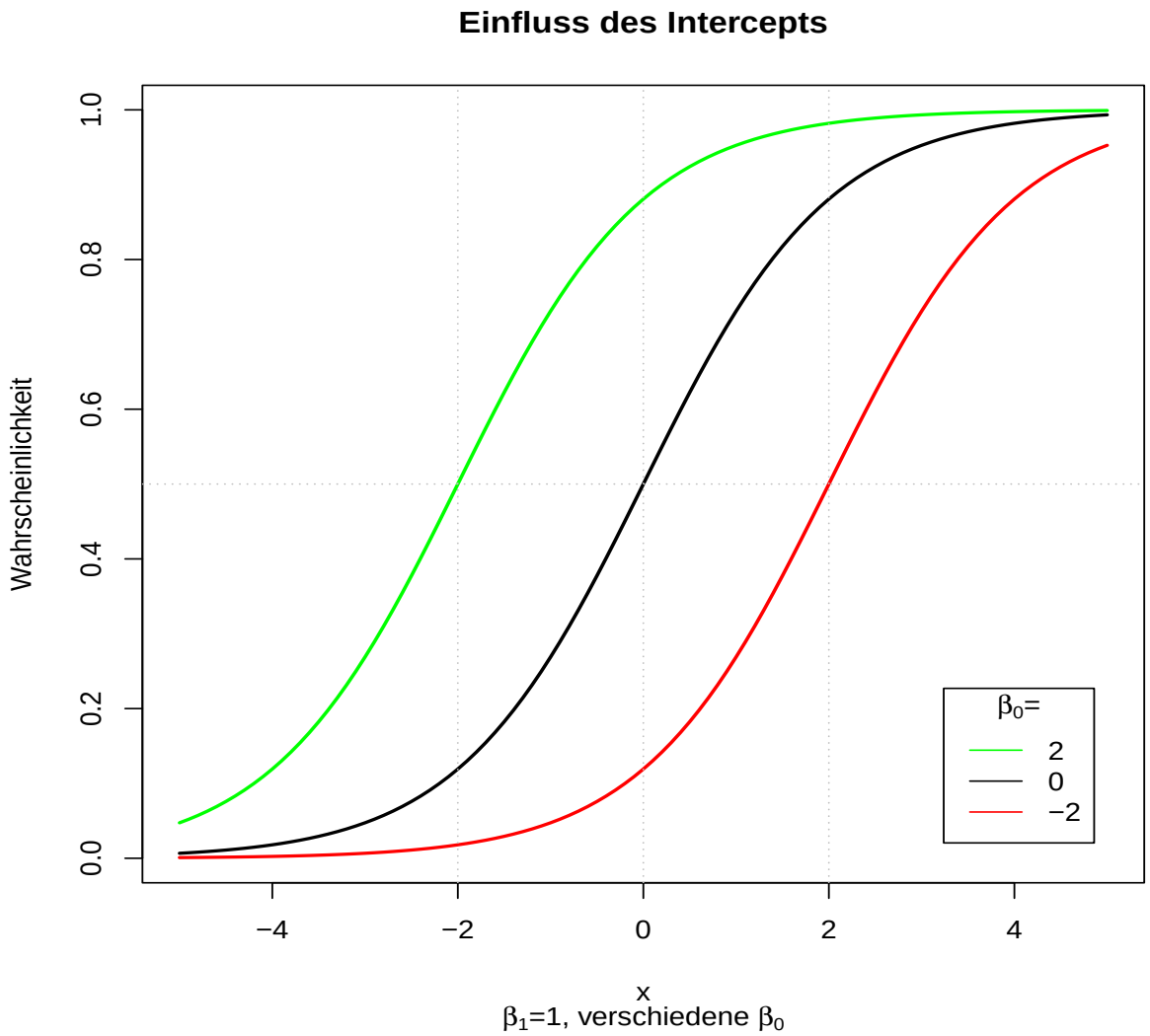
$$\text{logit}_x = \ln(o_x) = \beta_0 + \beta_1 \cdot x$$

- Ändert sich x um eine Einheit, dann ändert sich ...

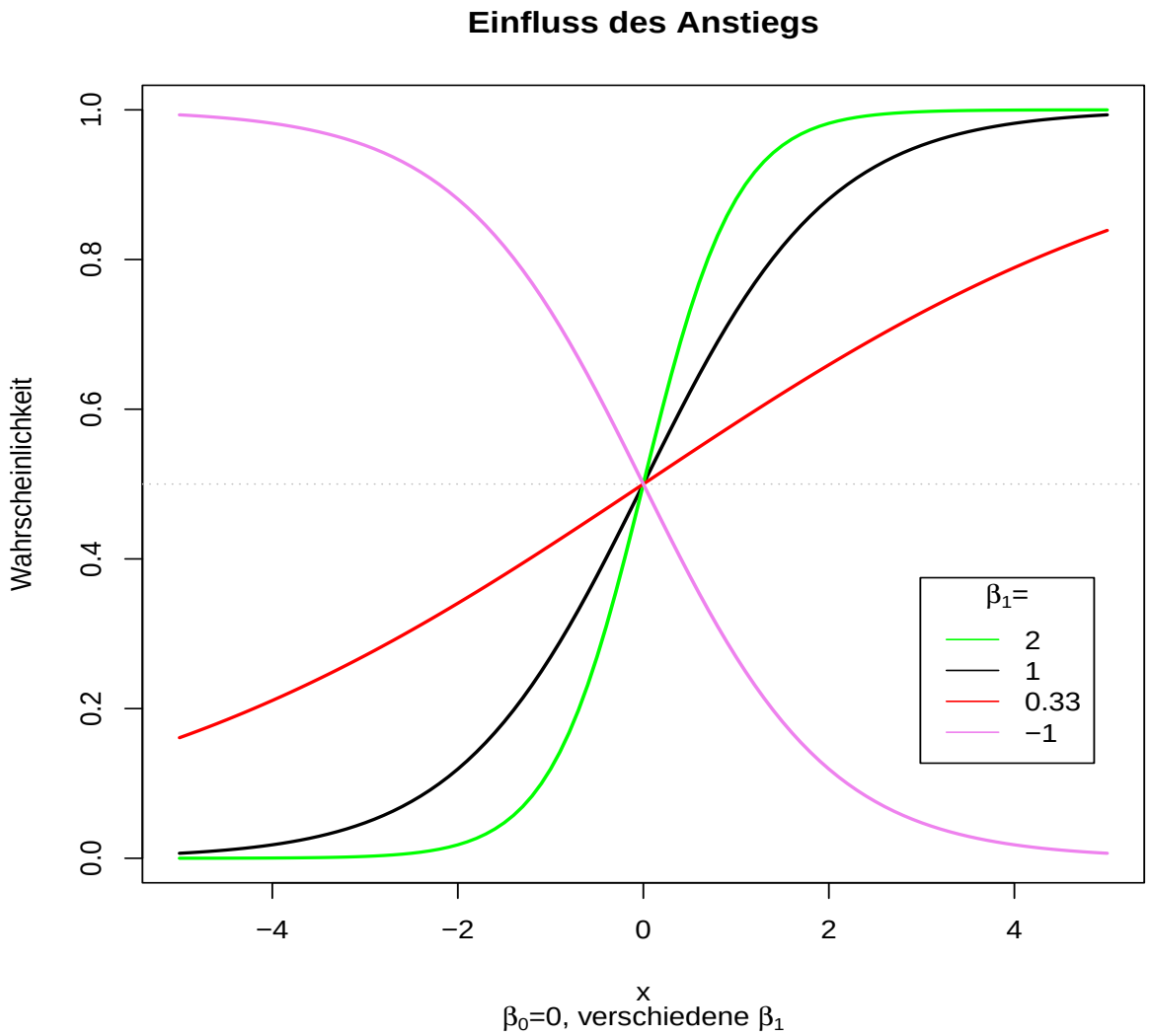
Logit	additiv	$+\beta_1$
Odds	multiplikativ	$\cdot e^{\beta_1}$

- e^{β_1} beschreibt also das Odds Ratio zwischen zwei Subjekten, die sich in x um 1 unterscheiden.

Intercept-Koeffizient β_0



Anstiegs-Koeffizient β_1



Anpassung an Daten

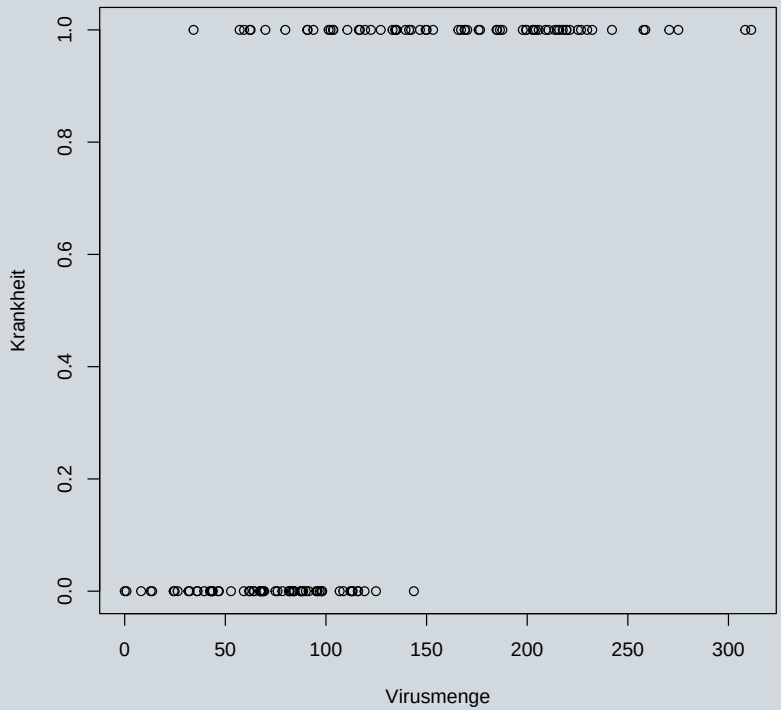
- Methode der kleinsten Quadrate nicht optimal, da Varianz von Y nicht homogen
- „Maximum-Likelihood“ (ML) Methode ist besser geeignet
  - „Welche Parameter machen die beobachteten Daten am wahrscheinlichsten?“
  - ML liefert Punkt-Schätzung $\hat{\beta}$
  - ML liefert Intervall-Schätzung über Standardfehler der $\hat{\beta}$

Logistische Modelle in R

Beispiel

Erkrankung (ja/nein) in Abhängigkeit von Virusmenge

```
plot(Krankheit ~ Virusmenge, data = disease)
```



Logistische Modelle in R

bei linearer Regression: "lm"-Funktion

```
log_model <- glm(Krankheit ~ Virusmenge,
    family = binomial, data = disease)
```

- Funktion `glm()` „Generalized linear model“
- `family=binomial` für logistische Regression (binomiale Zielgröße)
- Zielvariable angeben als
  - 2-stufiger Faktor (z. B. 0/1 oder krank/gesund) *oder*
  - Anzahl Erfolge und Misserfolge in 2-spaltiger Matrix

Das `glm`-Objekt

Analog zur linearen Regression:

- geschätzte Koeffizienten `coef(log_model)`
- Konfidenzintervalle zu den geschätzten Koeffizienten `confint(log_model)`
- Zusammenfassung `summary(log_model)`
- Vorhersagen `predict(log_model)`

Logistische Modelle in R

```
summary(log_model)


##
## Call:
## glm(formula = Krankheit ~ Virusmenge, family = binomial, data = disease)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8619  -0.5822  -0.0553   0.3607   2.5862
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.82351    0.82888   -5.82  5.9e-09 ***
## Virusmenge   0.04428    0.00779    5.68  1.3e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 194.081  on 139  degrees of freedom
## Residual deviance:  99.294  on 138  degrees of freedom
## AIC: 103.3
```

Logistische Modelle in R

Um das Odds Ratio zu bekommen, müssen die Koeffizienten exponenziert werden:

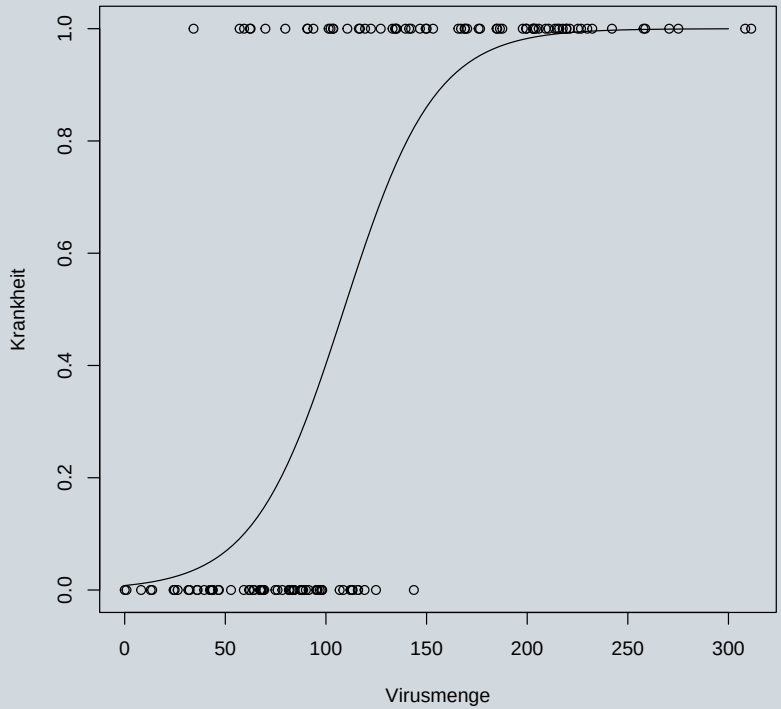
```
exp(coef(log_model))

## (Intercept)  Virusmenge
##     0.00804     1.04528
```

Das Odds Ratio (pro Einheit der Virusmenge) liegt bei 1,04, d. h: Erhöht sich die Virusmenge um eine Einheit, steigt die Chance (das Odds) zu erkranken um 4%.

Logistische Modelle in R

Beispiel

geschätzte Wahrscheinlichkeit für Erkrankung in Abhängigkeit von Virusmenge:



Testen auf statistische Signifikanz

- Signifikanztest eines Parameters $\Rightarrow$ Modellvergleich
- *Relativer* Vergleich eines kleineren mit einem größeren Modell
- Nullhypothese H_0 : „Ist das kleinere Modell ausreichend?“
- Verschiedene Tests bei logistischer Regression möglich
  - **Wald-Test** in R: `summary()`
  - **Likelihood-Ratio-Test (LRT)**
    in R: `drop1(., test="Chisq")`
- LRT bei logistischer Regression bevorzugt

Regressionsmodelle

Teil 2: Einführung in die Logistische Regression
II. Modelldiagnostik

Thomas Zerjatke
thomas.zerjatke@tu-dresden.de

TU Dresden
Institut für Medizinische Informatik und Biometrie

WS 2023/24

Interpretation des Koeffizienten bei metrischem Prädiktor

- Einfache logistische Regression mit metrischer Einflussgröße x

$$\text{logit}_x = \ln(o_x) = \beta_0 + \beta_1 \cdot x$$

- Ändert sich x um eine Einheit, dann ändert sich ...

Logit	additiv	$+\beta_1$
Odds	multiplikativ	$\cdot e^{\beta_1}$

- e^{β_1} beschreibt also das OR zwischen zwei Subjekten, die sich in x um eine Einheit unterscheiden.

Modelldiagnostik: metrische Prädiktoren

Voraussetzung der logistischen Regression

Lineare Abhängigkeit der Logits von metrischer Größe:

$$\text{logit}_x = \ln(o_x) = \beta_0 + \beta_1 \cdot x$$

Überprüfen der Linearität:

- Kategorisierung der metrischen Größe,
  z.B. in vier Kategorien anhand der Quartile
- Anpassen eines neuen Modells mit kategorialem Prädiktor
- Überprüfen, ob geschätzte Koeffizienten ungefähr auf Geraden liegen
- falls nicht:
  - Transformation der metrischen Variablen, z.B. Log oder Wurzel
  - Kategorisierung

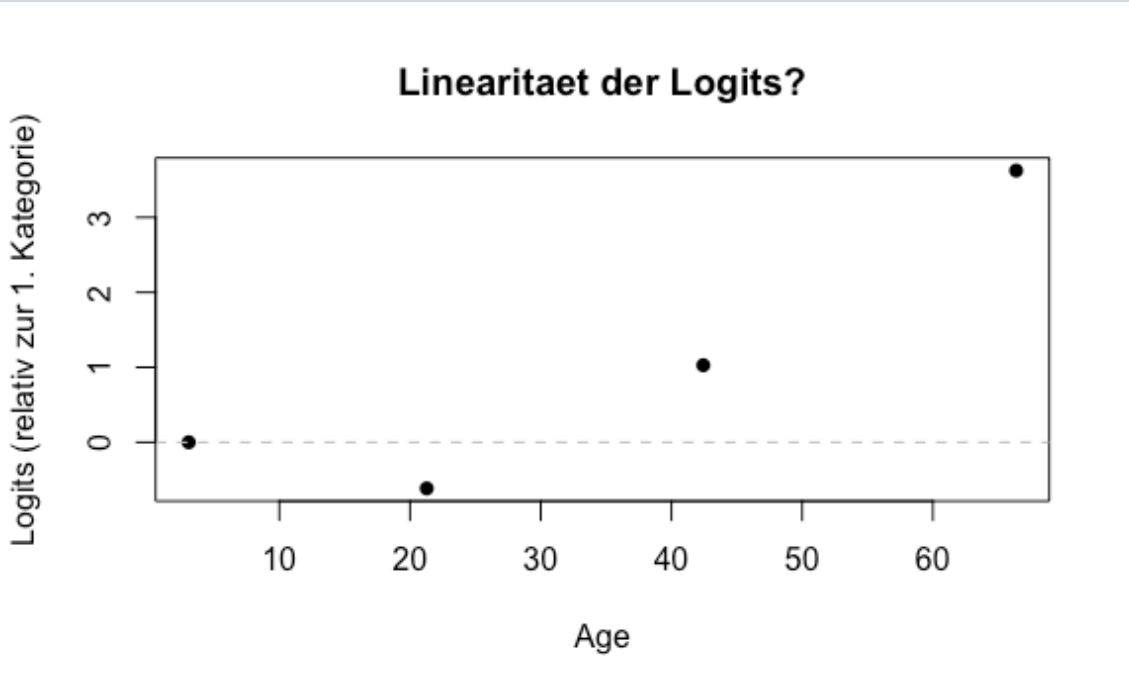
Modelldiagnostik: metrische Prädiktoren

Voraussetzung der logistischen Regression

Lineare Abhängigkeit der Logits von metrischer Größe.

$$\text{logit}_x = \ln(o_x) = \beta_0 + \beta_1 \cdot x$$

Beispiel



Modelldiagnostik: Vorhersagekraft

„Wie gut passt die Vorhersage zu den Daten?“

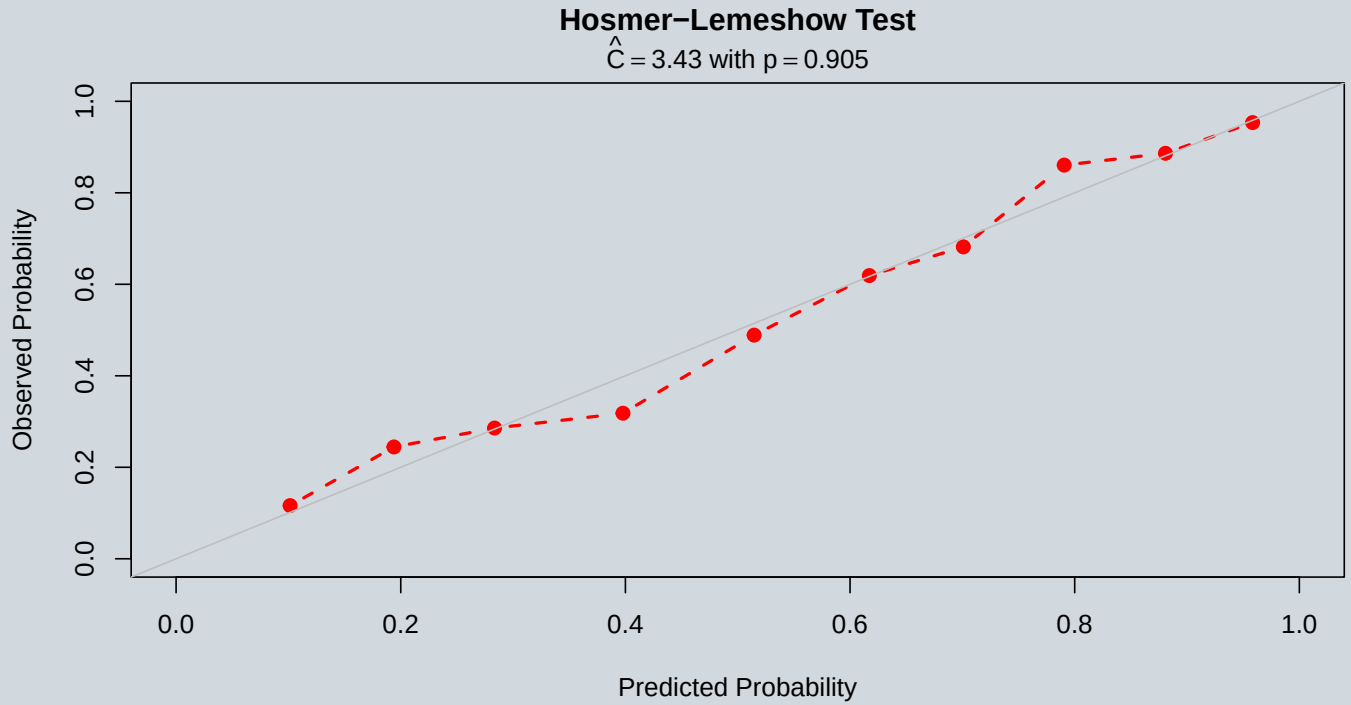
Hosmer-Lemeshow-Anpassungstest

- Gruppiere die Fälle nach der *vorhergesagten* Wahrscheinlichkeit
- Meistens $g = 10$ Gruppen (mindestens $g \geq 6$)
- H_0: beobachtete Wahrscheinlichkeiten $\sim$ vorhergesagte Wahrscheinlichkeiten in den Gruppen
- Kleiner p-Wert heißt schlechte Anpassung

Modelldiagnostik: Vorhersagekraft

„Wie gut passt die Vorhersage zu den Daten?“

Hosmer-Lemeshow-Anpassungstest



Modelldiagnostik: Vorhersagekraft

Klassifizierung

- Logistisches Modell sagt bei gegebenen Prädiktoren X_i die Wahrscheinlichkeit für $Y = 1$ vorher
- Klassifizierung durch Festlegen eines Schwellwerts, z.B. 0.5
- Bewertung eines Modells anhand seiner Klassifizierungsgüte
  - Sensitivität
  - Spezifität

Einführung in die Statistik von Metaanalysen

Wintersemester 2023/24

PD Dr. rer. med. Ingmar Glauche

(*ingmar.glauche@tu-dresden.de*)

Institut für Medizinische Informatik und Biometrie (IMB)
Medizinische Fakultät Carl Gustav Carus
TU Dresden

Motivation

I think it is preferable to accustom a baby to sleeping on his stomach from the beginning if he is willing.

Benjamin Spock in seinem Bestseller *Baby and Child Care* 1946

Motivation

Advice to put infants to sleep on the front for nearly a half century was contrary to evidence available from 1970 that this was likely to be harmful. Systematic review of preventable risk factors for SIDS from 1970 would have led to earlier recognition of the risks of sleeping on the front and might have prevented over 10 000 infant deaths in the UK and at least 50 000 in Europe, the USA, and Australasia.

Ruth Gilbert et al., International Journal of Epidemiology 2005

Inhalt des Teilbereiches

Vorlesung 1

- historische Einordnung
- Effektstärken
- Genauigkeit

Vorlesung 2

- fixed effect Modelle
- random effects Modelle

Vorlesung 3

- Heterogenität

Übung 1

- fixed effect Modelle
- random effects Modelle

Inhalt des Teilbereiches

Vorlesung 4

- Publikationsbias
- Subgruppenanalyse
- Meta-Regression
- *Individual Patient Data (IPD) meta analysis*

Übung 2

- Publikationsbias
- Subgruppenanalyse
- Meta-Regression

Literatur

- Introduction to Meta-Analysis
  Michael Borenstein, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein
  John Wiley & Sons, 2009
- Metaanalyse - Einführung und kritische Diskussion
  Martin Eisend
  Diskussionsbeiträge des Fachbereichs Wirtschaftswissenschaft der Freien Universität Berlin Nr. 2004/8

Historische Einordnung

Universelles Wissen

Gottfried Wilhelm Leibniz (1646 - 1716)

Historische Einordnung

Überblick

- seit dem 17. Jahrhundert unterliegt der Umfang wissenschaftlicher Erkenntnis einem exponentiellen Wachstum
- 1750 gab es etwa zehn wissenschaftliche Zeitschriften
- bis zum Ende des zwanzigsten Jahrhunderts hat sich diese Anzahl alle fünfzig Jahre in etwa verzehnfacht (Zuwachs der Weltbevölkerung im gleichen Zeitraum um einen Faktor 2)
- es ist kaum noch möglich, einen Überblick über alle Forschungsergebnisse selbst in einem klar abgegrenzten Forschungsgebiet zu bekommen und zu behalten
- zu einer Fragestellung liegen oftmals mehrere Untersuchungen vor, die uneinheitliche und manchmal sogar widersprüchliche Befunde ausweisen

Historische Einordnung

Überblick

- aus dieser Entwicklung leitet sich ein Bedarf an Möglichkeiten der Informationsverdichtung und -bewertung von wissenschaftlichen Forschungsergebnissen ab, der schließlich zur Entwicklung unterschiedlicher Methoden der Ergebniszusammenfassung führte
- neben den traditionellen Formen wie den Reviews haben seit Mitte der 1970er Jahre auch quantitative Ergebniszusammenfassungen, so genannte Metaanalysen, immer mehr an Bedeutung in den verschiedensten Disziplinen mit empirischer Ausrichtung gewonnen

Historische Einordnung

Entwicklung von Metaanalysen

- Karl Pearson (1857-1936): Fragestellungen zum Erfolg von Impfungen gegen Typhus
- Zusammenfassung der Resultate kleinerer Studien (Korrelationen zwischen Impfung und Todeswahrscheinlichkeit) um so eine Verbesserung der Parameterschätzung auf der Basis einer größeren Stichprobe zu erhalten (Pearson 1904)
- methodische Weiterentwicklungen in den folgenden Jahrzehnten vor allem im Bereich der Agrarforschung und der Biostatistik (bspw. Berücksichtigung von Stichprobengrößen oder alternative Zusammenfassung von Signifikanzniveaus), ab den fünfziger Jahren auch im Bereich der Psychologie und den Erziehungswissenschaften

Historische Einordnung

Entwicklung von Metaanalysen

- Gene V. Glass prägte Mitte der siebziger Jahre den Begriff "Metaanalyse" im Bereich der Erziehungswissenschaften/Psychologie für die von ihm entwickelte Methode zur quantitativen Ergebnisintegration
- danach deutliche Zunahme der Anzahl der durchgeführten quantitativen Ergebniszusammenfassungen und der systematische Auseinandersetzung mit metaanalytischen Methoden
- Übernahme der Methodik in andere Wissensgebiete, z.B. Medizin, empirische Sozialforschung, Betriebswirtschaftslehre

Historische Einordnung

Definition

- Gene Glass (1976): "Meta-analysis refers to the analysis of analyses"
- Arno Drinkmann (1990) zur Definition der Metaanalyse: "eine an den Kriterien empirischer Forschung orientierte Methode zur quantitativen Integration der Ergebnisse empirischer Untersuchungen sowie zur Analyse der Variabilität dieser Ergebnisse" → Wo kommen die Unterschiede her?
- Metaanalysen beruhen immer auf empirischen Untersuchungen und können daher auch keine Integration theoretischer oder konzeptioneller Arbeiten leisten.
- Metaanalysen benötigen quantitative empirische Ergebnisse, so dass auch Ergebnisse qualitativer Untersuchungsformen wie beispielsweise aus Fallstudien nicht metaanalysierbar sind.

Historische Einordnung

Abgrenzung

- **Narrative Reviews** bieten meist einen breiten Überblick zu einem bestimmten Thema
- **Systematischer Review** umfasst die Erstellung eines detaillierten Studienprotokolls und Auswerteplans, sowie eine Literaturrecherche geeigneter Studien nach a priori definierten Ein- und Ausschlusskriterien
- **Metaanalyse** erweitert das Vorgehen um eine quantitative Zusammenfassung der Ergebnisse

Aufbau

Prototypischer Aufbau einer Metaanalyse

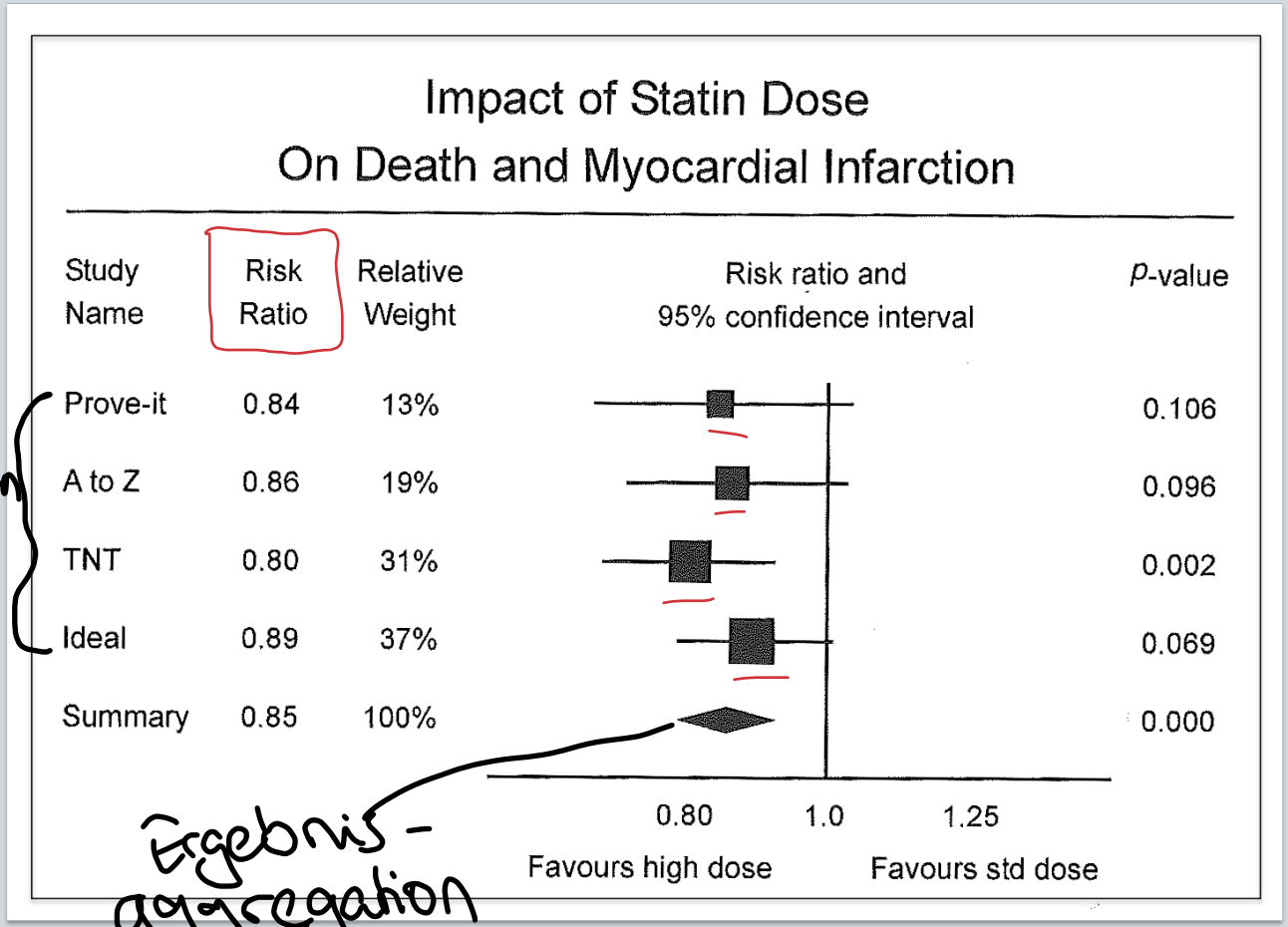
Vorgehensweise und Ablauf einer Metaanalyse sind mit dem entsprechenden Vorgehen in Einzeluntersuchungen vergleichbar

- Konkretisierung des Forschungsvorhabens
- Sammlung relevanter Untersuchungen
- Codierung und Bewertung der Untersuchungen
- **Datenanalyse**
- **Darstellung und Interpretation** der Ergebnisse

Überblick Datenanalyse

H0 : Risk Ratio = 1

Elemente einer Metaanalyse - Forest Plot



4 Studien

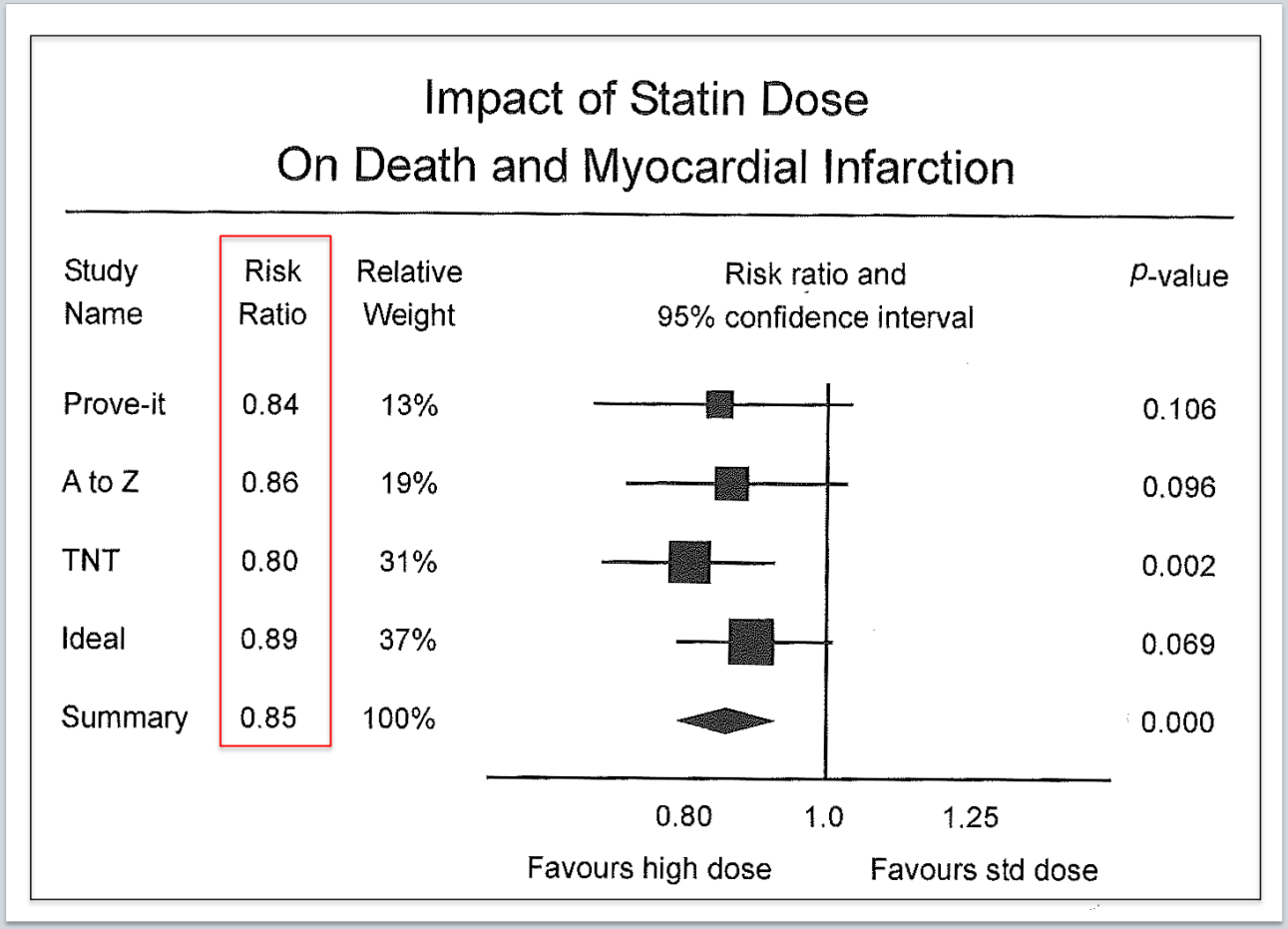
Ergebnis-aggregation

je größer die Studien, desto genauer kann geschätzt und desto höher wird gewichtet

⇒ Größe des Quadrats

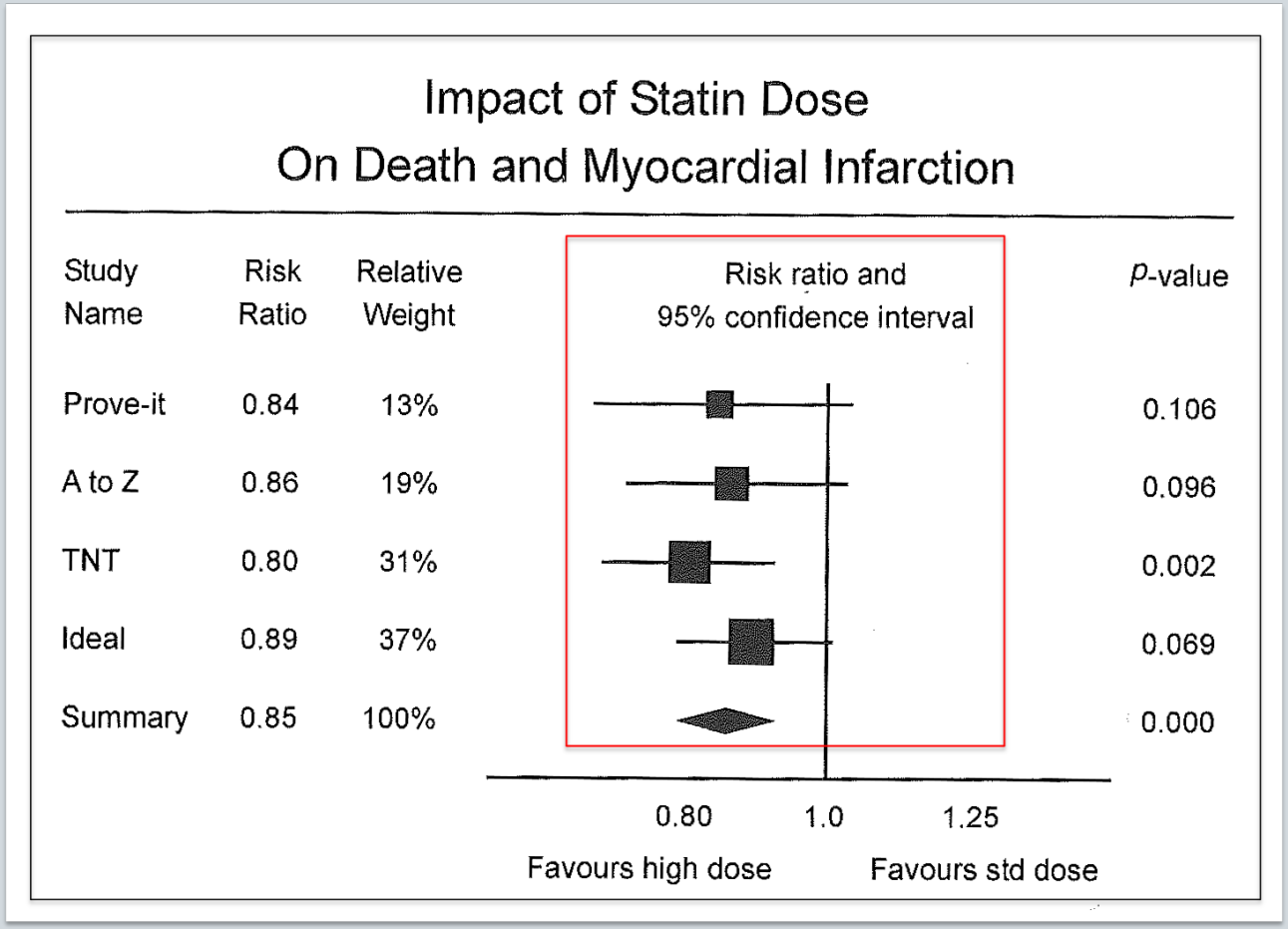
Überblick Datenanalyse

Elemente einer Metaanalyse - Effektstärken



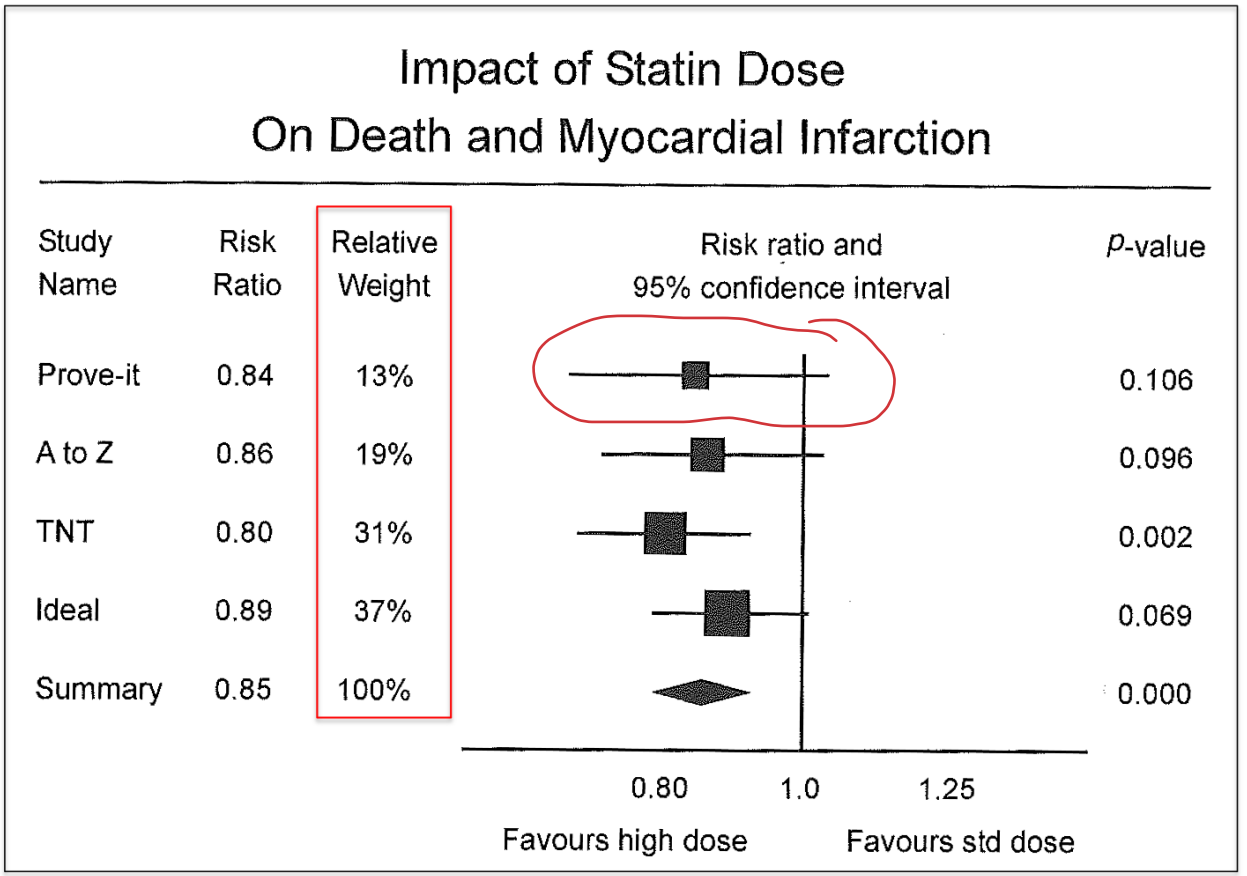
Überblick Datenanalyse

Elemente einer Metaanalyse - Genauigkeit



Überblick Datenanalyse

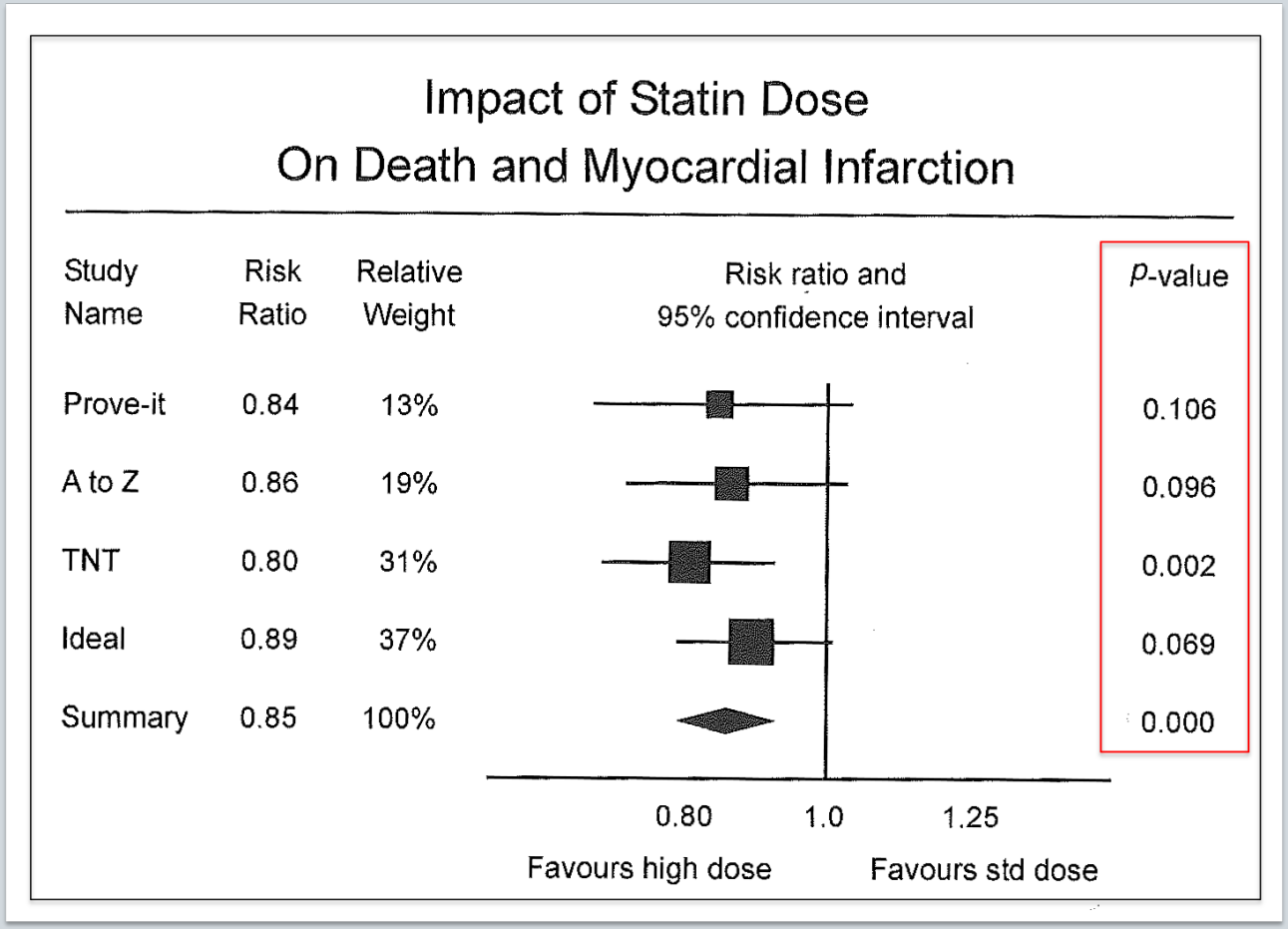
Elemente einer Metaanalyse - Gewichte



Studie mit größtem Whisker hat kleinstes Gewicht

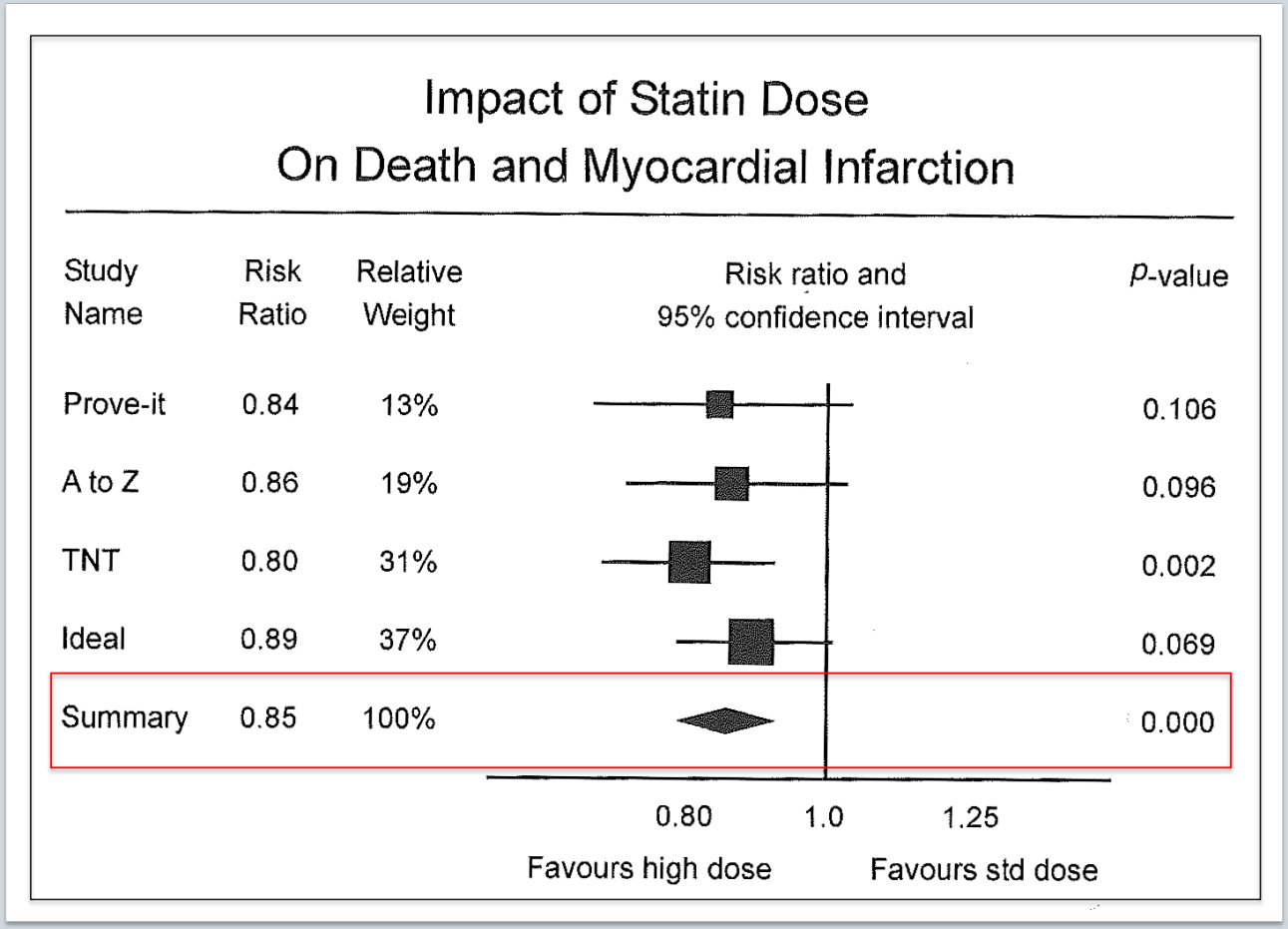
Überblick Datenanalyse

Elemente einer Metaanalyse - p-Werte



Überblick Datenanalyse

Elemente einer Metaanalyse - Summary effect



Effektstärken und Genauigkeit

Effektstärken

- Effektstärken (engl: effect size or treatment effect) dienen der Beschreibung der Beziehung zweier Variablen, sind als solche primäre Zielgrößen der Metaanalyse
- Auswahl der geeigneten Zielgröße für die Beschreibung der Effektstärke hängt von der Fragestellungen und von technischen Überlegungen ab:
  - Vergleichbarkeit
  - Berechenbarkeit
  - Interpretierbarkeit

Effektstärken: x Mittelwertdifferenzen
x Binäre Messungen
(OR, RR)

Effektstärken: Mittelwerte

Effektstärkenbeschreibung durch Mittelwerte

- für Studien, die Mittelwerte und Standardabweichung als primäres Resultat angeben, werden häufig Differenzen bzw. standardisierte Differenzen als Effektgrößen verwendet
- die Beschreibung der Effektstärke als Mittelwertdifferenz ist geeignet, wenn Meßwerte auf üblichen Skalen angegeben werden

Beispiel: Effektstärkenbeschreibung durch Mittelwerte

- Blutdruck
- Benotungen Abitur

Effektstärken: Mittelwerte

unstandardisierte Differenz von Mittelwerten

- Differenz der (Populations-) Mittelwerte

$$\Delta = \mu_1 - \mu_2$$

- gesucht ist ein Schätzer D für diesen Parameter
- aus den Stichprobenmittelwerte $\bar{X}_1$ und $\bar{X}_2$ zweier unabhängiger Stichproben (bspw. Treatment and Control) lässt sich der Schätzer für die Differenz D berechnen:

$$D = \bar{X}_1 - \bar{X}_2$$

muss für jede Studie in einer Metanalyse berechnet werden

Einschub: Varianzen

nochmal anschauen !

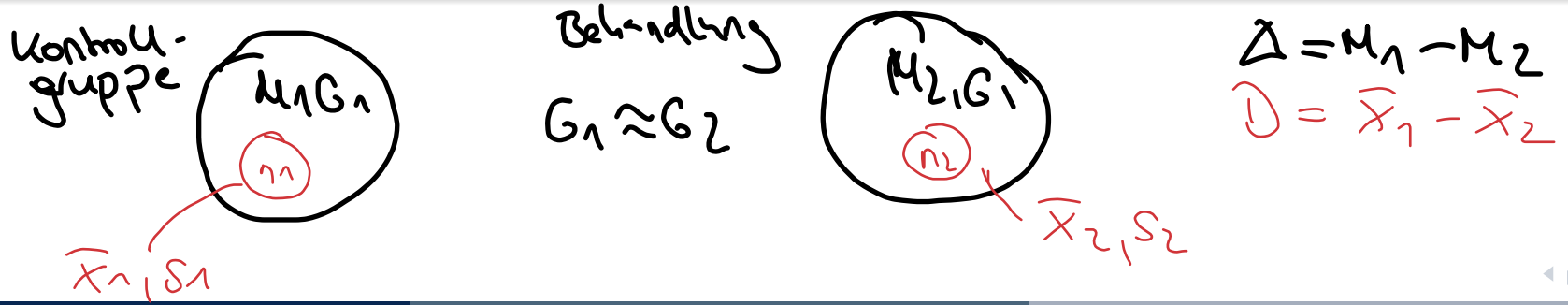
Varianz

die Varianz bewertet die zu erwartende Streuung einer Zufallsvariablen

$$Var[x] = \sigma^2 = \sum_i^n (x_i - \mu)^2 \cdot f(x_i)$$
$$= E[(x - \mu)^2]$$

Standardabweichung

$$\sigma = \sqrt{Var[x]} = \sqrt{\sigma^2}$$



Einschub: Varianzen

$$Var(D) = V_D = Var(\bar{x}_1 - \bar{x}_2) = Var(\bar{x}_1) + Var(\bar{x}_2)$$
$$= \frac{s_{p1}^2}{n_1} + \frac{s_{p2}}{n_2}$$
$$= \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

Varianz für Einzelmessung und Mittelwert

Zufallsvariable Einzelmessung x Mittelwert $\bar{x}$

Varianz

$$Var[x] = s_x^2 = \frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2$$

$$Var[\bar{x}] = s_{\bar{x}}^2 = \frac{s_x^2}{n}$$

Standardabweichung

$$s_x = \sqrt{Var[x]}$$

$$s_{\bar{x}} = \sqrt{Var[\bar{x}]} = \frac{s_x}{\sqrt{n}}$$

Effektstärken: Mittelwerte

Varianz der unstandardisierten Differenz von Mittelwerten

- unter Annahme von Varianzgleichheit ergibt sich für zwei Gruppen mit Standardabweichung S_1 und S_2 und Fallzahlen n_1 und n_2 die Varianz der Differenz D:

$$V_D = \frac{n_1 + n_2}{n_1 n_2} S^2_{\text{pooled}}$$

mit

$$S_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

- der Standardfehler von D ist definiert als

$$SE_D = \sqrt{V_D}$$

Effektstärken: Mittelwerte

Beispiel: unstandardisierten Differenz von Mittelwerten

- Studie mit zwei Gruppen: Mittelwerte $\bar{X}_1 = 103.0$ und $\bar{X}_2 = 100.0$ mit Standardabweichung $S_1 = 5.5$ und $S_2 = 4.5$ und Stichprobenumfang $n_1 = n_2 = 50$
- Mittelwertdifferenz

$$D = 103.0 - 100.0 = 3.0$$

- gepoolte Standardabweichung

$$S_{\text{pooled}} = \sqrt{\frac{(50-1)5.5^2 + (50-1)4.5^2}{50+50-2}} = 5.0249$$

- Varianz

$$V_D = \frac{50+50}{50 \cdot 50} \cdot 5.0249^2 = 1.01$$

Effektstärken: Mittelwerte

standardisierte Differenz von Mittelwerten

- für Studien, in denen unterschiedliche Skalen definiert sind, bietet es sich an, die Differenz der Mittelwerte mit der Standardabweichung zu normieren
- für zwei unabhängige Gruppen mit den (Populations-) Mittelwerten μ_1 und μ_2 und den Standardabweichungen σ_1 und σ_2 ergibt sich die standardisierte Differenz der Mittelwerte:

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} = \frac{\text{Mittelwertdifferenz}}{\text{Standardabweichung}}$$

wobei angenommen wird, dass die Standardabweichung der beiden Gruppen gleich ist ($\sigma_1 = \sigma_2 = \sigma$)

Effektstärken: Mittelwerte

standardisierte Differenz von Mittelwerten

- aus diesen Überlegungen ergibt sich der Schätzer für die standardisierte Differenz der Mittelwerte in einer Studie:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{S_{\text{within}}}$$

mit

$$S_{\text{within}} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Effektstärken: Mittelwerte

Varianz der standardisierten Differenz von Mittelwerten

- unter Annahme von Varianzgleichheit wird die Varianz der Differenz d (in guter Approximation) geschätzt durch:

$$V_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}$$

- der erste Term beschreibt die Unsicherheit in der Schätzung der mittleren Differenz, der zweite Term die Unsicherheit von S_{within}
- der Standardfehler von d ist definiert als

$$SE_d = \sqrt{V_d}$$

Effektstärken: Mittelwerte

Korrektur: Hedges' g

- für kleine Fallzahlen überschätzt d den wahren Wert $\delta \rightarrow$ bias-korregierter Schätzer: Hedges' g

$$g = J \cdot d$$

= Korrekturfaktor(J) × $\frac{\bar{x}_1 - \bar{x}_2}{SD}$

und

$$V_g = J^2 \cdot V_d$$

- Korrekturfaktor

$$J = 1 - \frac{3}{4df - 1}$$

mit Freiheitsgraden $df = n_1 + n_2 - 2$

von einer Studie ?

Genauigkeit

Varianz als Maß der Genauigkeit

- Varianz V_y wird für jede Effektstärke Y unterschiedlich ermittelt
- Berechnung des Standardfehler der Effektstärke

$$SE_Y = \sqrt{V_Y}$$

- damit können die Grenzen des 95% Konfidenzintervalls berechnet werden (Annahme: Normalverteilung)

$$LL_Y = Y - 1.96 \cdot SE_Y$$

$$UL_Y = Y + 1.96 \cdot SE_Y$$

Genauigkeit

Faktoren

- Stichprobenumfang ist ein zentraler Faktor für die Genauigkeit eines Schätzers
- ebenso beeinflusst das Studiendesign die Genauigkeit des Schätzers (matched groups vs. independent groups)

Bedeutung für die Metaanalyse

- Stichprobenumfang und Studiendesign gehen in die Wichtung der Studien für die Metaanalyse ein
- Studien mit höhere Genauigkeit werden stärker berücksichtigt

Genauigkeit

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p} \qquad Vd = \frac{n_1 + n_2}{n_1 \times n_2} + \frac{d^2}{2(n_1 + n_2)}$$

= weil $d = 0$

Beispiel: Einfluss des Stichprobenumfang

I II III

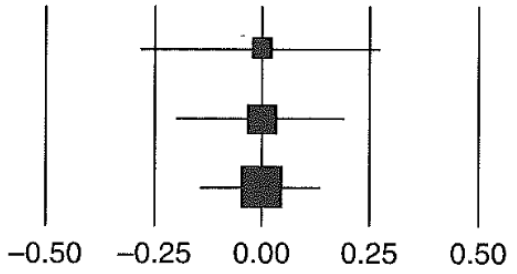
- 3 Studien ($n = 100, 200, 400$) mit Mittelwertvergleich in je zwei unabhängigen Gruppen
- standardisierte Mittelwertdifferenz $d = 0.0$ → für alle 3 Studien

I: $V_{d_1} = \frac{100+100}{10.000} = \frac{2}{100} = \frac{1}{50}$

II: $V_{d_2} = \frac{200+200}{20.000} = \frac{4}{40.000} = \frac{1}{100}$

Impact of Sample Size on Variance

N per group	Standard Error	Variance
N = 100	0.141	0.020
N = 200	0.100	0.010
N = 400	0.071	0.005



$SEM_{d_1} = \frac{1}{\sqrt{50}}$

$SEM_{d_2} = \frac{1}{\sqrt{100}}$

Von I zu II: n wurde verdoppelt

- Fläche der Box ist invers proportional zur Varianz
- CI ist proportional zum Standardfehler

Vorlesung 2

- fixed effect Modelle
- random effects Modelle

Durchführung von Metaanalysen

Vorbetrachtungen

- **Ziel der Metaanalyse**: Berechnung eines gewichteten Mittelwertes
- Metaanalysen werden im Wesentlichen als *fixed-effect* oder *random-effects* Modelle berechnet
- **fixed-effect Modelle** basieren auf der Annahme, dass eine *wahre Effektstärke* gibt; alle Unterschiede der beobachteten Effektstärken sind Stichprobenfehler
- bei einem **random-effects Modell** wird angenommen, dass der wahre Effekt von Studie zu Studie variieren kann (*unterschiedliche Effektstärken*); die beobachteten Effektstärken sind eine zufällige Stichprobe aus der Verteilung der zu Grunde liegenden Effektstärken

Durchführung von Metaanalysen

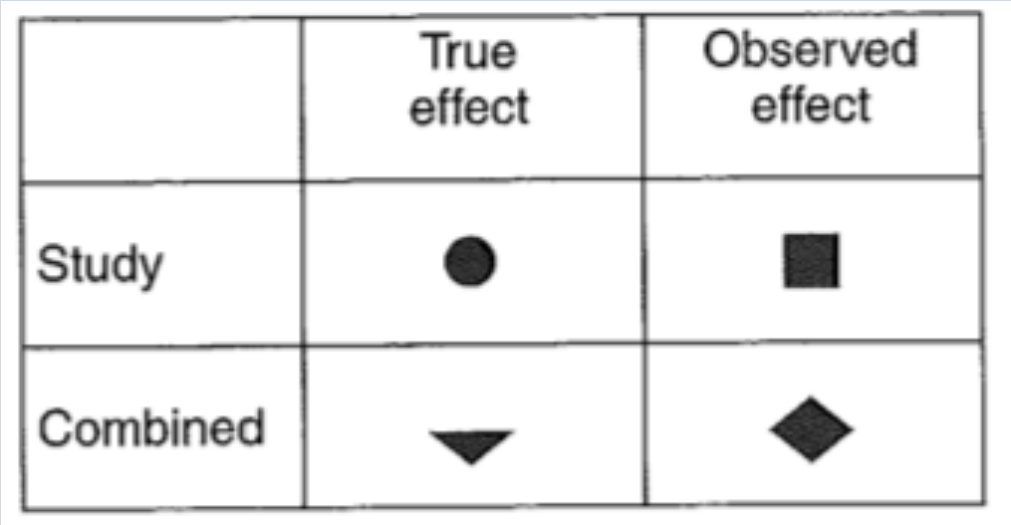
Aufbau der Vorlesung

- fixed-effect Modell
- random-effects Modell
- exemplarische Berechnung für ein Beispiel (standardisierte Differenz von Mittelwerten)
- konzeptuelle Unterschiede zwischen den Modellen
- weiteres Beispiel (Odds ratio)

Durchführung von Metaanalysen

Nomenklatur

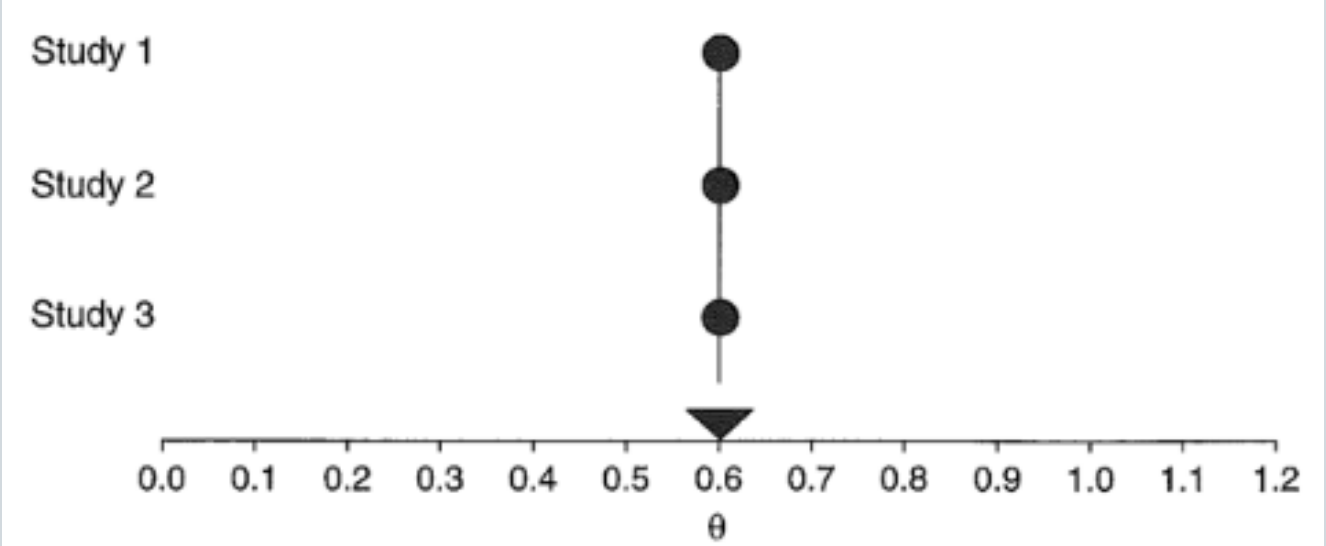
Unterscheidung zwischen *wahren Effektstärken* und *gemessenen Effektstärken*



fixed-effect Modell

wahren Effektstärken

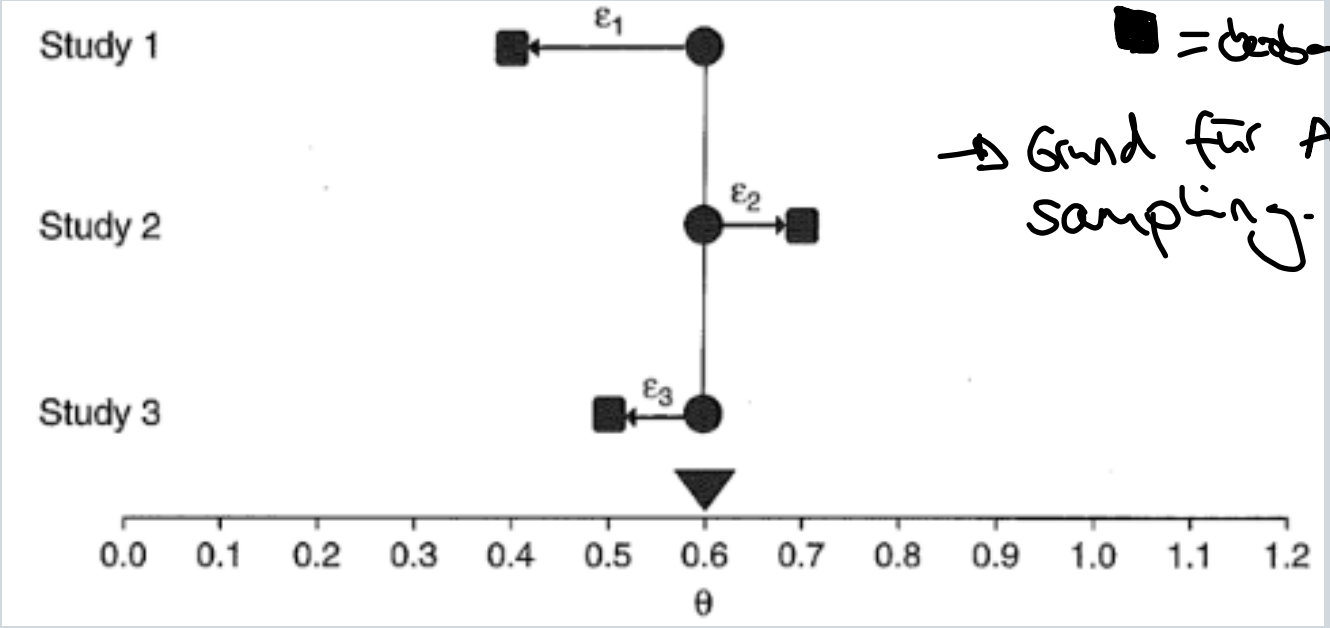
alle Studien weisen eine gemeinsame, wahre (unbekannte) Effektstärke θ auf



fixed-effect Modell

beobachtete Effektstärken

Variationen zwischen den Studien beruhen nur auf zufälligen Fehlern ϵ_i durch den endlichen Stichprobenumfang



● = wahrer Wert

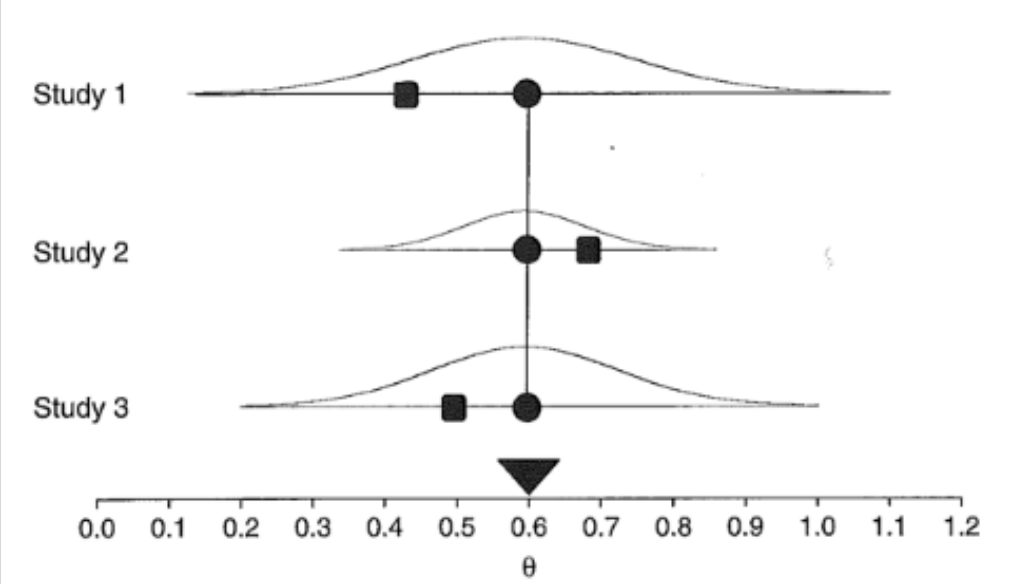
■ = beobachteter Wert

→ Grund für Abweichung: sampling-effect

fixed-effect Modell

beobachtete Effektstärken

- beobachtete Effektstärke $Y_i = \theta + \epsilon_i$
- Fehler ϵ_i ist zufällig für jede Studie, allerdings kann die Verteilung der Fehler durch eine Normalverteilung approximiert werden
- Breite der Normalverteilung skaliert mit dem Standardfehler



fixed-effect Modell

Berechnungsvorschrift

- genauere Studien sollten höheres Gewicht bekommen

$$\text{(Gewicht)}\ W_i = \frac{1}{V_{Y_i}}\ \text{(Varianz)}$$

bei großer Studie: → kleine Varianz → großes Gewicht

- das gewichtete Mittel M berechnet sich nach

Gewichtungsfaktor

$$M = \frac{\sum_{i=1}^{k} W_i Y_i}{\sum_{i=1}^{k} W_i}$$

fixed-effect Modell

Berechnungsvorschrift

- Varianz der gemittelten Effektstärke (summary effect) wird abgeschätzt durch

$$V_M = \frac{1}{\sum_{i=1}^{k} W_i}$$

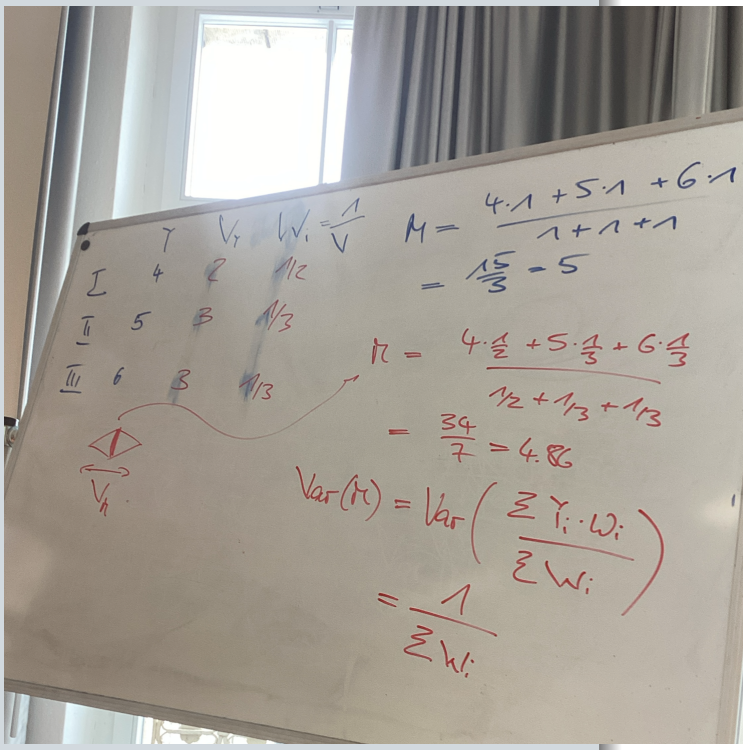
- Standardfehler der gemittelten Effektstärke

$$SE_M = \sqrt{V_M}$$

- Grenzen des 95% Konfidenzintervalls

$$LL_M = M - 1.96 \cdot SE_M$$

$$UL_M = M + 1.96 \cdot SE_M$$



fixed-effect Modell

Testvorschrift

- Test gegen die Nullhypothese, dass der wahre Effekt $\theta = 0$ ist
- z-Wert berechnen

$$z = \frac{M}{SE_M}$$

- für einen zweiseitigen Test ergibt sich

$$p = 2(1 - \Phi(|z|))$$

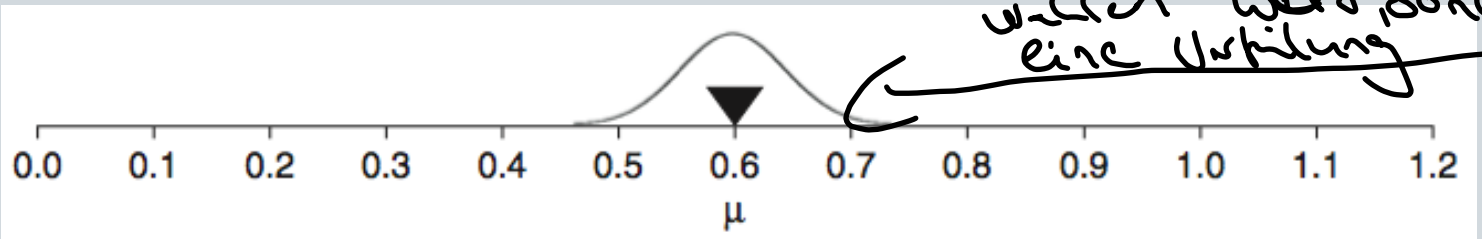
wobei Φ der kumulativen Standardnormalverteilung entspricht

random-effects Modell

Effektstärken

- allgemeinere Annahme: Effektstärken sind ähnlich, aber nicht identisch
- Zusammensetzung der Stichproben (beispielsweise hinsichtlich Alter, Bildungsgrad etc.) können zu unterschiedlichen Effektstärken führen
- diese Kovariablen sind häufig unbekannt und lassen sich schlecht approximieren
- Annahme: individuelle Effektstärken sind um einen Mittelwert μ normalverteilt

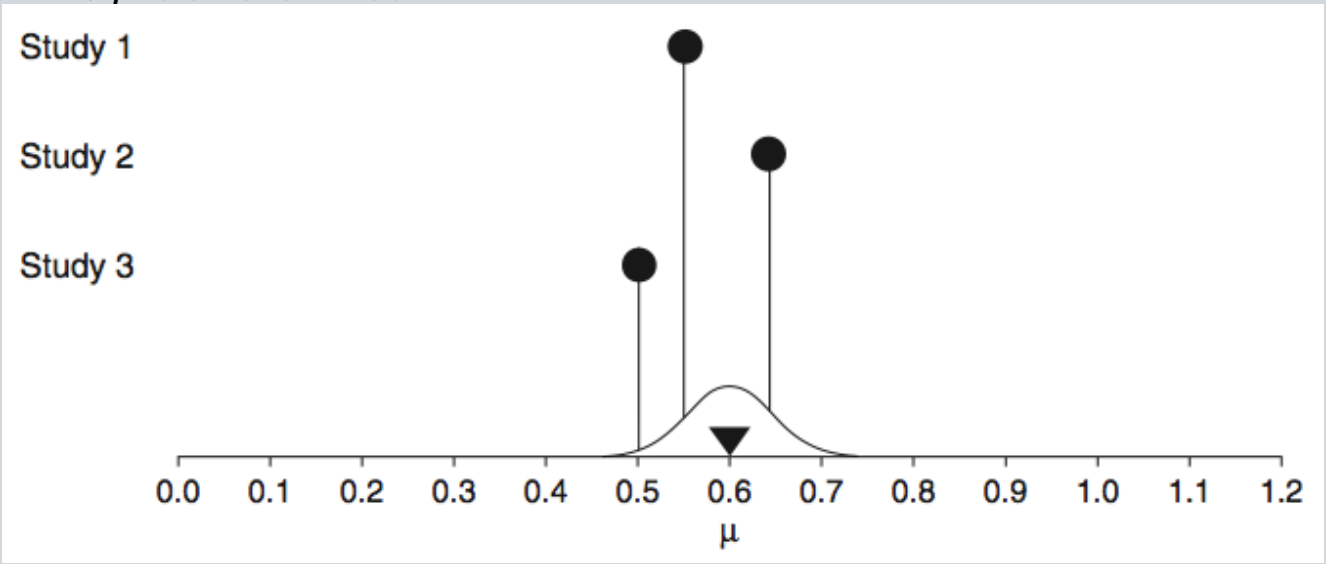
es gibt nicht einen wahren Wert sondern eine Verteilung



random-effects Modell

wahre Effektstärken

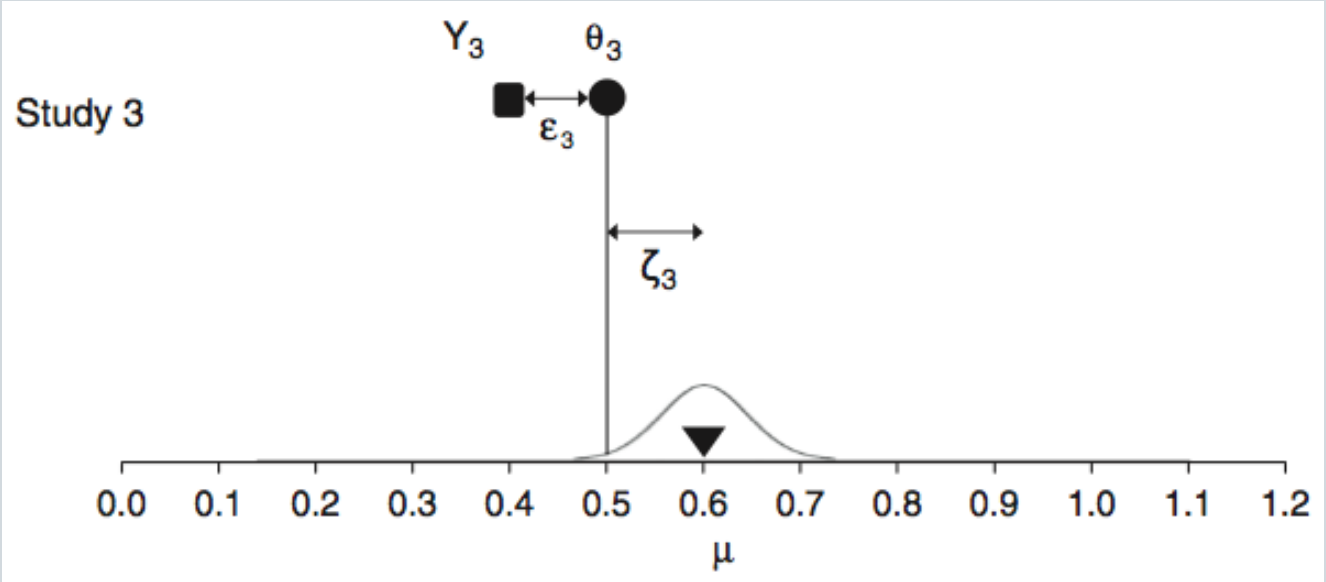
- für Studien mit unendlicher Stichprobenanzahl stellen die Effektstärken (hier: θ_1, θ_2 und θ_3) Realisierung aus der Verteilung der wahren Effekte θ dar
- die Abweichungen vom Mittelwert der wahren Effekte μ wird mit $\zeta_i = \mu - \theta_i$ bezeichnet



random-effects Modell

beobachtete Effektstärken

durch den endlichen Stichprobenumfang in jeder Studie gibt es einen zusätzlichen Stichprobenfehler ϵ_i



random-effects Modell

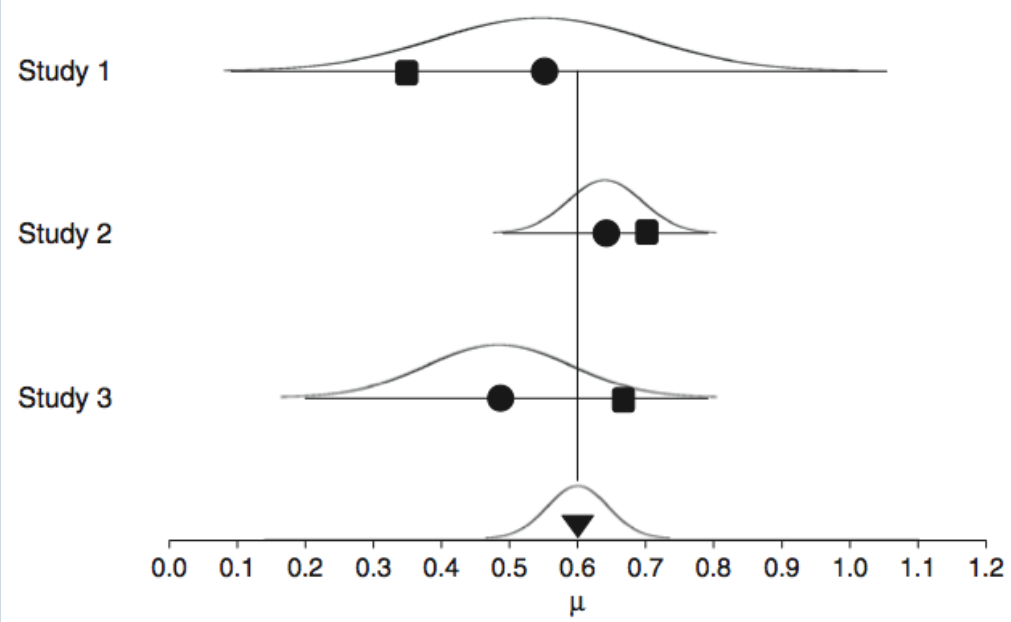
beobachtete Effektstärken

- beobachtete Effektstärke Y_i ergibt sich als Summe der Abweichungen ζ_i und ϵ_i:

$$Y_i = \mu + \zeta_i + \epsilon_i$$

↳ Mittelwert des wahren Werte

- die Verteilung der ζ_i wird durch die Varianz τ^2 beschrieben; die Varianz τ^2 ist damit für alle Studien gleich



random-effects Modell

Berechnungsvorschrift

- verwenden der beobachteten Y_i um den Mittelwert der Effektstärke μ zu schätzen
- Schätzung der Gewichte aus der Varianz der einzelnen Studien UND der Varianz τ^2 der Verteilung der Effektstärken zwischen den Studien

random-effects Modell

Varianz der Studien

- für bekannte, wahre Effektstärken θ_i und eine unendliche Anzahl von Studien könnte man die Varianz τ^2 bestimmen
- Abschätzung von τ^2 mit der Methode von DerSimonian und Laird

Q = total variation

$$T^2 = \frac{Q - df}{C}$$

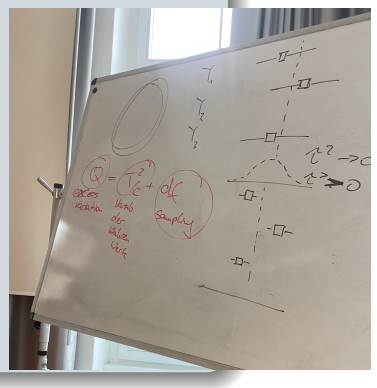
wobei

$$Q = \sum_{i=1}^{k} W_i (Y_i - M)^2 = \sum_{i=1}^{k} W_i Y_i^2 - \frac{(\sum_{i=1}^{k} W_i Y_i)^2}{\sum_{i=1}^{k} W_i}$$

$$df = k - 1$$

und

$$C = \sum_{i=1}^{k} W_i - \frac{\sum_{i=1}^{k} W_i^2}{\sum_{i=1}^{k} W_i}$$



random-effects Modell

Berechnungsvorschrift

- für die Schätzung der Gewichte sind jetzt beide Varianzquellen zu berücksichtigen

$$W_i^* = \frac{1}{V_{Y_i}^*}$$

- $V_{Y_i}^*$ entspricht der summierten Varianz

$$V_{Y_i}^* = V_{Y_i} + T^2$$

- das gewichtete Mittel M^* berechnet sich nach

$$M^* = \frac{\sum_{i=1}^{k} W_i^* Y_i}{\sum_{i=1}^{k} W_i^*}$$

random-effects Modell

Berechnungsvorschrift

- Varianz der gemittelten Effektstärke (summary effect) wird abgeschätzt durch

$$V_{M^*} = \frac{1}{\sum_{i=1}^{k} W_i^*}$$

- Standardfehler der gemittelten Effektstärke

$$SE_{M^*} = \sqrt{V_{M^*}}$$

- Grenzen des 95% Konfidenzintervalls

$$LL_{M^*} = M^* - 1.96 \cdot SE_{M^*}$$

$$UL_{M^*} = M^* + 1.96 \cdot SE_{M^*}$$

fixed-effect Modell

Rechenbeispiel

- 6 Studien sollen in einer Metaanalyse zusammengeführt werden; bias-korrigierte, standardisierte Mittelwertdifferenz (Hedges' g) als Maß der Effektstärke
- Zusammenfassung der Studien mittels fixed-effect Modell

Study	Treated			Control		
	Mean	*SD*	*n*	Mean	*SD*	*n*
Carroll	94	22	60	92	20	60
Grant	98	21	65	92	22	65
Peck	98	28	40	88	26	40
Donat	94	19	200	82	17	200
Stewart	98	21	50	88	22	45
Young	96	21	85	92	22	85

Study	Treated			Control		
	Mean	*SD*	*n*	Mean	*SD*	*n*

fixed-effect Modell

Rechenbeispiel

- Berechnung der standardisierten Differenz der Mittelwerte

$$d = \frac{\bar{X}_1 - \bar{X}_2}{S_{\text{within}}}, \; S_{\text{within}} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Swithin ≙ Spooled

- für die *Carroll* Studie ergibt sich

$$S_{\text{within}} = \sqrt{\frac{(60 - 1)22^2 + (60 - 1)20^2}{60 + 60 - 2}} = 21.0238$$

und

$$d = \frac{94 - 92}{21.0238} = 0.0951$$

fixed-effect Modell

Rechenbeispiel

- Varianz

$$V_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)} = \frac{60 + 60}{60 \cdot 60} + \frac{0.0951^2}{2(60 + 60)} = 0.0334$$

- der Korrekturfaktor J ergibt sich zu

$$J = 1 - \frac{3}{4df - 1} = 1 - \frac{3}{4 \cdot 118 - 1} = 0.9936$$

fixed-effect Modell

R. Packet : metafor "escale"

Rechenbeispiel

damit berechnet sich die bias-korrigierte, standardisierte Mittelwertdifferenz (Hedges' g) und deren Varianz

$$g = J \cdot d = 0.9936 \cdot 0.0951 = 0.0945 = Y_1$$

und

$$V_g = J^2 \cdot V_d = 0.9936^2 \cdot 0.0334 = 0.0329 = \text{Varianz}$$

das Gewicht der Studien ergibt sich aus der inversen Varianz

$$W = \frac{1}{V_Y} = \frac{1}{0.0329} = 30.3515$$

fixed-effect Modell

Rechenbeispiel

Study	Effect size	Variance Within	Weight	Calculated quantities		
	Y	V_Y	W	WY	WY^2	W^2
Carroll	0.095	0.033	30.352	2.869	0.271	921.214
Grant	0.277	0.031	32.568	9.033	2.505	1060.682
Peck	0.367	0.050	20.048	7.349	2.694	401.931
Donat	0.664	0.011	95.111	63.190	41.983	9046.013
Stewart	0.462	0.043	23.439	10.824	4.999	549.370
Young	0.185	0.023	42.698	7.906	1.464	1823.115
Sum			244.215	101.171	53.915	13802.325

Study	Effect size	Variance Within	Weight	Calculated quantities		
	Y	V_Y	W	WY	WY^2	W^2
Carroll	0.095	0.033	30.352	2.869	0.271	921.214
Grant	0.277	0.031	32.568	9.033	2.505	1060.682
Peck	0.367	0.050	20.048	7.349	2.694	401.931

fixed-effect Modell

Rechenbeispiel

- das gewichtete Mittel ergibt sich aus

$$M = \frac{\sum_{i=1}^{k} W_i Y_i}{\sum_{i=1}^{k} W_i}$$

und die Varianz der gemittelten Effektstärke

$$V_M = \frac{1}{\sum_{i=1}^{k} W_i}$$

- im konkreten Fall:

$$M = \frac{101.171}{244.215} = 0.4143$$

und

$$V_M = \frac{1}{244.215} = 0.0041$$

fixed-effect Modell

Rechenbeispiel

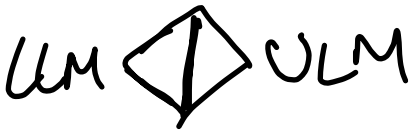
- aus der Varianz errechnet sich der Standardfehler

$$SE_M = \sqrt{V_M} = \sqrt{0.0041} = 0.0640$$

- für die Grenzen des 95%-Konfidenzintervalles ergibt sich

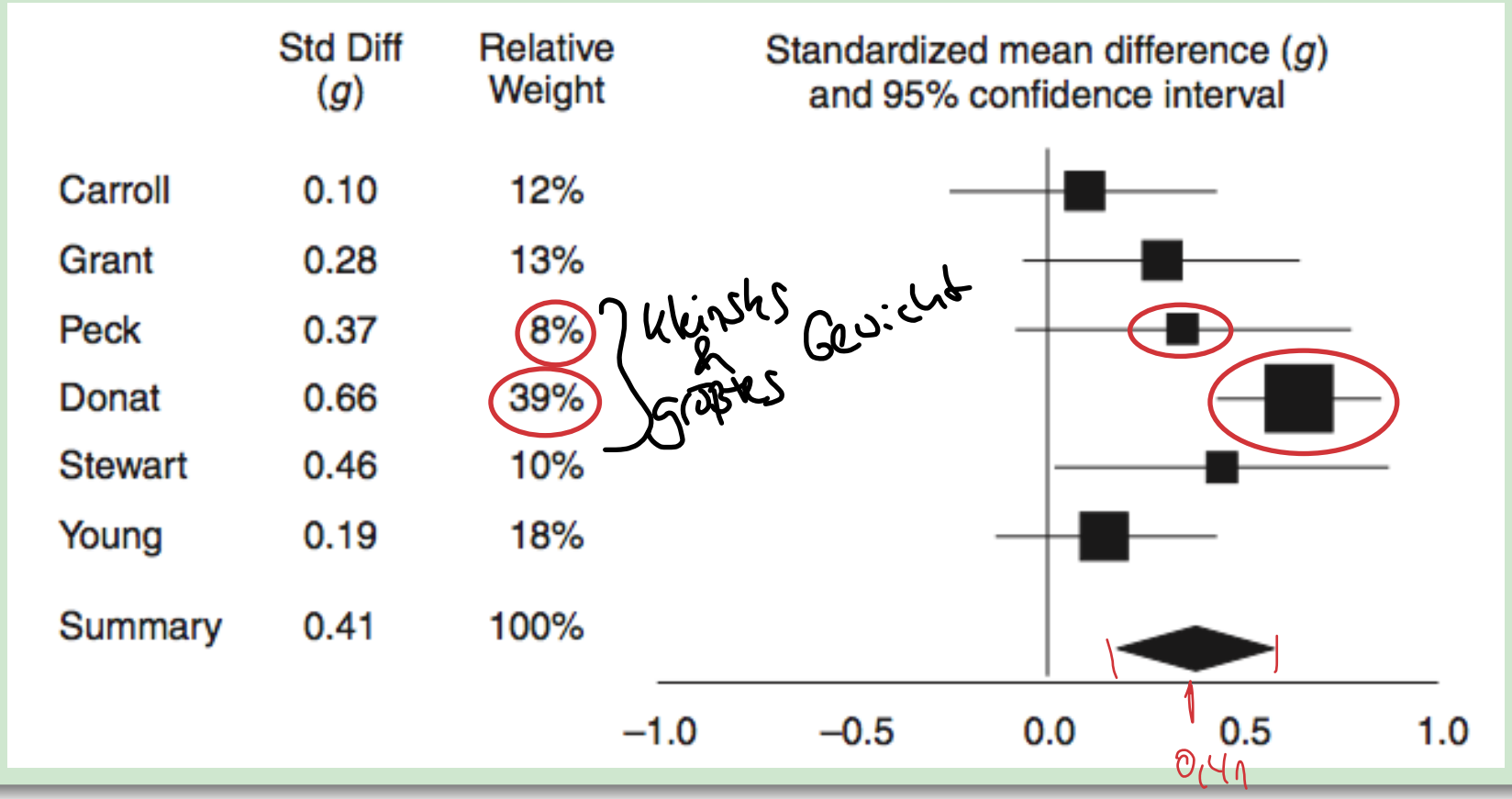
$$LL_M = M - 1.96 \cdot SE_M = 0.4143 - 1.96 \cdot 0.0640 = 0.2889$$

$$UL_M = M + 1.96 \cdot SE_M = 0.4143 + 1.96 \cdot 0.0640 = 0.5397$$



fixed-effect Modell

Rechenbeispiel



random-effects Modell

Rechenbeispiel

- identische Analyse mittels eines random-effects Modells
- Schätzung der Varianz τ^2 der wahren standardisierten Mittelwerte mittels der *DerSimonian und Laird* Methode:

$$T^2 = \frac{Q - df}{C}$$

random-effects Modell

Rechenbeispiel

- Berechnung von Q, df und C:

$$Q = \sum_{i=1}^{k} W_i Y_i^2 - \frac{(\sum_{i=1}^{k} W_i Y_i)^2}{\sum_{i=1}^{k} W_i} = 53.915 - \frac{101.171^2}{244.215} = 12.0033$$

$$df = k - 1 = 6 - 1 = 5$$

$$C = \sum_{i=1}^{k} W_i - \frac{\sum_{i=1}^{k} W_i^2}{\sum_{i=1}^{k} W_i} = 244.215 - \frac{13802.325}{244.215} = 187.698$$

- daraus folgt

$$T^2 = \frac{Q - df}{C} = \frac{12.0033 - 5}{187.698} = 0.0373$$

random-effects Modell

Rechenbeispiel

Study	Effect size Y	Variance Within V_Y	Variance Between T^2	Variance Total $V_Y + T^2$	Weight W^*	Calculated quantities W^*Y
Carroll	0.095	0.033	0.037	0.070	14.233	1.345
Grant	0.277	0.031	0.037	0.068	14.702	4.078
Peck	0.367	0.050	0.037	0.087	11.469	4.204
Donat	0.664	0.011	0.037	0.048	20.909	13.892
Stewart	0.462	0.043	0.037	0.080	12.504	5.774
Young	0.185	0.023	0.037	0.061	16.466	3.049
Sum					90.284	32.342

Study	Effect size Y	Variance Within V_Y	Variance Between T^2	Variance Total $V_Y + T^2$	Weight W^*	Calculated quantities W^*Y
Carroll	0.095	0.033	0.037	0.070	14.233	1.345
Grant	0.277	0.031	0.037	0.068	14.702	4.078
Peck	0.367	0.050	0.037	0.087	11.469	4.204
Donat	0.664	0.011	0.037	0.048	20.909	13.892
Stewart	0.462	0.043	0.037	0.080	12.504	5.774

random-effects Modell

Rechenbeispiel

- damit berechnet sich das gewichtete Mittel und die Varianz:

$$M^* = \frac{\sum_{i=1}^{k} W_i^* Y_i}{\sum_{i=1}^{k} W_i^*} = \frac{32.342}{90.284} = 0.3582$$

$$V_{M^*} = \frac{1}{\sum_{i=1}^{k} W_i^*} = \frac{1}{90.284} = 0.0111$$

- Standardfehler der gemittelten Effektstärke

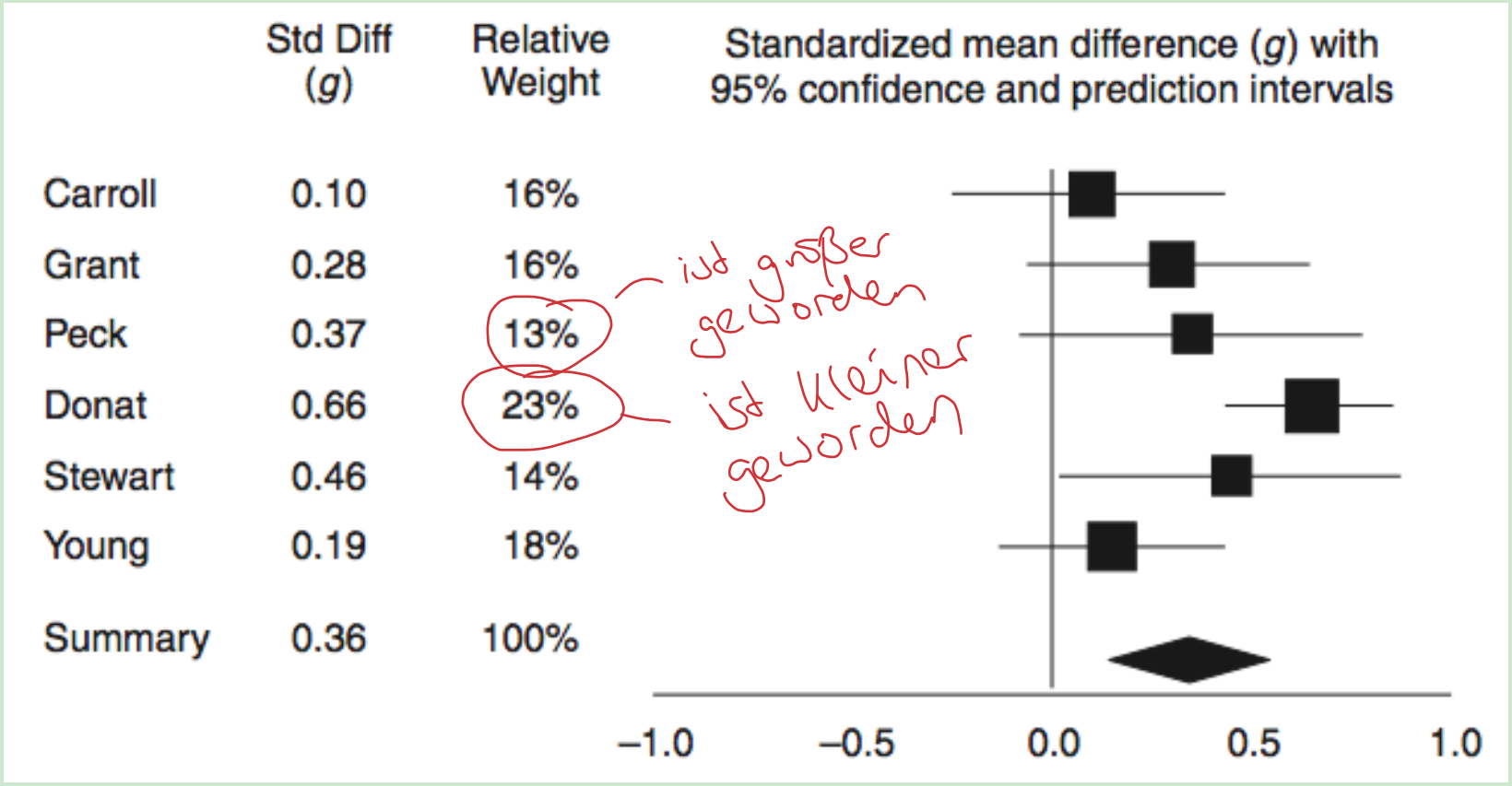
$$SE_{M^*} = \sqrt{V_{M^*}} = \sqrt{0.0111} = 0.1052$$

- Grenzen des 95% Konfidenzintervalls

$$LL_{M^*} = M^* - 1.96 \cdot SE_{M^*} = 0.3582 - 1.96 \cdot 0.1052 = 0.1520$$
$$UL_{M^*} = M^* + 1.96 \cdot SE_{M^*} = 0.3582 + 1.96 \cdot 0.1052 = 0.5645$$

random-effects Modell

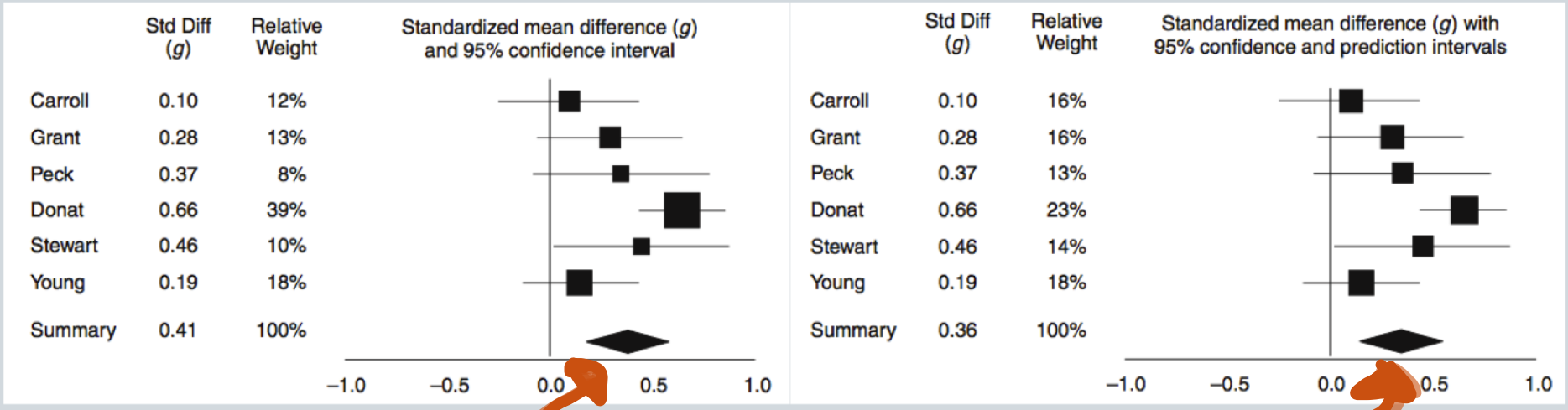
Rechenbeispiel



fixed-effect vs. random-effects Modell

Summary Effect

- Bedeutung des Summary Effect in den beiden berechneten Modellen?



- geschätzte Effektstärke vs. geschätzter Mittelwert der Effektstärke

fixed-effect vs. random-effects Modell

Summary Effect für das fixed-effect Modell

- identische "wahre" Effektstärke θ für alle Studien
- lediglich der Fehler durch die Auswahl der Stichprobe führt zur Variabilität der Resultate der einzelnen Studien $\rightarrow$ kleinere Studien können nahezu vernachlässigt werden (geringeres Gewicht)
- große Studien beeinflussen den Summary Effect stärker

Summary Effect für das random-effects Modell

- Schätzung des Mittelwertes einer Verteilung von Effekten
- jede Studie besitzt Informationsgehalt bezüglich dieser Verteilung
- kleine Studien können zwar ungenauer sein, sind aber dennoch wichtig $\rightarrow$ höheres Gewicht als im fixed-effect Modell
- Gewichte sind insgesamt balancierter als im fixed-effect Modell

fixed-effect vs. random-effects Modell

Konfidenzintervalle

- für random-effects Modelle gibt es jeweils eine zusätzliche Quelle der Unsicherheit (*between-study variance*)
- $\rightarrow$ größerer Standardfehler und CI für random-effects Modell im Vergleich zum fixed-effect Modell (gegeben $T^2 > 0$)

Standardfehler

- k Studien mit jeweils Stichprobenumfang n und Standardabweichung σ
- Standardfehler im fixed-effect Modell $SE_M = \sqrt{\frac{\sigma^2}{k \cdot n}}$
- Standardfehler im random-effects Modell $SE_M = \sqrt{\frac{\sigma^2}{k \cdot n} + \frac{\tau^2}{k}}$

fixed-effect vs. random-effects Modell

Auswahl des richtigen Modells

- die Modellauswahl MUSS an unsere Erwartungen bezüglich das Vorliegens / Nicht-vorliegens einer gemeinsamen Effektstärke gekoppelt sein
- fixed-effect Modell:
  - funktionelle Identität der Studien
  - Aussage für die untersuchte Population
- random-effects Modell:
  - unabhängige Studien verschiedener Wissenschaftler
  - Annahme einer gemeinsamen Effektstärke ist nicht gegeben
  - generalisierbar für andere Populationen
  - CAVE: für geringe Anzahl von Studien $k \rightarrow$ schlechte Schätzung von τ^2

Modellauswahl sollte nicht an einen Homogenitätstest gekoppelt werden

Effektstärken: Binärdaten

Vierfeldertafeln

- Vierfeldertafeln sind eine häufige Darstellung für unterschiedliche Ereignisse in zwei Gruppen
- Berechnung des Relative Risiken oder des Odds Ratios
- Nomenklatur

	Ereignis	kein Ereignis	N
Behandelt	A	B	n_1
Kontrolle	C	D	n_2

Beispiel: Vierfeldertafel

	verstorben	lebend	N
Behandelt	5	95	100
Kontrolle	10	90	100

Effektstärken: Binärdaten

Relatives Risiko

- das Verhältnis zweier Risiken wird als Relatives Risiko bezeichnet
- Relative Risiken sind intuitiv einfach interpretierbar
- das Relative Risiko RR wird berechnet

$$RR = \frac{A/n_1}{C/n_2}$$

- für das logarithmische Relative Risiko $LogRR$ gilt

$$LogRR = \ln(RR)$$

bzw.

$$RR = exp(LogRR)$$

Effektstärken: Binärdaten

Beispiel: Vierfeldertafel

- für unser Beispiel: Relatives Risiko in der Behandlungsgruppe zu versterben:

$$RR = \frac{5/100}{10/100} = 0.5$$

- logarithmisches Relatives Risiko

$$LogRR = \ln(0.5) = -0.6932$$

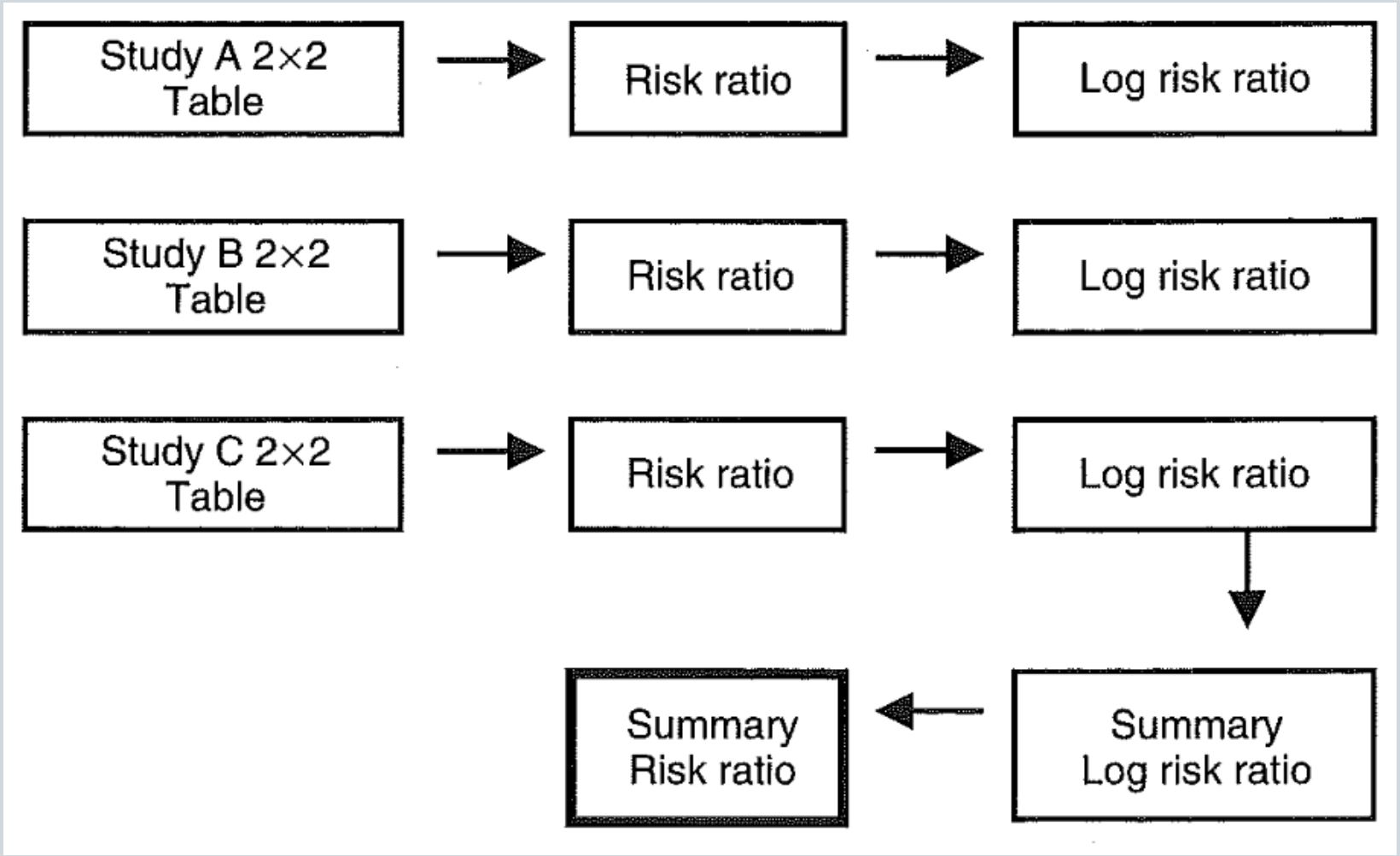
Effektstärken: Binärdaten

Logarithmische Transformation?

- logarithmische Transformation sichert die Symmetrie der Analyse
- Studie (I) beschreibt ein zweifach erhöhtes Risiko für Gruppe A $RR_I = 2$ und Studie (II) beschreibt eine zweifache Erhöhung in Gruppe B $RR_{\mathrm{II}} = 0.5$
- bei gleichem Gewicht (Gruppengröße) sollten sich die Effekte aufheben $\rightarrow RR_{\mathrm{comb}} = 1.0$
- auf der Skala der relativen Risiken wäre der Mittelwert bei $RR_{\mathrm{comb}} = 1.25$
- für die logarithmischen relativen Risiken ergibt sich $LogRR_I = -0.693$ und $LogRR_{\mathrm{II}} = 0.693$
- daraus folgt $LogRR_{\mathrm{comb}} = 0.0 \rightarrow RR_{\mathrm{comb}} = 1.0$

Effektstärken: Binärdaten

Logarithmische Transformation?



Effektstärken: Binärdaten

Varianzabschätzung Relatives Risiko

- zur Berechnung eines Summary Effect werden Varianzmaße nicht in der originalen Metrik des Relativen Risikos berechnet, sondern in der logarithmischen Transformation
- Varianzschätzung

$$Var_{\mathrm{LogRR}} = \frac{1}{A} - \frac{1}{n_1} + \frac{1}{C} - \frac{1}{n_2}$$

- approximierter Standardfehler

$$SE_{\mathrm{LogRR}} = \sqrt{Var_{\mathrm{LogRR}}}$$

Effektstärken: Binärdaten

Odds ratio

- als Odds ratio wird ein Chancenverhältnis bezeichnet
- das Odds ratio OR wird berechnet

$$OR = \frac{A/B}{C/D} = \frac{AD}{BC}$$

- für das logarithmische Odds ratio $LogOR$ gilt

$$LogOR = \ln(OR)\ ;\ OR = exp(LogOR)$$

- das Odds ratio ist als Maß weniger intutitiv als das Relative Risiko, hat aber gute statistische Eigenschaften für Metaanalysen
- für kleine Risiken nähert sich das Odds ratio dem Relativen Risiko an

Effektstärken: Binärdaten

Beispiel: Vierfeldertafel

- für unser Beispiel: Chance ("Wahrscheinlichkeit") in der Behandlungsgruppe zu versterben ist 5/95, in der Kontrollgruppe 10/90
- damit ergibt sich das Odds ratio

$$OR = \frac{5/95}{10/90} = 0.4737$$

- logarithmisches Odds ratio

$$LogOR = \ln(0.4737) = -0.7472$$

Effektstärken: Binärdaten

Varianzabschätzung Odds Ratio

- analog zum Relativen Risiko werden Varianzmaße in der logarithmischen Metrik bevorzugt
- Varianzschätzung

$$Var_{\mathrm{LogOR}} = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}$$

- approximierter Standardfehler

$$SE_{\mathrm{LogOR}} = \sqrt{Var_{\mathrm{LogOR}}}$$

Binärdaten

Rechenbeispiel: Odds Ratio

6 Studien mit binären Daten

Study	Treated			Control		
	Events	Non-events	*n*	Events	Non-events	*n*
Saint	12	53	65	16	49	65
Kelly	8	32	40	10	30	40
Pilbeam	14	66	80	19	61	80
Lane	25	375	400	80	320	400
Wright	8	32	40	11	29	40
Day	16	49	65	18	47	65

Rechenbeispiel: Odds Ratio

- Berechnung des Odds ratio

$$OR = \frac{AD}{BC} = \frac{12 \cdot 49}{53 \cdot 16} = 0.6934$$

- für das logarithmische Odds ratio ergibt sich

$$LogOR = \ln(OR) = \ln(0.6934) = -0.3662$$

- Varianzschätzung

$$Var_{\mathrm{LogOR}} = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D} = \frac{1}{12} + \frac{1}{53} + \frac{1}{16} + \frac{1}{49} = 0.1851$$

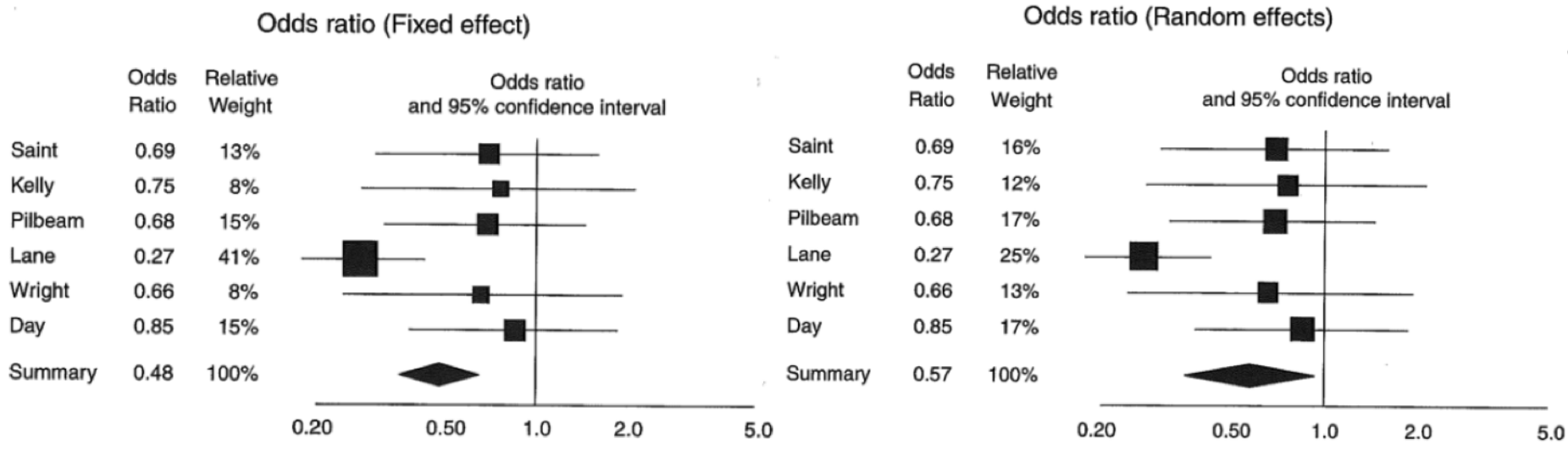
Binärdaten

fixed-effect & random-effects Berechnungen

Study	Effect size	Variance Within	Weight	Calculated quantities			Variance Between	Variance Total	Weight	Calculated quantities
	Y	V_Y	W	WY	WY^2	W^2	T^2	V_Y+T^2	W^*	W^*Y
Saint	−0.366	0.185	5.402	−1.978	0.724	29.184	0.173	0.358	2.793	−1.023
Kelly	−0.288	0.290	3.453	−0.993	0.286	11.925	0.173	0.462	2.162	−0.622
Pilbeam	−0.384	0.156	6.427	−2.469	0.948	41.300	0.173	0.329	3.044	−1.169
Lane	−1.322	0.058	17.155	−22.675	29.971	294.298	0.173	0.231	4.325	−5.717
Wright	−0.417	0.282	3.551	−1.480	0.617	12.607	0.173	0.455	2.200	−0.917
Day	−0.159	0.160	6.260	−0.998	0.159	39.190	0.173	0.333	3.006	−0.479
Sum			42.248	−30.594	32.705	428.503			17.531	−9.928

Binärdaten

Forest plots



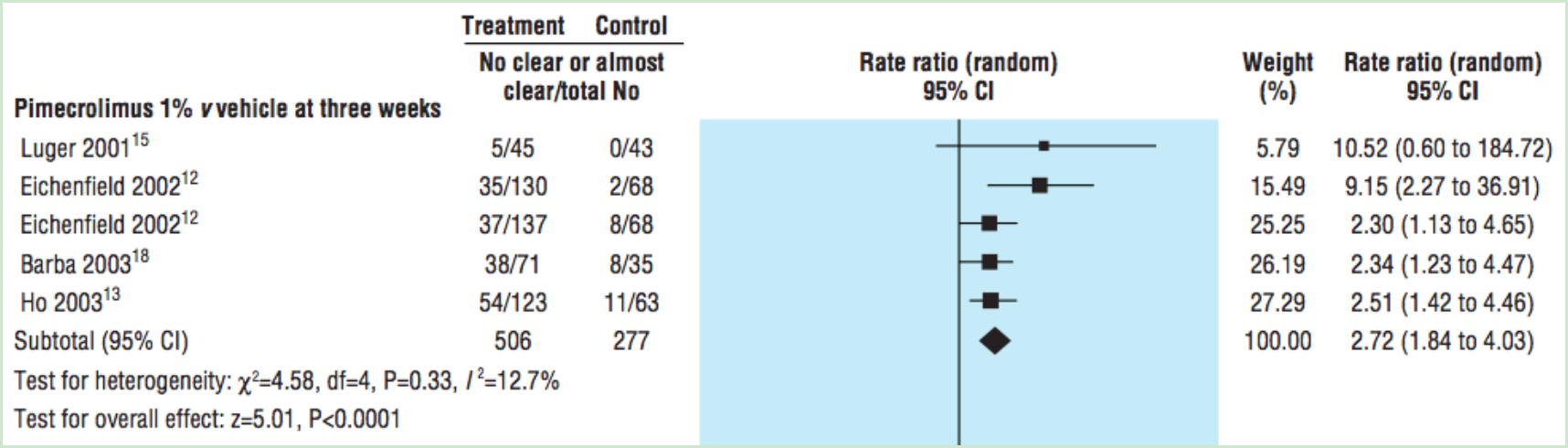
Vorlesung 3

- Heterogenität

Heterogenität

Vorbetrachtungen

- Definition der Metaanalyse als "eine an den Kriterien empirischer Forschung orientierte Methode zur quantitativen Integration der Ergebnisse empirischer Untersuchungen sowie zur **Analyse der Variabilität** dieser Ergebnisse" (Drinkmann (1990))
- Metaanalysen gehen über die Schätzung eines Summary Effects hinaus $\rightarrow$ Erklärung der "Muster" von Effekten



Pimecrolimus 1% *v* vehicle at three weeks	Treatment No clear or almost clear/total	Control No	Rate ratio (random) 95% CI	Weight (%)	Rate ratio (random) 95% CI
Luger 2001[15]	5/45	0/43		5.79	10.52 (0.60 to 184.72)
Eichenfield 2002[12]	35/130	2/68		15.49	9.15 (2.27 to 36.91)
Eichenfield 2002[12]	37/137	8/68		25.25	2.30 (1.13 to 4.65)
Barba 2003[18]	38/71	8/35		26.19	2.34 (1.23 to 4.47)
Ho 2003[13]	54/123	11/63		27.29	2.51 (1.42 to 4.46)
Subtotal (95% CI)	506	277		100.00	2.72 (1.84 to 4.03)
Test for heterogeneity: χ^2=4.58, df=4, P=0.33, I^2=12.7%					
Test for overall effect: z=5.01, P<0.0001					

Heterogenität von Metaanalysen und Publikationsbias

Aufbau der Vorlesung

- Q-statistik
- abgeleitete Maße der Heterogenität
- exemplarische Berechnung für ein Beispiel (standardisierte Differenz von Mittelwerten)
- Zusammenfassung Heterogenität

Heterogenität

Vorbetrachtungen

- Variabilität der geschätzten Effektstärken ist schwer interpretierbar, da es eine zugrundeliegende Variabilität der wahren Effektstärken gibt, die zusätzlich mit einem Stichprobenfehler überlagert ist
- $\rightarrow$ Identifikation der *wahren* Varianz der Effektstärken
- $\rightarrow$ Maßzahlen zur Beschreibung der Dispersion
  - Q-statistik
  - *between-study variance* T^2
  - Verhältnis der *wahren* Heterogenität zur gesamten, gemessenen Variabilität (I^2)

$$Y_i = \mu + \theta_i + \varepsilon_i$$

θ_i: Variabilität der wahren Werte

ε_i: Stichproben-fehler

Heterogenität

Fragen

- Gibt es Hinweise für Heterogenität der wahren Effektstärken?
- Wie groß ist die Varianz der wahren Effekte?
- Welche Implikationen hat diese Varianz?

Nomenklatur

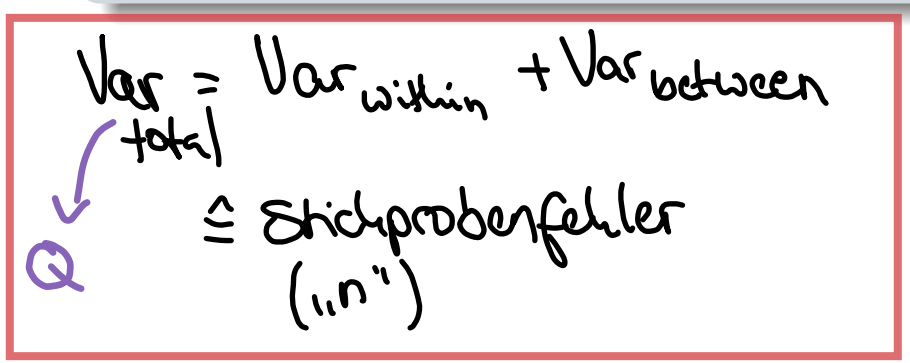
- *wahre Effektstärke* beschreibt die Effektstärke in der unterliegenden Population (unendliche Stichprobengröße $\rightarrow$ kein Stichprobenfehler)
- *gemessene Effektstärke* ist der ermittelte Wert einer Studie mit endlicher Stichprobengröße
- *Heterogenität* beschreibt die Heterogenität der wahren Effektstärken, *Variabilität* und *Dispersion* werden allgemeiner verwendet, um Unterschiede in (wahren und gemessenen) Effektstärken zu beschreiben

quantitative Heterogenität

Vorgehen

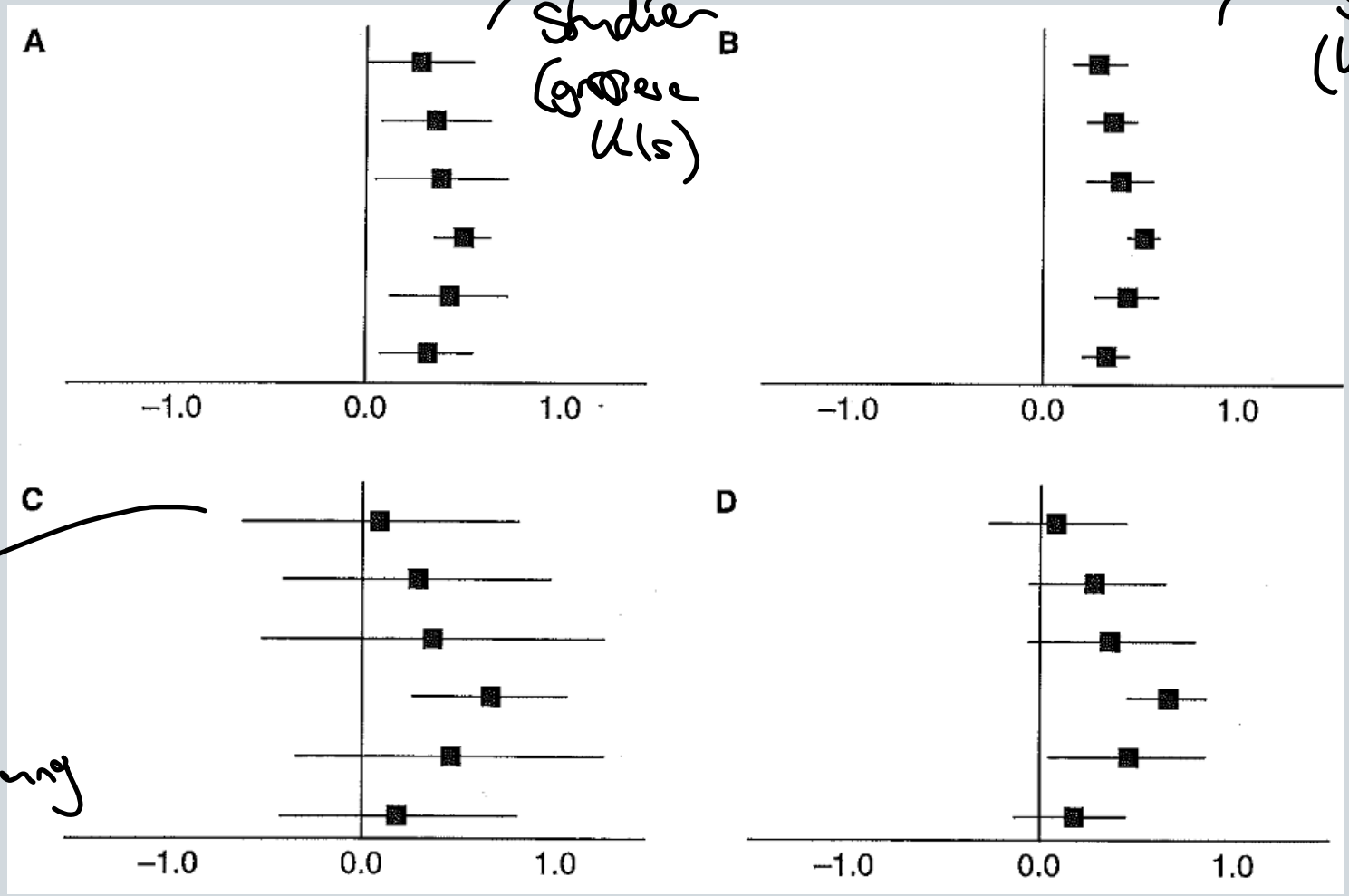
Aufteilung der gemessenen Variabilität in *Heterogenität der Effektstärken* und *Stichprobenfehler* (*within-study variance*)

1. Berechnungen der absoluten Variabilität zwischen allen Studien
2. Abschätzung der Variabilität, die zu erwarten wäre, wenn wir gleiche *wahre* Effekte für alle Studien annehmen würden
3. *excess variation* (Differenz zwischen den beiden Variabilitäten) ist ein Schätzer für die Heterogenität (d.h. die *wahre* Heterogenität der Effektstärken)



quantitative Heterogenität

Dispersion vs. Stichprobenfehler



quantitative Heterogenität

$$Q = \sum w_i \times (Y_i - M)^2$$

$$w_i = \frac{1}{V_{Yi}} = \frac{1}{S_{Yi}^2}$$

Berechnung von Q

- Berechnung von Q als gewichtete Summe der quadratischen Abweichungen der Effektstärken Y_i vom Summary Effect M:

$$\text{Varianz} \quad Q = \sum_{i=1}^{k} W_i (Y_i - M)^2$$

- Q ist ein standardisiertes Maß; es ist daher unabhängig von der Metrik der Effektstärke
- äquivalente Formulierung

$$Q = \sum_{i=1}^{k} W_i Y_i^2 - \frac{(\sum_{i=1}^{k} W_i Y_i)^2}{\sum_{i=1}^{k} W_i} \quad .$$

quantitative Heterogenität

Erwartungswert von Q

- Abschätzung des Erwartungswertes für Q unter der Annahme, dass alle Studien eine identische Effektstärke aufweisen (Variation der Effektstärken wird nur durch die Stichprobenfehler bestimmt)
- da Q ein standardisiertes Maß ist, ist der Erwartungswert unabhängig von der Metrik der Effektstärke und entspricht der Anzahl der Freiheitsgrade *df*

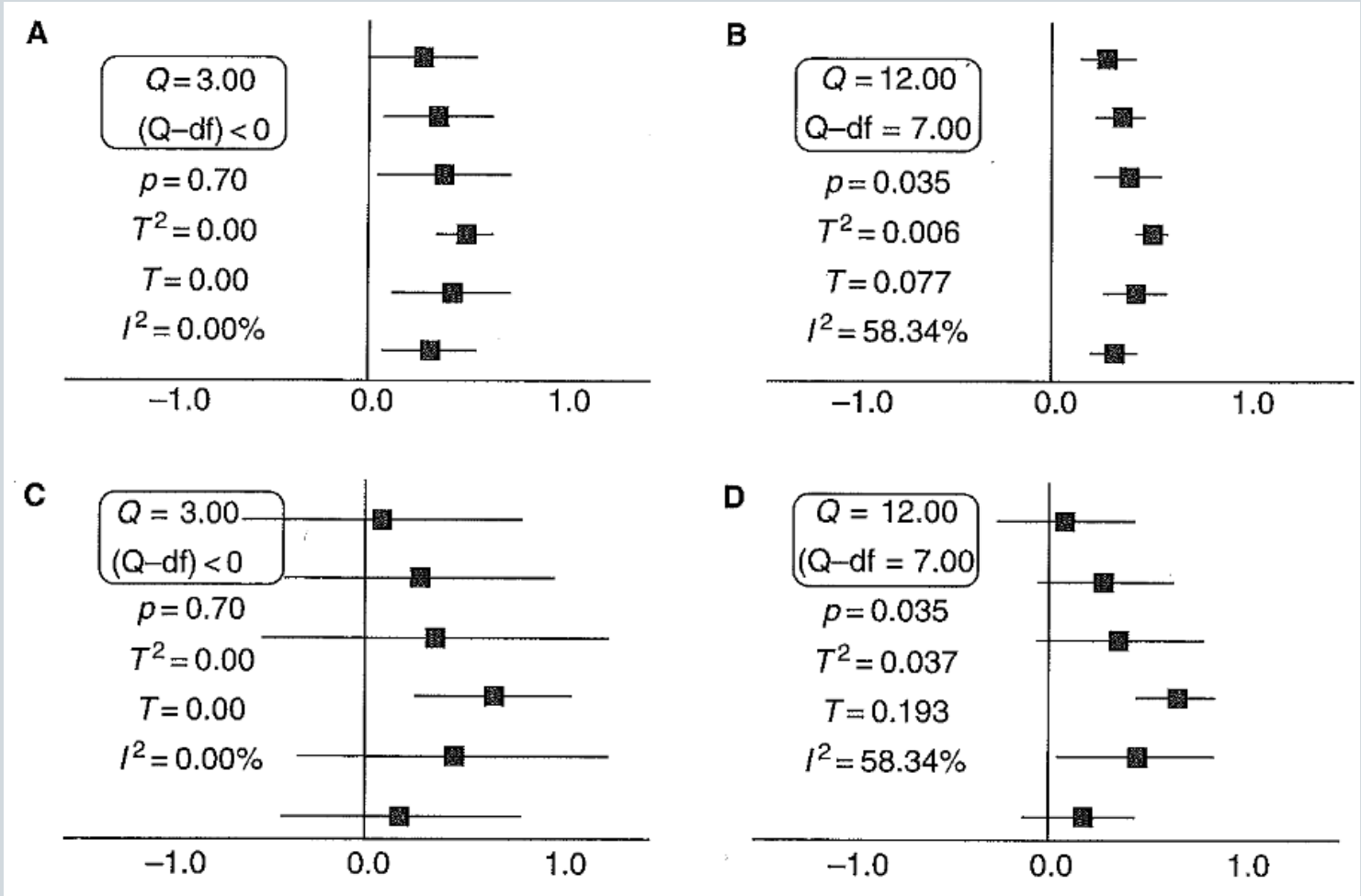
$$df = k - 1$$

Abschätzung der *excess variation*

- die Differenz $Q - df$ beschriebt die *excess variation*
- entspricht dem Anteil der Variabilität, der durch die Heterogenität (d.h. die zugrunde liegende Verteilung der wahren Effektstärken) verursacht wird

quantitative Heterogenität

Q vs. df



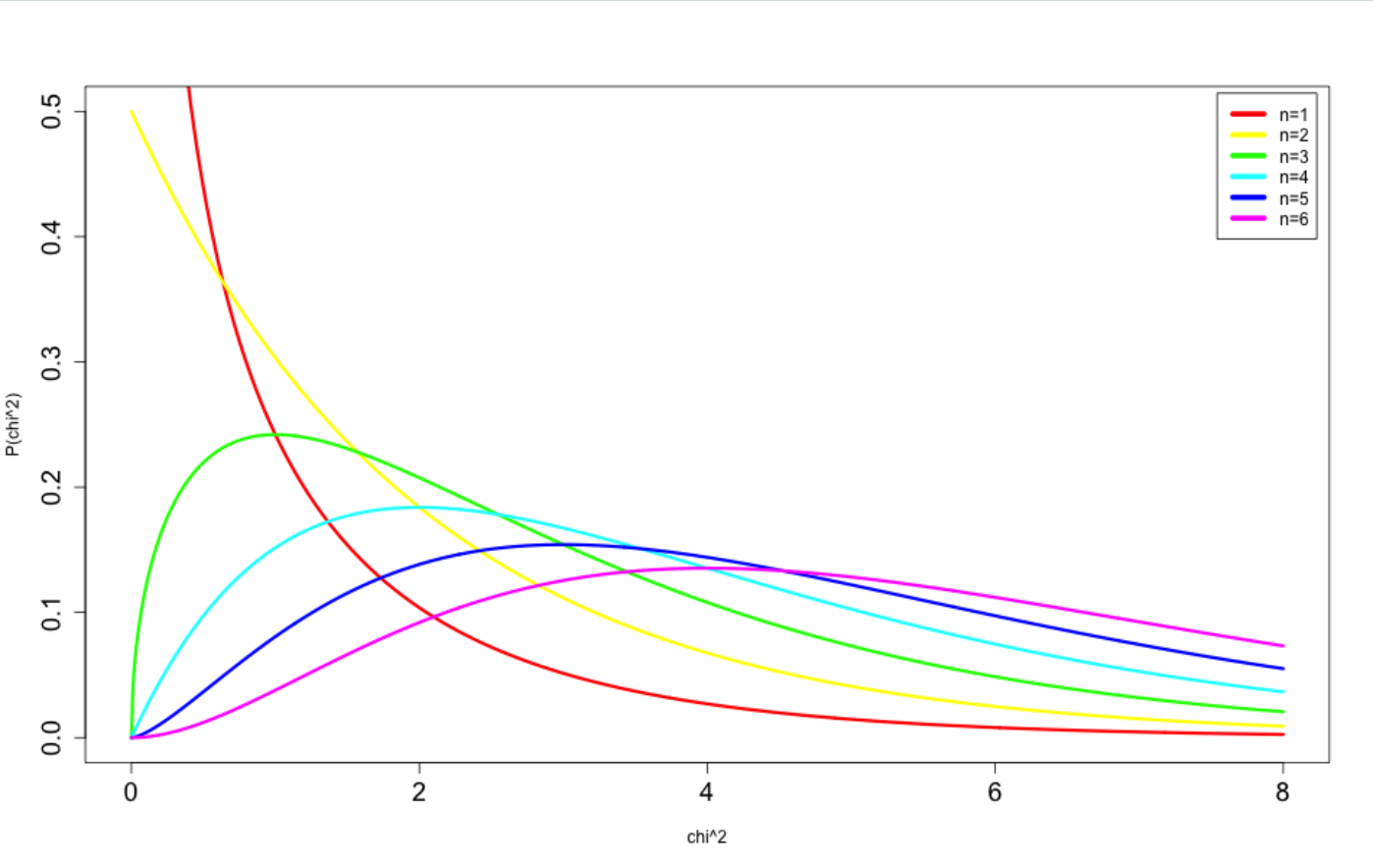
quantitative Heterogenität

Statistischer Test auf Homogenität

- Frage nach der *Signifikanz* der Heterogenität
- formal wird die Null-hypothese aufgestellt, dass alle Studien (nur) einen gemeinsamen, wahren Effekt aufweisen
- unter der Null-hypothese ist Q χ^2-verteilt, wobei die Freiheitsgrade $df = k - 1$ entsprechen
- Ermittlung eines p-Wertes für jedes (Q, df) - entspricht der Wahrscheinlichkeit, ein solches oder ein extremeres Ereignis unter der Null-hypothese zu beobachten

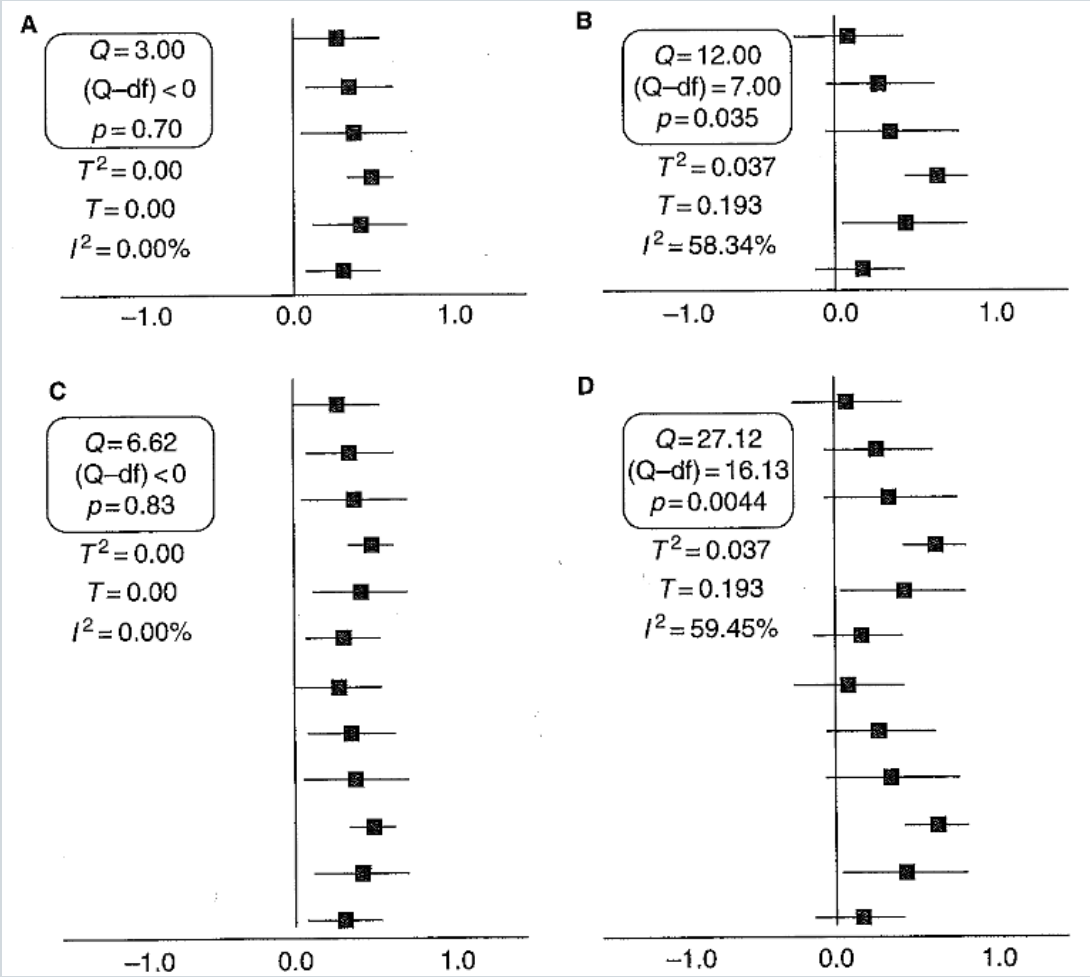
quantitative Heterogenität

Einschub: χ^2-Verteilung



quantitative Heterogenität

Q und Studienanzahl



quantitative Heterogenität

p-Wert

- ein nicht-signifikanter p-Wert ist kein Beweis, dass die Effektstärken ähnlich sind; das könnte beispielsweise auch an geringer Studienanzahl liegen
- p-Werte adressieren die Signifikanz der Resultate, nicht die Stärke der Effekte

→ Die Excess Variation ist wichtiges als der p-Wert
↳ v.a. bei Studien mit kleinen n

Einschub: statistische Tests in Metaanalysen

Signifikantes Ergebnis für die mittlere Effektstärke

- Null-hypothese: kein (Behandlungs-)effekt ($M = 0$ bzw. $M = 1$ für RR und OR)
- p-Wert für M vermittelt eine Eindruck, ob eine solche mittlere Effektstärke wahrscheinlich ist oder nicht

Signifikantes Ergebnis für die beobachtete Heterogenität

- Null-hypothese: keine Heterogenität, die über den zu erwartenden Stichprobenfehler (bei einem gemeinsamen, wahren Effekt) hinaus geht
- p-Wert für Q vermittelt eine Eindruck, ob die beobachtete Heterogenität unter der Annahme eines gemeinsamen, wahren Effektes für alle Studien erwartbar ist oder nicht

quantitative Heterogenität

Schätzung der Varianz der wahren Effekte

- τ^2 bezeichnet die Varianz der wahren Effektstärken
- τ^2 ist bestimmbar für eine unendliche Anzahl von Studien mit jeweils unendlicher Stichprobengröße (d.h. es wird jeweils der wahre Effekt θ_i in jeder Studie bestimmt)
- Abschätzung durch T^2

$$T^2 = \frac{Q - df}{C}$$

quantitative Heterogenität

Schätzung der Varianz der wahren Effekte

- die Konstante C entspricht einer Normierung, die die *excess variation* (a) in die ursprüngliche Metrik ("quadratisch") zurück transformiert und (b) einen Mittelwert bildet

$$C = \sum_{i=1}^{k} W_i - \frac{\sum_{i=1}^{k} W_i^2}{\sum_{i=1}^{k} W_i}$$

- für $Q < df$ wird T^2 negativ und wird auf $T^2 = 0$ gesetzt

↳ Variabilität wird nur durch den Stichprobenfehler erklärt

quantitative Heterogenität

Standardabweichung der wahren Effekte

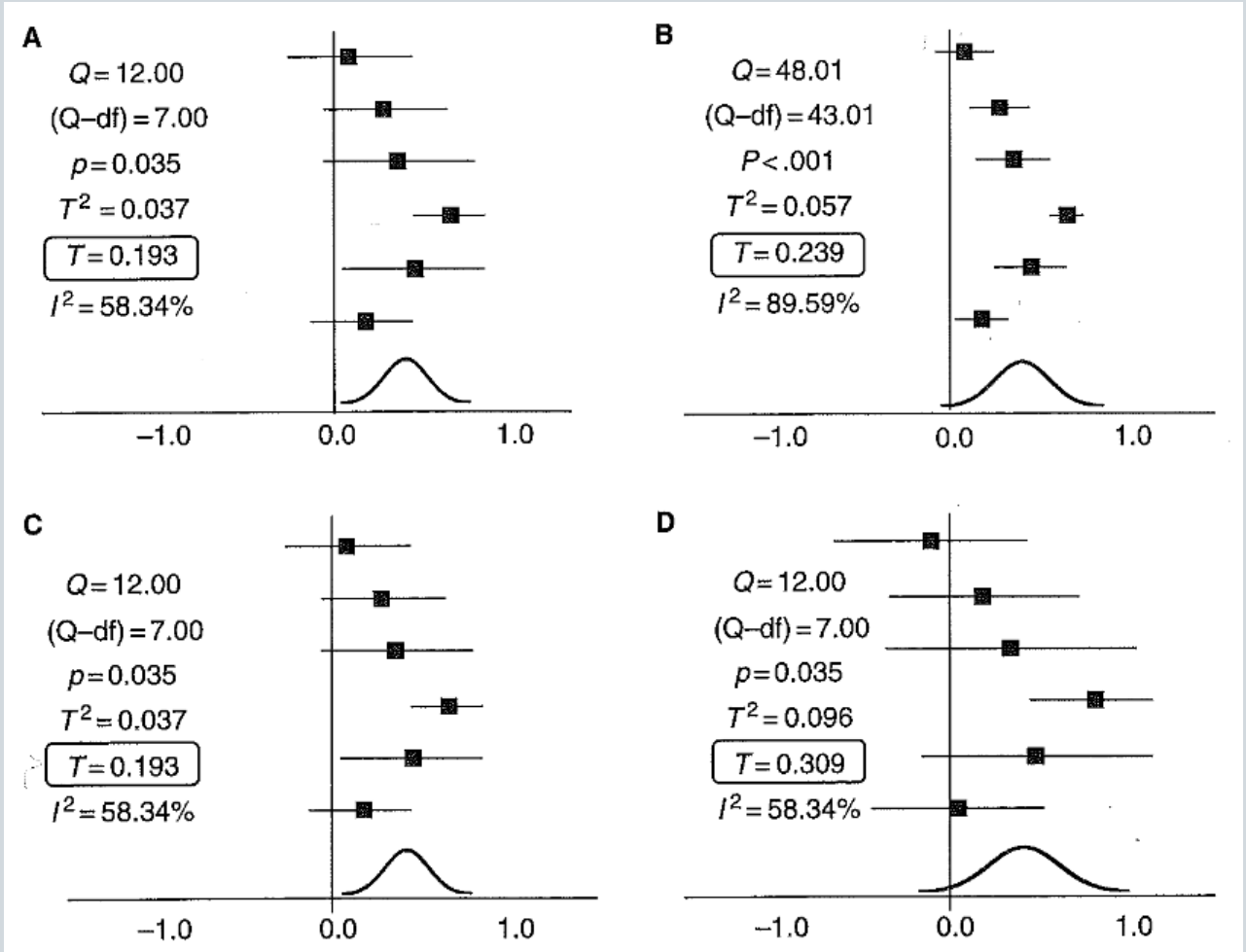
- τ bezeichnet die Standardabweichung der wahren Effektstärken, und T ist der korrespondierende Schätzer

$$T = \sqrt{T^2}$$

- T ist in der gleichen Metrik wie die Effektstärke
- ein hinreichend genauer Schätzer für T erlaubt es, die Verteilung der wahren Effektstärken anzugeben (bzw. die entsprechenden Konfidenzintervalle)

quantitative Heterogenität

Abhängigkeit von T



quantitative Heterogenität

I^2 Statistik

- T^2 und T sind absolute Maße, welche die Abweichung in der Metrik der Effektstärken quantifizieren
- $\rightarrow$ Beschreibung der Heterogenität auf einer unabhängigen Skala ist wünschenswert: *Welcher Anteil der gemessenen Variabilität lässt sich auf die Variabilität der Effektstärken zurückführen?*
- I^2 ist ein relatives Maß, dass diesen Anteil der *excess variation* an der gesamten gemessenen Dispersion beschreibt:

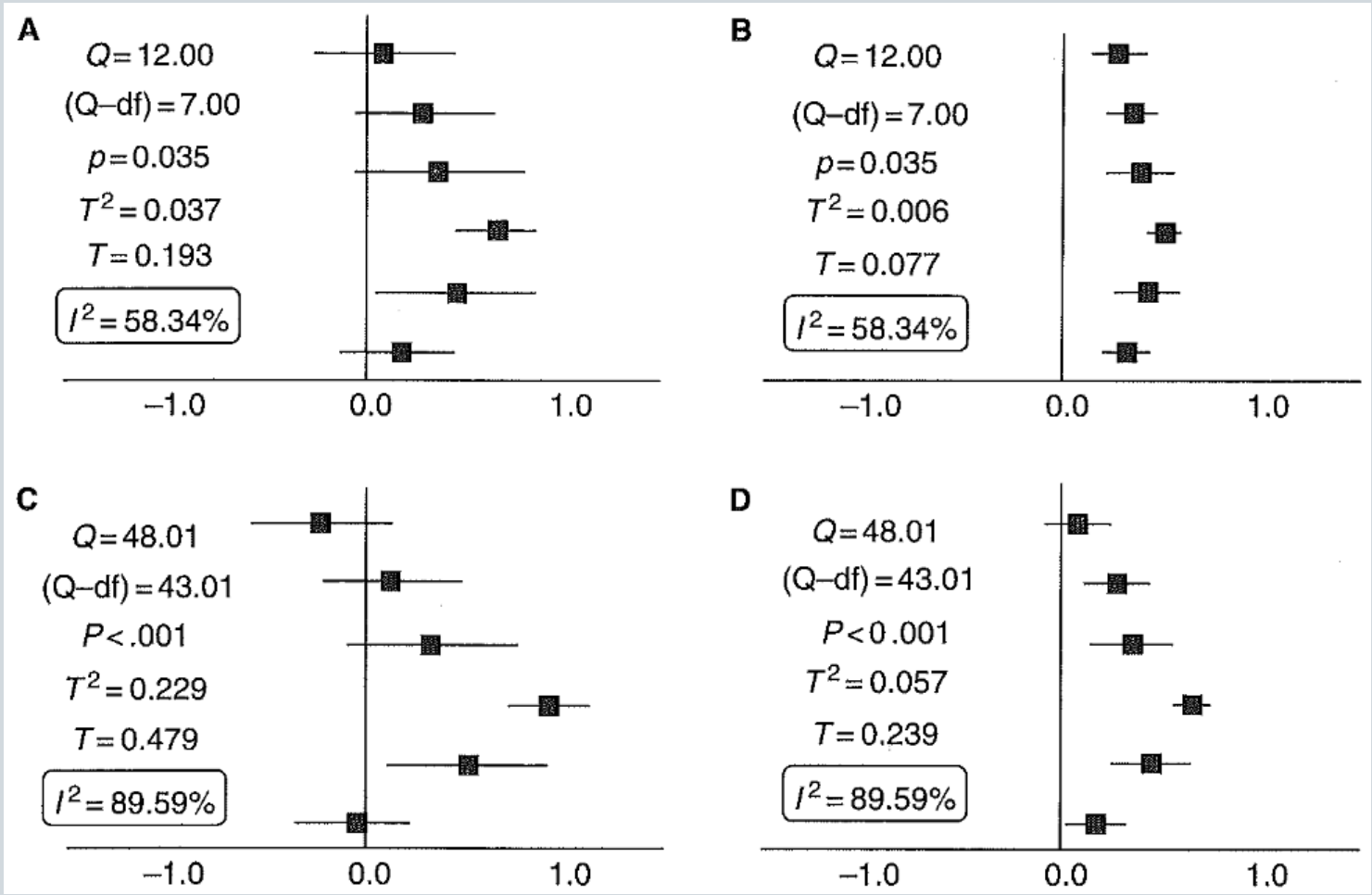
$$I^2 = \frac{Q - df}{Q} \cdot 100\%$$

- I^2 lässt sich konzeptionell beschreiben als

$$I^2 = \frac{Var_{between}}{Var_{total}} \cdot 100\% = \frac{\tau^2}{\tau^2 + V_y} \cdot 100\%$$

quantitative Heterogenität

Abhängigkeit von I^2



quantitative Heterogenität

I^2 Statistik

- für $I^2 \to 0$ wird die Variabilität der Daten nahezu vollständig durch die auftretenden Stichprobenfehler beschrieben
- für größere Werte von I^2 wird es interessant, die Gründe für die vorliegende Heterogenität zu erforschen ($\to$ Subgruppenanalyse, Metaregression)
- in *Higgins et al. (2003)* wurde vorgeschlagen, I^2-Werte von 25 %, 50% und 75% als *niedrig*, *moderate* und *hoch* einzustufen

$I^2 \to 100\%$ sagt lediglich aus, dass die meiste Variabilität der Daten auf eine Heterogenität der wahren Effekte zurückzuführen ist, sagt aber nicht, dass die Effekte weit streuen!

quantitative Heterogenität

weitere Maßzahlen zur Heterogenität in Metaanalysen

- H^2 beschreibt das Verhältnis der beobachteten Heterogenität zur erwarteten Heterogenität unter der Annahme einer gleichen, wahren Effektstärke für alle Studien

$$H^2 = \frac{Q}{df}$$

- R^2 beschreibt das das Verhältnis des Konfidenzintervalls unter dem random effects Models mit dem Konfidenzintervall des fixed effect Modells

$$R^2 = \frac{CI_{RE}}{CI_{FE}}$$

quantitative Heterogenität

Rechenbeispiel

- Analyse der Heterogenität der 6 Studien aus Vorlesung 2
- bias-korrigierte, standardisierte Mittelwertdifferenz (Hedges' g) als Maß der Effektstärke

Study	Treated			Control		
	Mean	*SD*	*n*	Mean	*SD*	*n*
Carroll	94	22	60	92	20	60
Grant	98	21	65	92	22	65
Peck	98	28	40	88	26	40
Donat	94	19	200	82	17	200
Stewart	98	21	50	88	22	45
Young	96	21	85	92	22	85

quantitative Heterogenität

Rechenbeispiel

- Berechnung der Effektstärken $Y_i = g$ und der zugehörigen Varianzen für alle Studien

Study	Effect	Variance	Weight	Calculated quantities			
	Y	V_Y	W	WY	WY^2	W^2	W^3
Carroll	0.095	0.033	30.352	2.869	0.271	921.21	27960.25
Grant	0.277	0.031	32.568	9.033	2.505	1060.68	34544.41
Peck	0.367	0.050	20.048	7.349	2.694	401.93	8058.00
Donat	0.664	0.011	95.111	63.190	41.983	9046.01	860371.10
Stewart	0.462	0.043	23.439	10.824	4.999	549.37	12876.47
Young	0.185	0.023	42.698	7.906	1.464	1823.12	77843.29
Sum			244.215	101.171	53.915	13802.33	1021653.52

quantitative Heterogenität

Rechenbeispiel

- das gewichtete Mittel wurde berechnet aus

$$M = \frac{\sum_{i=1}^{k} W_i Y_i}{\sum_{i=1}^{k} W_i} = \frac{101.171}{244.215} = 0.4143$$

- Berechnung von Q:

$$Q = \sum_{i=1}^{k} W_i Y_i^2 - \frac{(\sum_{i=1}^{k} W_i Y_i)^2}{\sum_{i=1}^{k} W_i} = 53.915 - \frac{101.171^2}{244.215} = 12.0033$$

quantitative Heterogenität

Rechenbeispiel

- unter der Annahme, dass alle Studien eine gemeinsame wahre Effektstärke aufweisen, wird der Erwartungswert für Q abgeschätzt durch

$$df = k - 1 = 6 - 1 = 5$$

- damit ergibt sich die *excess variation*

$$Q - df = 12.0033 - 5 = 7.0033$$

- Bestimmung des p-Wertes mittels der χ^2-Verteilung aus $Q = 12.0033$ und $df = 5$
- $p = 0.035$ ist auf $\alpha = 5\%$-Niveau als signifikant zu bewerten

quantitative Heterogenität

Rechenbeispiel

- unter Berücksichtigung der Wichtung C

$$C = \sum_{i=1}^{k} W_i - \frac{\sum_{i=1}^{k} W_i^2}{\sum_{i=1}^{k} W_i} = 244.215 - \frac{13802.325}{244.215} = 187.698$$

- ergibt sich für Varianz und Standardabweichung der Verteilung der wahren Effektstärken

$$T^2 = \frac{Q - df}{C} = \frac{12.0033 - 5}{187.698} = 0.0373$$

und

$$T = \sqrt{T^2} = \sqrt{0.0373} = 0.1932$$

quantitative Heterogenität

Rechenbeispiel

- das Verhältnis I^2 von *excess variation* zur gesamten Variabilität berechnet sich aus

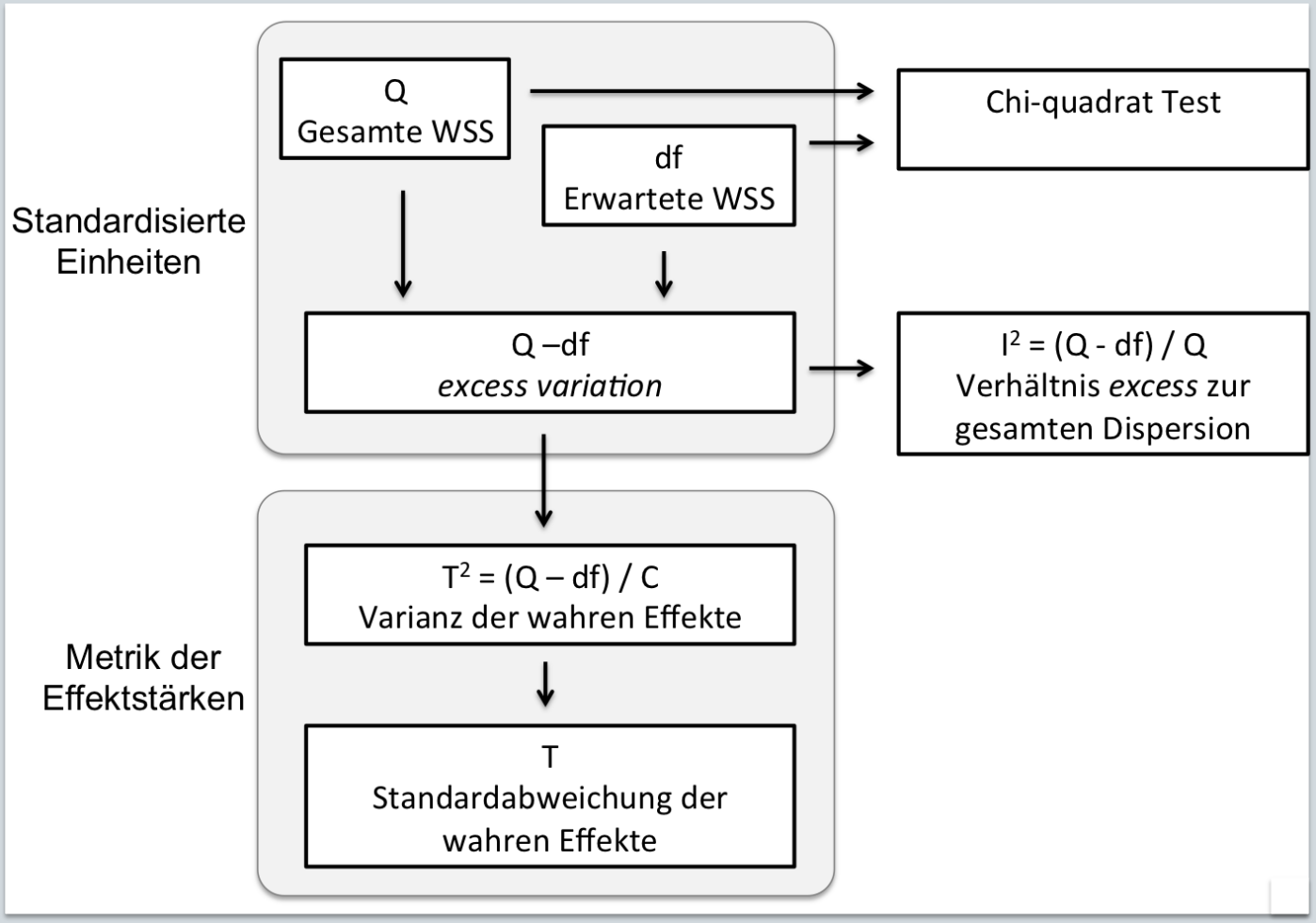
$$I^2 = \frac{Q - df}{Q} \cdot 100\%$$

- im konkreten Fall

$$I^2 = \frac{12.0033 - 5}{12.0033} \cdot 100\% = 58.34\%$$

quantitative Heterogenität

Übersicht der Maße



quantitative Heterogenität

Vergleich der Maße

- Q-**Statistik und** p-**Wert** dienen als Signifikanzmaße
- diese Maße sind abhängig von der Studienanzahl, aber unabhängig von der Metrik der Effektstärke

- T^2 **und** T schätzen die Varianz τ^2 und die Standardabweichung τ der Verteilung der wahren Effektstärken in den untersuchten Studien
- diese Maße sind unabhängig von der Studienanzahl

- I^2 ist das Verhältnis der wahren Heterogenität zur gesamten Variabilität der Resultate in den einzelnen Studien
- das Verhältnis ist unabhängig von der Metrik und von der Anzahl der untersuchten Studien

quantitative Heterogenität - Hinweise

p-Wert

- der p-Wert der Q-Statistik ist kein Maß der Effektstärke

Standardabweichung T

- T ist ein geeignetes Maß, um die Variabilität der wahren Effekte abzuschätzen

Genauigkeit

- die Unsicherheit bei der Schätzung von T^2 und I^2 ist häufig sehr groß
- für Studien mit geringer Genauigkeit (große Konfidenzintervalle) kann das zu einer Verdeckung der wahren Heterogenität führen ($T^2 \to 0$ und $I^2 \to 0$) $\to$ die Schlussfolgerung, dass die Effektstärken gleich sind, wäre dann nicht gerechtfertigt

Vorlesung 4

- Publikationsbias
- Subgruppenanalyse
- Meta-Regression
- *Individual Patient Data (IPD) meta analysis*

Publikationsbias

Publikationsbias

Vorbetrachtungen

- Metaanalysen gestatten eine statistisch sinnvolle Aggregation von Studienergebnissen
- sind allerdings die Studien eine nicht repräsentative, systematisch verzerrte ("biased") Stichprobe, wird der Summary Effect der Metaanalyse auch diesen Bias abbilden
- problematisch: Studien mit höheren Effektstärken werden häufiger publiziert

Publikationsbias

Das Problem fehlender Studien

- zur Durchführung eines Systematische Reviews, und insbesondere einer Metaanalyse, werden Einschlusskriterien für Studien definiert
- ABER: vollständiger Einschluss aller relevanter Studien scheitert in den meisten Fällen
- sind die fehlenden Studien eine *zufällige* Untergruppe aller relevanten Studien, so führt dieser Mangel zu geringerem Informationsgehalt und ungenaueren Schätzungen
- bei einer *systematischen* Abweichung der fehlenden Studien, wird das Gesamtergebnis der Untersuchung verzerrt

Publikationsbias

Das Problem fehlender Studien

tagesschau.de

Suche in tagesschau.de

Startseite | Videos & Audios | Inland | Ausland | Investigativ | Wirtschaft | Wahlen | Wetter | Ihre Meinung | Mehr

Startseite › Investigativ › Unis verheimlichen medizinische Studienergebnisse

Medizinische Forschung

Unis verheimlichen Studienergebnisse

Stand: 30.12.2019 06:48 Uhr

Bei 93 Prozent aller medizinischen Studien an deutschen Unis werden die Ergebnisse nicht vorschriftsgemäß veröffentlicht. Das geht aus einer neuen Untersuchung hervor, die *NDR, WDR* und SZ vorliegt.

Von Markus Grill, NDR/WDR

NDR

Ressort Investigation im NDR

| ndr

Publikationsbias

Signifikante vs. nicht-signifikante Resultate

- signifikante Resultate werden häufiger publiziert
- für eine gegebene Stichprobengröße ist es wahrscheinlicher ein signifikantes Ergebnis zu erhalten, wenn die Effektstärke größer ist
- $\rightarrow$ Studien mit größeren Effektstärken sind häufiger signifikant und werden demnach auch häufiger publiziert

Publikationsstatus sollte nie als Qualitätskriterium / Auswahlkriterium innerhalb einer Metaanalyse dienen

weitere Arten von Bias

Sprache, Zitation, *familiarity*, ...

Publikationsbias

Publikationsbias analysieren

- Publikationsbias wird sich nicht vollständig vermeiden lassen
- sophistizierte statistische Modelle erlauben die quantitative Analyse und die Kompensation von Publikationsbias, werden aber derzeitig wenig angewendet
- $\rightarrow$ Abschätzung des vorliegenden Bias:
  - Gibt es Hinweise für Publikationsbias?
  - Welchen Einfluss hat der Publikationsbias auf die Ergebnisse?

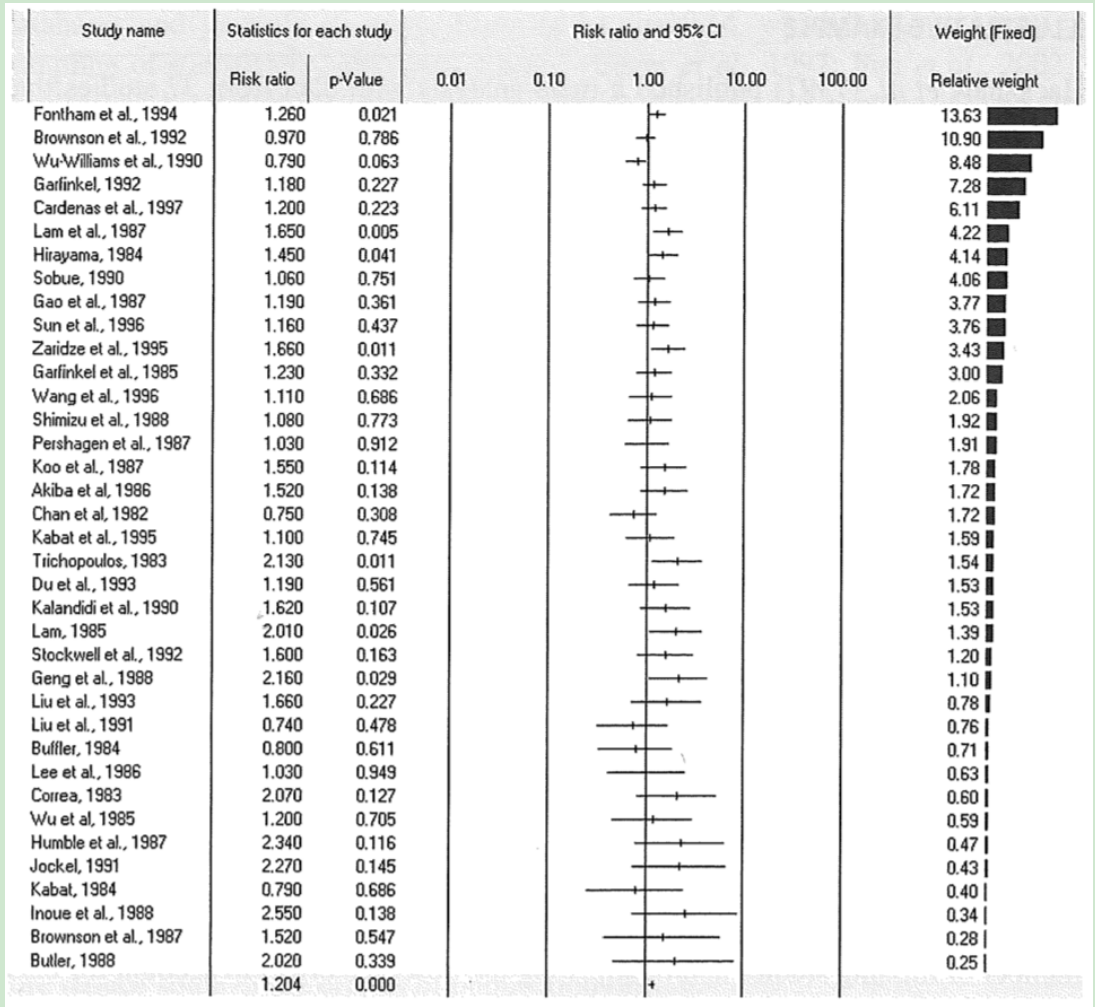
Publikationsbias

Bias konzeptuell verstehen

- Verständnis, welche Studien wahrscheinlich fehlen werden:
  - große Studien werden wahrscheinlich publiziert, unabhängig von ihrer statistischen Signifikanz
  - mittelgroße Studien werden vorrangig unpubliziert bleiben, wenn sie keine signifikanten Effekte aufweisen
  - kleine Studien haben das größte Risiko, nicht publiziert zu werden; nur sehr große Effektstärken führen (mit größerer Wahrscheinlichkeit) zu signifikanten Ergebnissen
- Publikationsbias wird größer, je kleiner die Studien werden

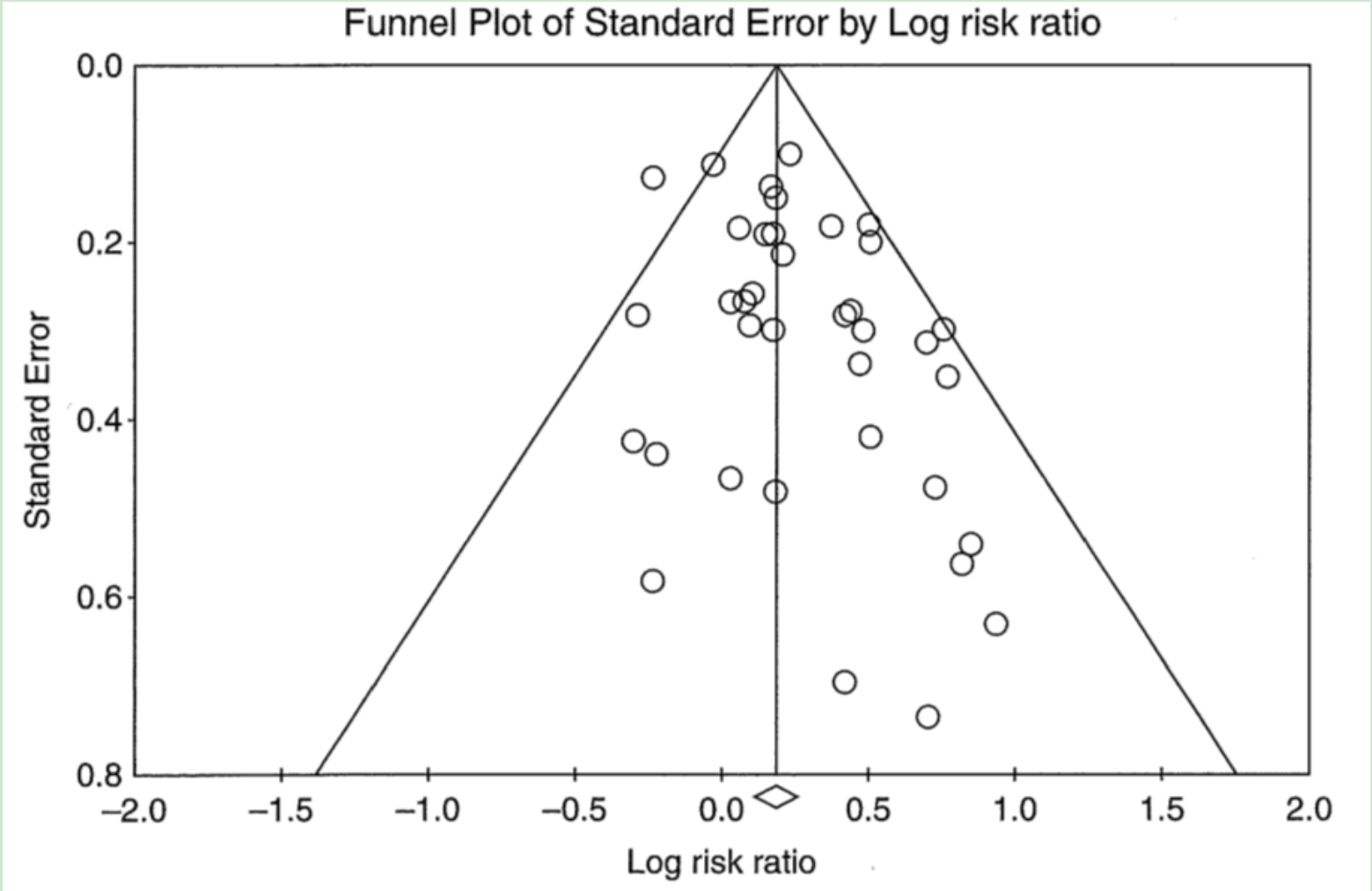
Publikationsbias

Passivrauchen und Lungenkrebs - Forest Plot



Publikationsbias

Passivrauchen und Lungenkrebs - Funnel Plot



Publikationsbias

Funnel Plot

- Funnel ("Trichter") Plots sind die gängigste Darstellung zur Adressierung von Bias
- Effektstärke wird traditionell auf der x-Achse dargestellt, Stichprobengröße oder Varianz auf der y-Achse
- größere Studien erscheinen "oben" (genauere Schätzung), kleinere "unten" (größerer Stichprobenfehler)
- Darstellung des Standardfehlers (anstatt der Varianz) auf der y-Achse streckt die Darstellung für die kleineren Studien
- $\rightarrow$ (visuelle) Einschätzung der Asymmetrie

Publikationsbias

Einfluss des Publikationsbias

Welchen Einfluss hat der Publikationsbias auf das Ergebnis der Metaanalyse?

a) Einfluss von Bias ist marginal. Selbst bei vollständigem Einschluss aller relevanten Untersuchungen bliebe der Summary Effekt weitgehend gleich.

b) Moderater Einfluss von Publikationsbias. Effektstärke wurde stärker variieren, aber die Grundaussage bliebe unverändert.

c) Publikationsbias hat großen Einfluss und verändert die Interpretation der Ergebnisse substantiell.

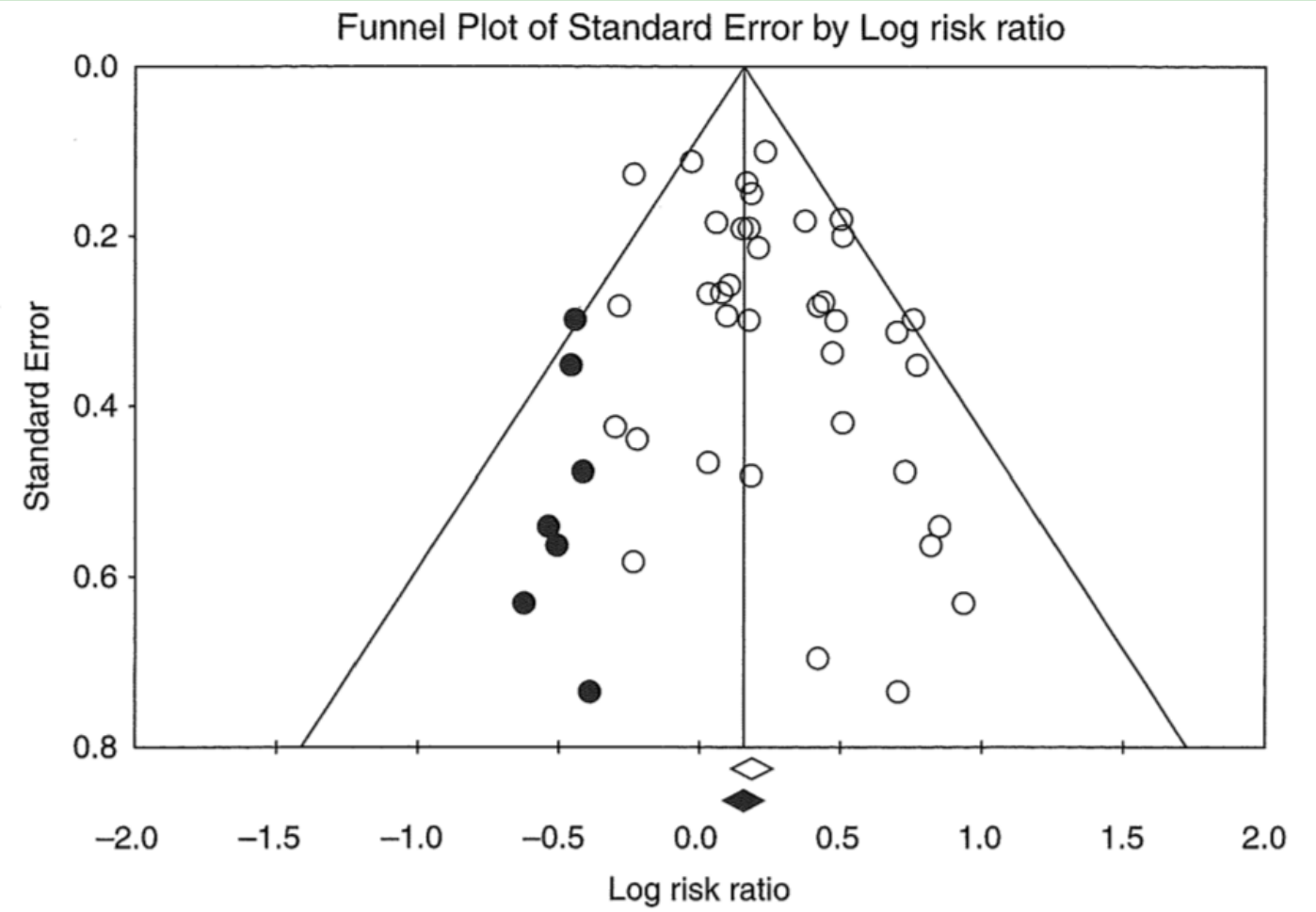
Publikationsbias

Quantifizierung des Publikationsbias

- *Rosenthal's* (und *Orwin's*) *Fail-safe N*:
  - Hinzufügen von N kleinen Studien mit Effektstärke $Y_i = 0$ ($Y_i = Y_{min}$) bis der Summary Effekt nicht mehr signifikant ist
  - große N vermitteln mehr Sicherheit, dass der Summary Effekt nicht zu stark durch Publikationsbias von kleinen Studien verzerrt wird
- *Trim and Fill*:
  - iteratives Entfernen kleiner Studien mit extremen Effektstärken $\rightarrow$ *unbiased* Schätzung der Effektstärke (aber zu stark reduzierte Varianz)
  - auffüllen aller originalen Studien und deren "Spiegelbilder" (Imputation) $\rightarrow$ kein Einfluss auf den Mittelwertschätzer der Effektstärke, aber Korrektur der Varianz

Publikationsbias

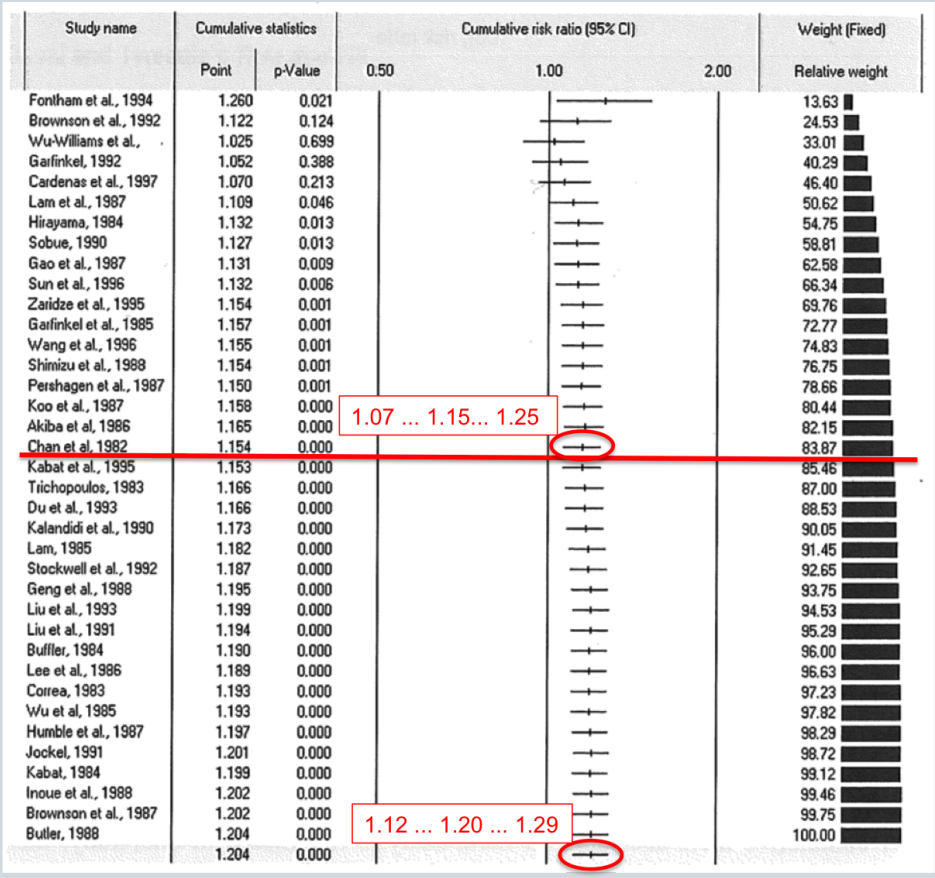
Trim and Fill



Publikationsbias

Quantifizierung des Publikationsbias

- Beschränkung auf hinreichend große Studien: Kumulativer Forest Plot



Publikationsbias

Zusammenfassung zum Publikationsbias

- Adressierung von Publikationsbias ist essentiell (Anfechtbarkeit der Untersuchung)
- Funnel Plots und sortierte Forest Plots sind geeignet für die visuelle Untersuchung von Asymmetrien in den eingeschlossenen Studien
- korrigierte Effektstärken (*Trim and Fill*, Beschränkung auf große Studien) vermitteln einen Eindruck, inwiefern die ermittelten Effektstärken von Publikationsbias beeinflusst werden $\rightarrow$ ändert sich die Interpretation der Ergebnisse?

Subgruppenanalyse

Subgruppenanalyse

Vorbetrachtungen

- Metaanalysen gehen über die Schätzung eines Summary Effects hinaus
  $\rightarrow$ Erklärung der "Muster" von Effekten

- Bestimmte Medikamente reduzieren das Sterberisiko bei Patienten mit kardialen Arhythmien. Allerdings wird spekuliert, dass die Stärke dieses Effektes davon abhängt, ob es sich um akute oder chronische Situationen handelt. Beide Gruppen sollen getrennt untersucht und verglichen werden.
- Wie unterscheiden sich die Ergebnisse einer Metaanalyse, wenn innerhalb der Studien verschiedene Randomisationsschemata verwendet wurden?

Subgruppenanalyse

Vorgehen

Drei Verfahren für die Subgruppenanalyse

- fixed effect
- random effects mit individuellem Schätzer für τ^2
- random effects mit einem gemittelten Schätzer für τ^2

Subgruppenanalyse

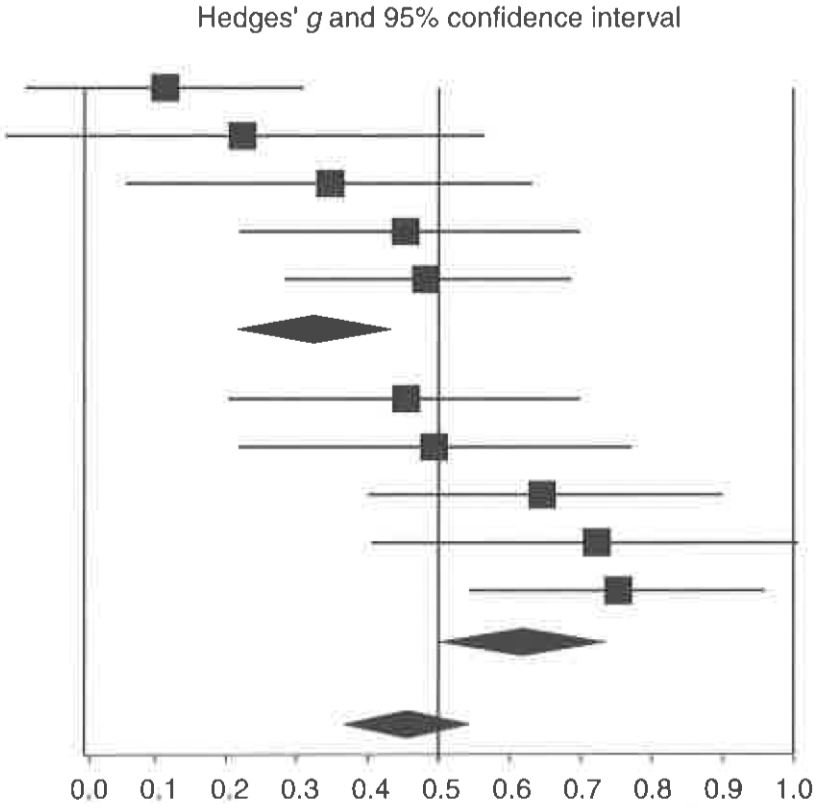
Beispieldatensatz

- Metaanalyse zur Untersuchung eines neuen Lernverfahren im Mathematikunterricht, wobei zwei verschiedene Lehrbücher zum Einsatz kommen. Beide sind wirksam, und es soll evaluiert werden, welches Lehrbuch effektiver ist.
- Standardisierten Differenz von Mittelwerten (Hedges' g) als Maß der Effektstärke.
- *Idealerweise würde man die beiden Lernverfahren in einer Studie direkt miteinander vergleichen.*

Subgruppenanalyse

Beispieldatensatz: $I^2 = 65\%$

	Study	g	Variance
A	Thornhill	0.110	0.010
	Kendall	0.224	0.030
	Vandamm	0.338	0.020
	Leonard	0.451	0.015
	Professor	0.480	0.010
	A	0.324	0.003
B	Jeffries	0.440	0.015
	Fremont	0.492	0.020
	Doyle	0.651	0.015
	Stella	0.710	0.025
	Thorwald	0.740	0.012
	B	0.611	0.003
	Combined	0.458	0.002



Subgruppenanalyse

Wie würde man vorgehen?

- Bei einzelnen Studien würde man den Mittelwert und die Varianz in jeder Gruppe berechnen. Mithilfe eines t-Tests könnte man dann die Mittelwertdifferenz ins Verhältnis zur beobachteten Varianz setzen.
- Grundsätzlich gleiches Vorgehen bei einer Subgruppenanalyse:
  - Mittlere Effektstärke und Varianzschätzung für jede Subgruppe.
  - Vergleich der mittleren Effekte zwischen den Subgruppen.

Subgruppenanalyse

Effektstärke für jede Subgruppe

	Study	Effect size Y	Variance Within V_Y	Variance Between T^2	Variance Total V	Weight W	Calculated quantities WY	WY^2	W^2
A	Thornhill	0.110	0.0100	0.0000	0.0100	100.000	11.000	1.210	10000.000
	Kendall	0.224	0.0300	0.0000	0.0300	33.333	7.467	1.673	1111.111
	Vandamm	0.338	0.0200	0.0000	0.0200	50.000	16.900	5.712	2500.000
	Leonard	0.451	0.0150	0.0000	0.0150	66.667	30.067	13.560	4444.444
	Professor	0.480	0.0100	0.0000	0.0100	100.000	48.000	23.040	10000.000
	Sum A					350.000	113.433	45.195	28055.556
B	Jefferies	0.440	0.0150	0.0000	0.0150	66.667	29.333	12.907	4444.444
	Fremont	0.492	0.0200	0.0000	0.0200	50.000	24.600	12.103	2500.000
	Doyle	0.651	0.0150	0.0000	0.0150	66.667	43.400	28.253	4444.444
	Stella	0.710	0.0250	0.0000	0.0250	40.000	28.400	20.164	1600.000
	Thorwald	0.740	0.0120	0.0000	0.0120	83.333	61.667	45.633	6944.444
	Sum B					306.667	187.400	119.061	19933.333
	Sum					656.667	300.833	164.255	47988.889

Subgruppenanalyse

Effektstärke für Subgruppe A

Vorgehen analog zur fixed effect Metaanalyse aus VL 2

- mittlere Effektstärke M_A berechnet sich nach

$$M_A = \frac{\sum_{i=1}^{k} W_{A_i} Y_{A_i}}{\sum_{i=1}^{k} W_{A_i}} = \frac{113.433}{350.0} = 0.3241$$

- Varianz der gemittelten Effektstärke

$$V_{M_A} = \frac{1}{\sum_{i=1}^{k} W_{A_i}} = \frac{1}{350.0} = 0.0029$$

Subgruppenanalyse

Effektstärke für Subgruppe A

daraus ergibt sich

- Standardfehler der gemittelten Effektstärke

$$SE_{M_A} = \sqrt{V_{M_A}} = \sqrt{0.0029} = 0.0535$$

- Grenzen des 95% Konfidenzintervalls

$$LL_{M_A} = M_A - 1.96 \cdot SE_{M_A} = 0.3241 - 1.96 \cdot 0.0535 = 0.2193$$

$$UL_{M_A} = M_A + 1.96 \cdot SE_{M_A} = 0.3241 + 1.96 \cdot 0.0535 = 0.4289$$

Subgruppenanalyse

Effektstärke für Subgruppenanalyse

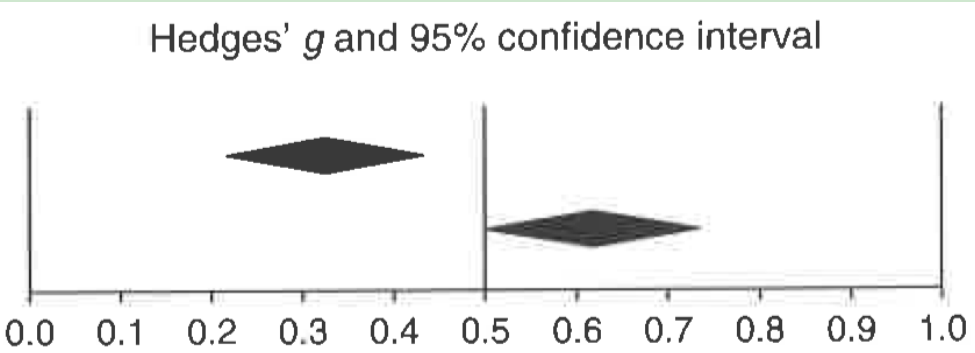
Analoge Berechnung für Subgruppe B und für beide Gruppen gemeinsam

	A	B	Gesamt
M	0.3241	0.6111	0.4581
V	0.0029	0.0033	0.0015
SE_M	0.0535	0.0571	0.0390
LL_M	0.2193	0.4992	0.3816
UL_M	0.4289	0.7230	0.5346

Subgruppenanalyse

Vergleich der Subgruppen

Subgroup	*g*	Variance
A	0.324	0.003
B	0.611	0.003



$\rightarrow$ mehrere Möglichkeiten, die Gruppen miteinander zu vergleichen

Subgruppenanalyse

Test-verfahren

- Für Subgruppenanalysen mit zwei Gruppen kann man die Differenz der Mittelwerte $Diff = M_B - M_A$ mithilfe eines t-Tests beurteilen.
- Mittelwertvergleiche von mehr als zwei Gruppen können mit einer ANOVA durchgeführt werden.
- Aufteilung der Varianzkomponenten in die Varianz zwischen den Gruppen ($Q_{between}$) und die Varianz innerhalb der Gruppen (Q_{within}).
- die genannten Methoden (t-Test und ANOVA) sind mathematisch äquivalent (gleicher p-Wert)
- zusätzlich zur Angabe der statistischen Signifikanz sollte die Differenz der Mittelwerte $Diff = M_B - M_A$ mit den zugehörigen 95% Konfidenzintervallen angegeben werden ($UL/LL_{Diff} = Diff \pm 1.96 \cdot SE_{Diff}$)

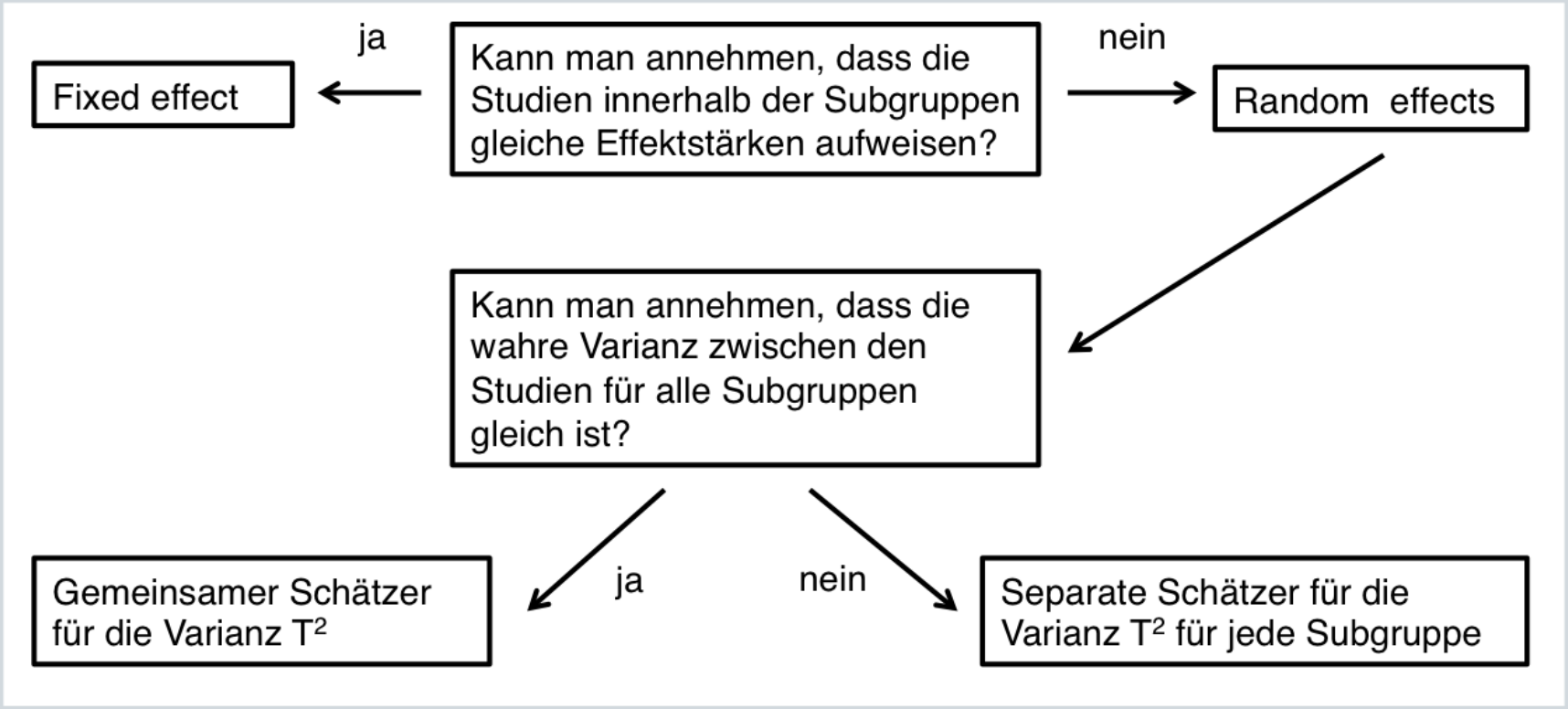
Subgruppenanalyse

random-effects model

- Die Anwendung eines fixed-effect model ist nur unter der Annahme einer gemeinsamen Effektstärke zulässig.
- Realistischer sind Szenarien, in denen die wahren Effektstärken einer Verteilung unterliegen. Für diese Szenarien sollte ein random-effects model angewendet werden.
- Dabei ist noch zu unterscheiden, ob die wahre Varianz τ^2 für jede Gruppe einzeln geschätzt wird, oder ob die Varianz als Abweichung aller Studien von den Gruppenmittelwerten bestimmt wird. Letztere Möglichkeit bietet sich an, wenn man annehmen kann, das die Varianz innerhalb der Gruppen gleich ist. Gleiches gilt bei einer kleinen Anzahl zur Verfügung stehender Studien pro Gruppe.

Subgruppenanalyse

Methodik



Subgruppenanalyse

random-effects model mit einem gemeinsamen Schätzer für τ^2_{within}

- Unter der Annahme, das die wahre Varianz der Studien innerhalb der Gruppen ähnlich ist (d.h. ähnliche Verteilung um die jeweilige mittlere Effektstärke in der Subgruppe), so kann eine gemeinsame (gepoolte) Varianz der wahren Effektstärken bestimmt werden.
- Abschätzung durch $T^2 = \frac{Q-df}{C}$, wobei $C = \sum_{i=1}^{k} W_i - \frac{\sum_{i=1}^{k} W_i^2}{\sum_{i=1}^{k} W_i}$
- Ein gemittelter Schätzer für τ^2_{within} ergibt sich aus der Summe der Q, df und C (für jede Sugruppe j) entsprechend:

$$T^2_{within} = \frac{\sum_{j=1}^{m} Q_j - \sum_{j=1}^{m} df_j}{\sum_{j=1}^{m} C_j}$$

Subgruppenanalyse

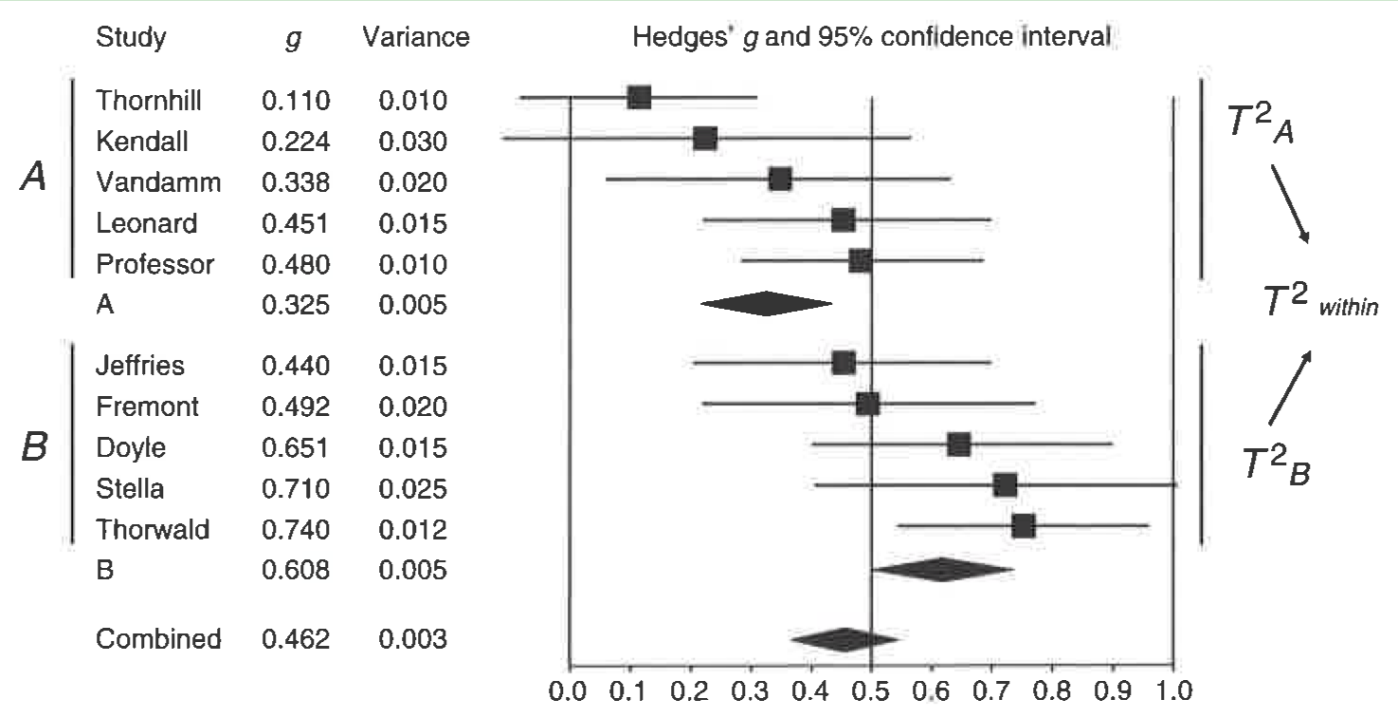
random-effects model mit einem gemeinsamen Schätzer für τ^2_{within}

Für den konrekten Fall ist die Spalte T^2 gleich für alle Studien.

	Study	Effect size Y	Variance Within V_Y	Variance Between T^2	Variance Total V	Weight W	Calculated quantities WY	WY^2	W^2
A	Thornhill	0.110	0.0100	0.0097	0.0197	50.697	5.577	0.613	2570.150
	Kendall	0.224	0.0300	0.0097	0.0397	25.173	5.639	1.263	633.678
	Vandamm	0.338	0.0200	0.0097	0.0297	33.642	11.371	3.843	1131.752
	Leonard	0.451	0.0150	0.0097	0.0247	40.445	18.241	8.226	1635.767
	Professor	0.480	0.0100	0.0097	0.0197	50.697	24.334	11.681	2570.150
	Sum A					200.652	65.161	25.627	8541.498
B	Jefferies	0.440	0.0150	0.0097	0.0247	40.445	17.796	7.830	1635.767
	Fremont	0.492	0.0200	0.0097	0.0297	33.642	16.552	8.143	1131.752
	Doyle	0.651	0.0150	0.0097	0.0247	40.445	26.329	17.140	1635.767
	Stella	0.710	0.0250	0.0097	0.0347	28.798	20.446	14.517	829.299
	Thorwald	0.740	0.0120	0.0097	0.0217	46.030	34.062	25.206	2118.721
	Sum B					189.358	115.185	72.837	7351.306
	Sum					390.010	180.346	98.463	15892.804

Subgruppenanalyse

random-effects model mit einem gemeinsamen Schätzer für τ^2_{within}



Achtung: Angabe eines summary effects sollte gut begründet sein

Subgruppenanalyse

random-effects model mit einem gemeinsamen Schätzer für τ^2_{within}

- in R werden Subgruppen als Kovariablen einbezogen
- ähnliches Vorgehen wie bei einer linearen Regression

```
> sampleMA <- rma(Y, V, mods = ~Type, data = bookData, method="DL", slab=Study); sampleMA

Mixed-Effects Model (k = 10; tau^2 estimator: DL)

tau^2 (estimated amount of residual heterogeneity):     0.0097 (SE = 0.0128)
tau (square root of estimated tau^2 value):             0.0986
I^2 (residual heterogeneity / unaccounted variability): 38.34%
H^2 (unaccounted variability / sampling variability):   1.62
R^2 (amount of heterogeneity accounted for):            67.45%

Test for Residual Heterogeneity:
QE(df = 8) = 12.9745, p-val = 0.1127

Test of Moderators (coefficient(s) 2):
QM(df = 1) = 7.8324, p-val = 0.0051

Model Results:

         estimate      se    zval    pval   ci.lb   ci.ub
intrcpt    0.3247  0.0706  4.6001  <.0001  0.1864  0.4631  ***
TypeB      0.2835  0.1013  2.7987  0.0051  0.0850  0.4821   **

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Subgruppenanalyse

Anteil der erklärten Varianz

- Das Maß R^2 wird häufig verwendet, um den Anteil der durch eine Kovariable erklärte Varianz an der Gesamtvarianz zu beschreiben.
- Im Kontext einer Metaanalyse wird R^2 über die wahre Varianz τ^2 definiert als:

$$R^2 = 1 - (\frac{T^2_{within}}{T^2_{total}})$$

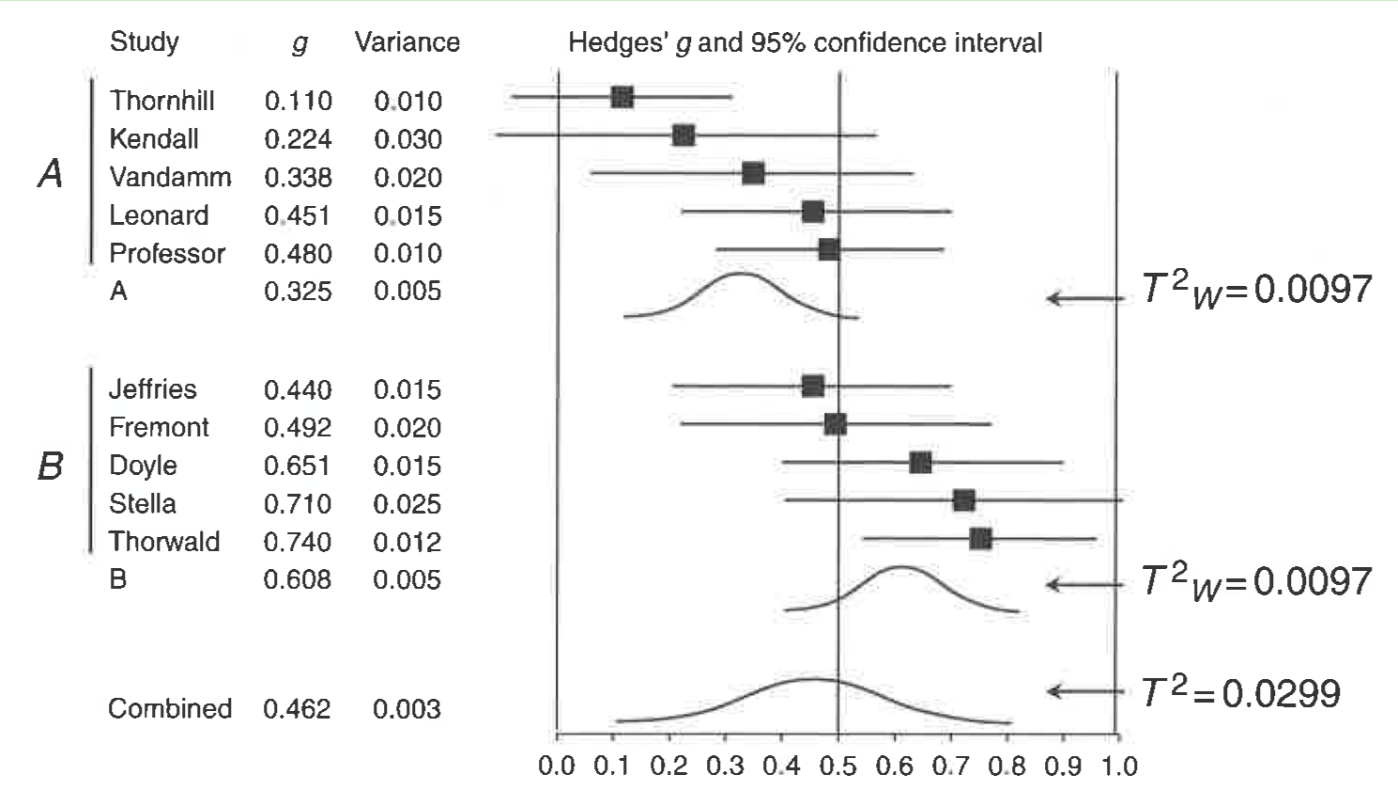
- Voraussetzung: gleiches τ^2 für alle Subgruppen

$$R^2 = 1 - (\frac{0.0097}{0.0299}) = 0.6745$$

67 % der gesamten Varianz können durch die Gruppenzugehörigkeit erklärt werden.

Subgruppenanalyse

Anteil der erklärten Varianz



Subgruppenanalyse

gemeinsamer summary effect?

- Es stellt sich die Frage, ob es sinnvoll ist, einen gemeinsamen summary effect für alle Studien anzugeben.
- Die Antwort ist von der Fragestellung abhängig: wenn es darum geht, den Unterschied zweier Methoden nachzuweisen, ist das vermutlich *nicht* sinnvoll. Für *ähnliche Ergebnisse* in den Subgruppen könnte es dennoch interessant sein, einen gemeinsamen summary effect anzugeben.

Subgruppenanalyse

gemeinsamer summary effect?

mehrere Möglichkeiten, einen gemeinsamen summary effect M anzugeben:

- kombinierter summary effect für die Subgruppen: fixed-effect Analyse für die mittlere Effektstärke und die Varianz jeder Subgruppe
- Analoges Vorgehen mit einer random-effects Analyse (mehr Studien erforderlich).
- Zusätzliche Metaanalyse über alle Studien, ohne die Zugehörigkeit zu einer Subgruppe zu berücksichtigen.

Metaregression

Metaregression

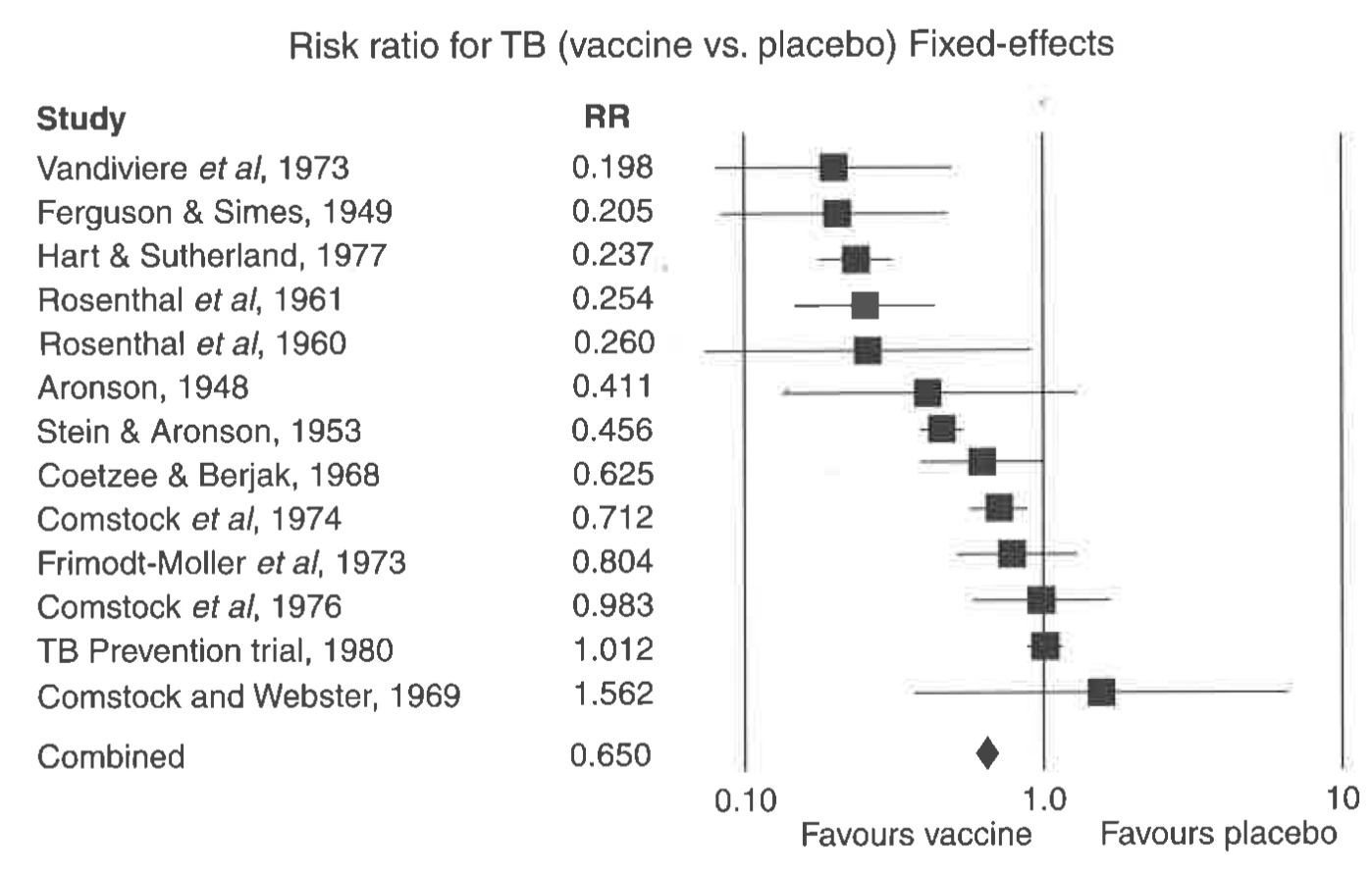
Einführung

- Regressionsverfahren werden verwendet, um den Einfluss einer metrischen Kovariable auf eine abhängige Größe zu beschreiben
- Bei einer Metaregression treten einzelne Studien an die Stelle von individuellen Messwerten.
- Ein sinnvolles Verhältnis der Anzahl an Studien zur Anzahl der Kovariablen ist erforderlich.

$\rightarrow$ Vorstellen anhand eines Beispieles

Metaregression

Metaanalyse für Tuberkulose-Impfung



RR von 0.2 bis 1.56; $I^2 = 92.12\%$

Metaregression

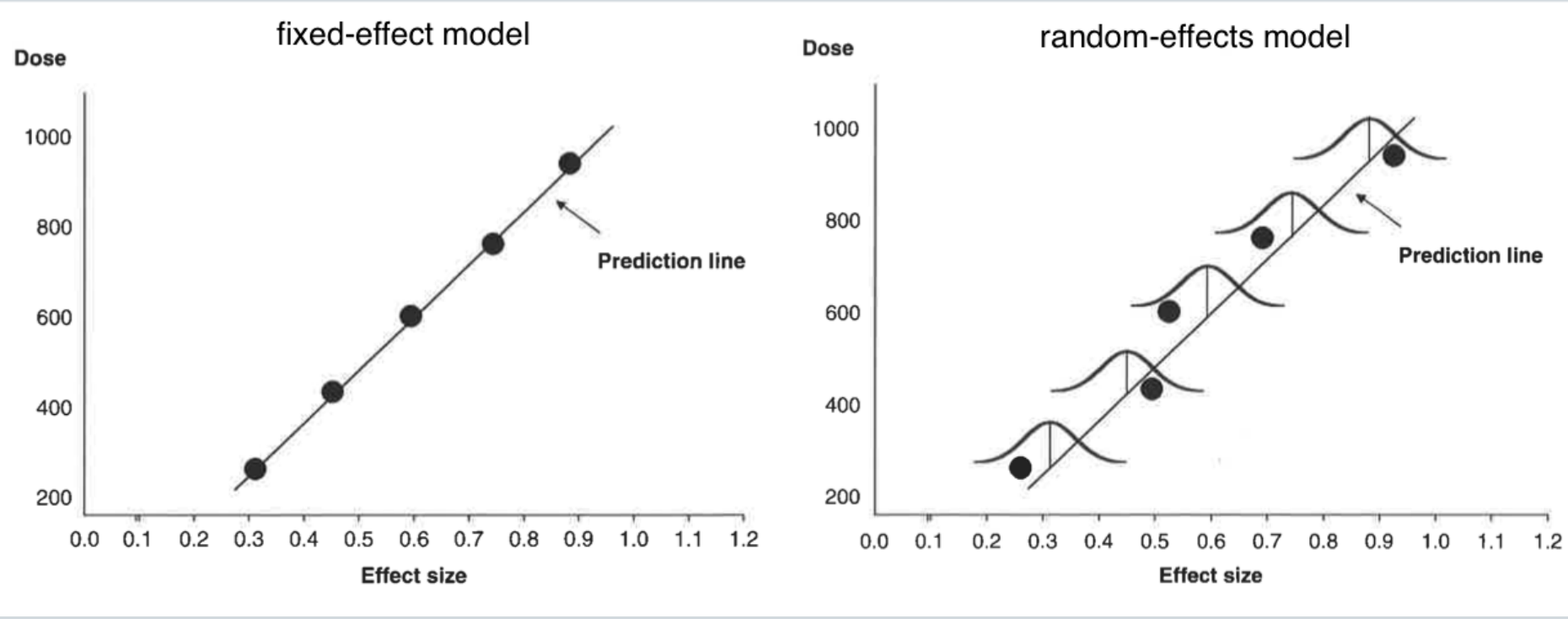
Metaanalyse für Tuberkulose-Impfung

Einfluss des Breitengrades als Maß für die Klimazone

	Vaccinated		Control					
	TB	Total	TB	Total	*RR*	ln*RR*	V_{lnRR}	Latitude
Vandiviere et al, 1973	8	2545	10	629	0.198	−1.621	0.223	19
Ferguson & Simes, 1949	6	306	29	303	0.205	−1.585	0.195	55
Hart & Sutherland, 1977	62	13598	248	12867	0.237	−1.442	0.020	52
Rosenthal *et al*, 1961	17	1716	65	1665	0.254	−1.371	0.073	42
Rosenthal *et al*, 1960	3	231	11	220	0.260	−1.348	0.415	42
Aronson, 1948	4	123	11	139	0.411	−0.889	0.326	44
Stein & Aaronson, 1953	180	1541	372	1451	0.456	−0.786	0.007	44
Coetzee & Berjak, 1968	29	7499	45	7277	0.625	−0.469	0.056	27
Comstock *et al*, 1974	186	50634	141	27338	0.712	−0.339	0.012	18
Frimodt-Moller *et al*, 1973.	33	5069	47	5808	0.804	−0.218	0.051	13
Comstock *et al*, 1976	27	16913	29	17854	0.983	−0.017	0.071	33
TB Prevention Trial, 1980	505	88391	499	88391	1.012	0.012	0.004	13
Comstock & Webster, 1969	5	2498	3	2341	1.562	0.446	0.533	33

Metaregression

Fixed- vs. random effects



Random-effects model für Tuberkulose-Impfung

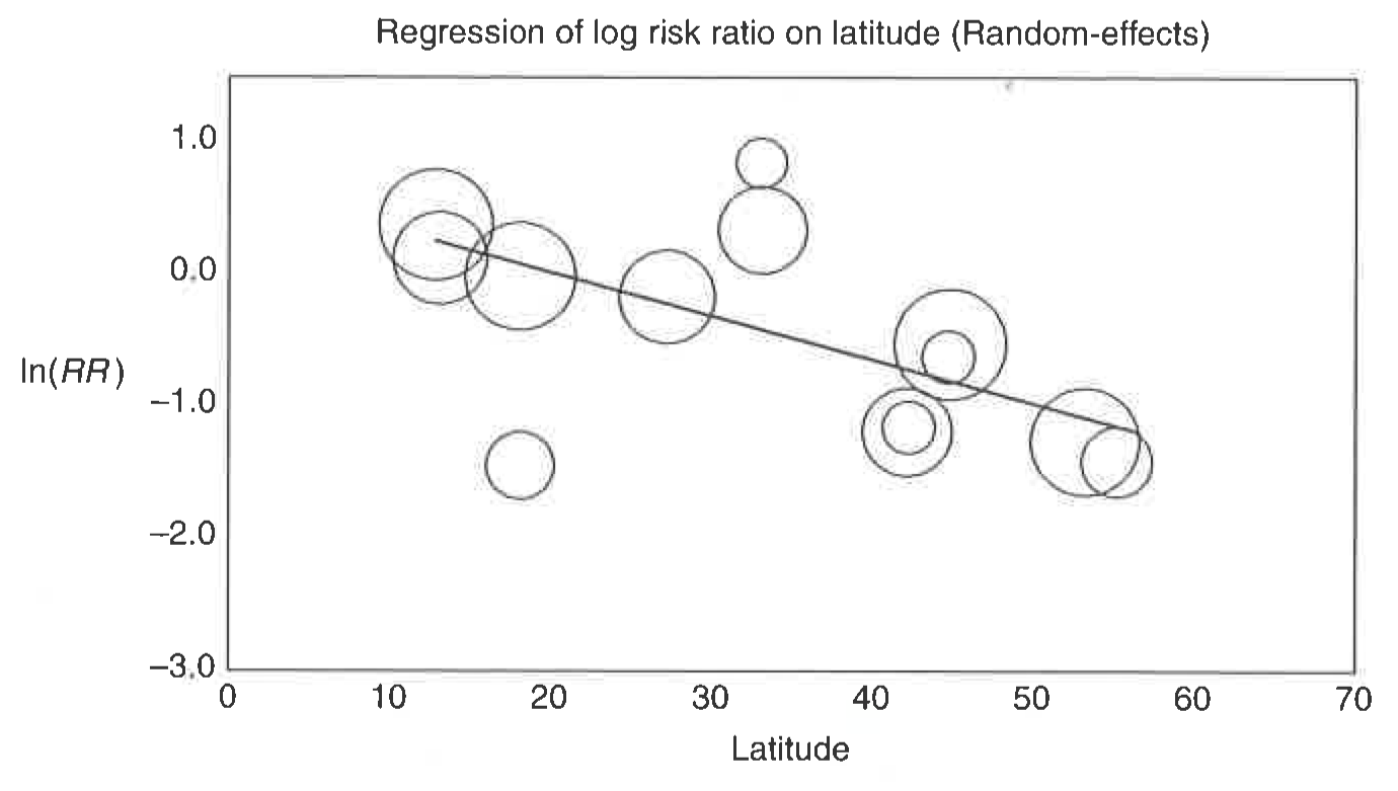
Ergebnisse der Regression

```
> res_bcg.RE_abl <- rma(yi, vi, mods = ~ablat, data = BCG, method="DL"); res_bcg.RE_abl

Mixed-Effects Model (k = 13; tau^2 estimator: DL)

tau^2 (estimated amount of residual heterogeneity):     0.0633 (SE = 0.0548)
tau (square root of estimated tau^2 value):             0.2516
I^2 (residual heterogeneity / unaccounted variability): 64.21%
H^2 (unaccounted variability / sampling variability):   2.79
R^2 (amount of heterogeneity accounted for):            79.50%

Test for Residual Heterogeneity:
QE(df = 11) = 30.7331, p-val = 0.0012

Test of Moderators (coefficient(s) 2):
QM(df = 1) = 18.8452, p-val < .0001

Model Results:

         estimate      se     zval    pval    ci.lb    ci.ub
intrcpt    0.2595  0.2323   1.1172  0.2639  -0.1958   0.7149
ablat     -0.0292  0.0067  -4.3411  <.0001  -0.0424  -0.0160  ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

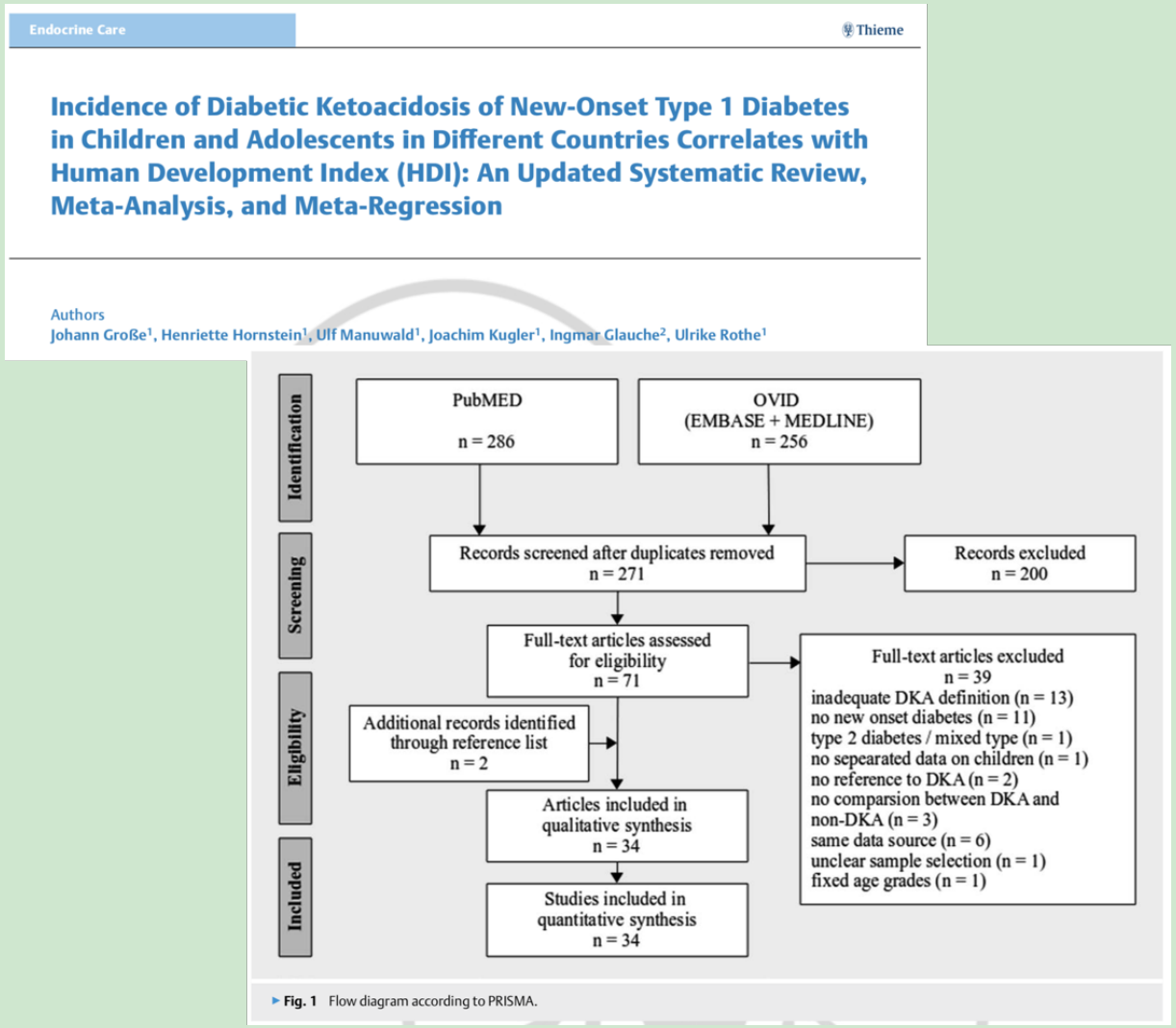
Metaregression

Random-effects model für Tuberkulose-Impfung

grafische Darstellung



Metaregression



Endocrine Care

Thieme

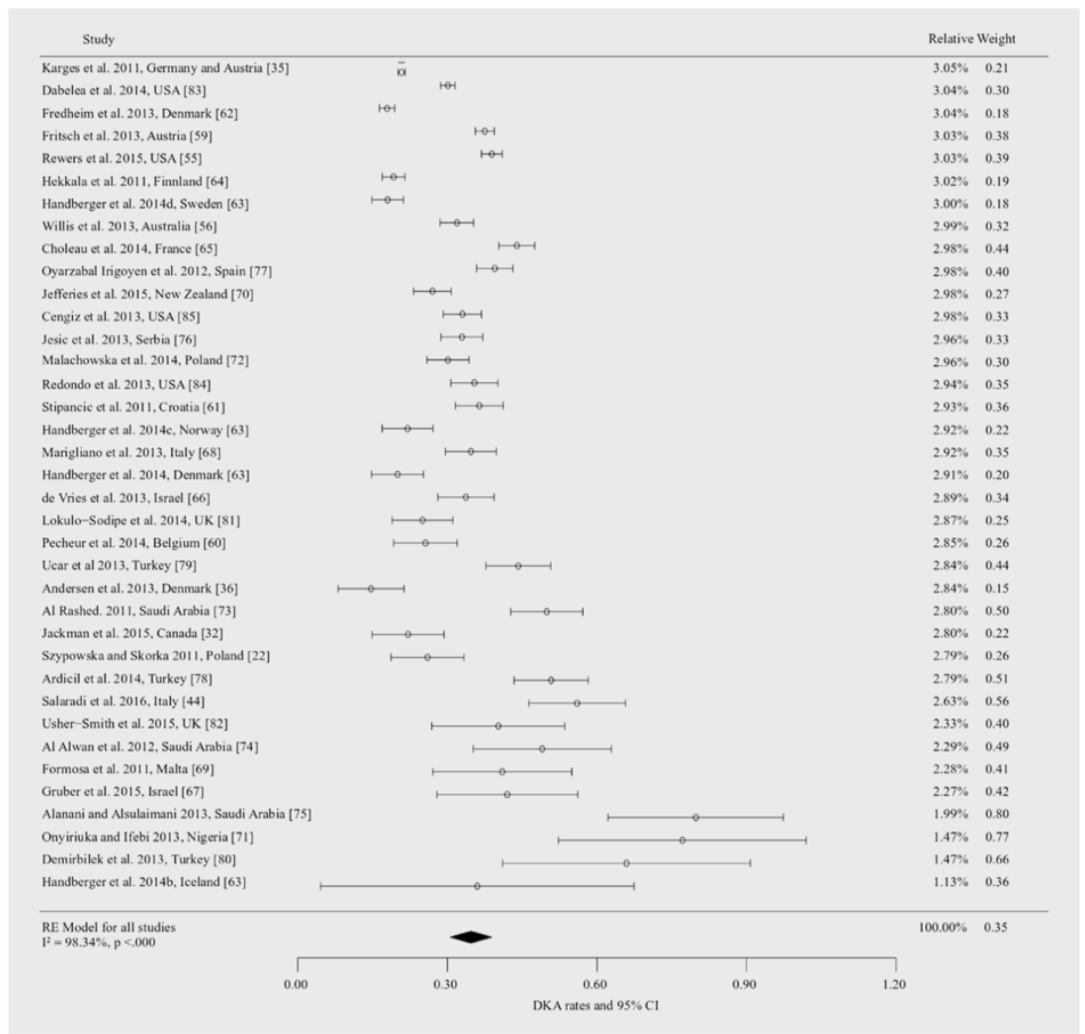
Incidence of Diabetic Ketoacidosis of New-Onset Type 1 Diabetes in Children and Adolescents in Different Countries Correlates with Human Development Index (HDI): An Updated Systematic Review, Meta-Analysis, and Meta-Regression

Authors

Johann Große[1], Henriette Hornstein[1], Ulf Manuwald[1], Joachim Kugler[1], Ingmar Glauche[2], Ulrike Rothe[1]

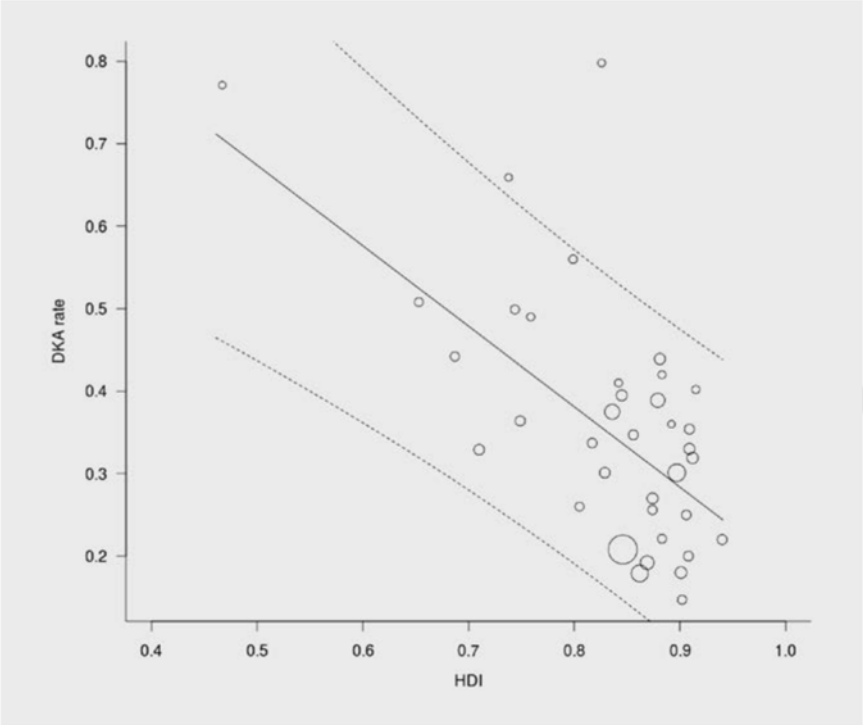
▸ **Fig. 1** Flow diagram according to PRISMA.

Metaregression

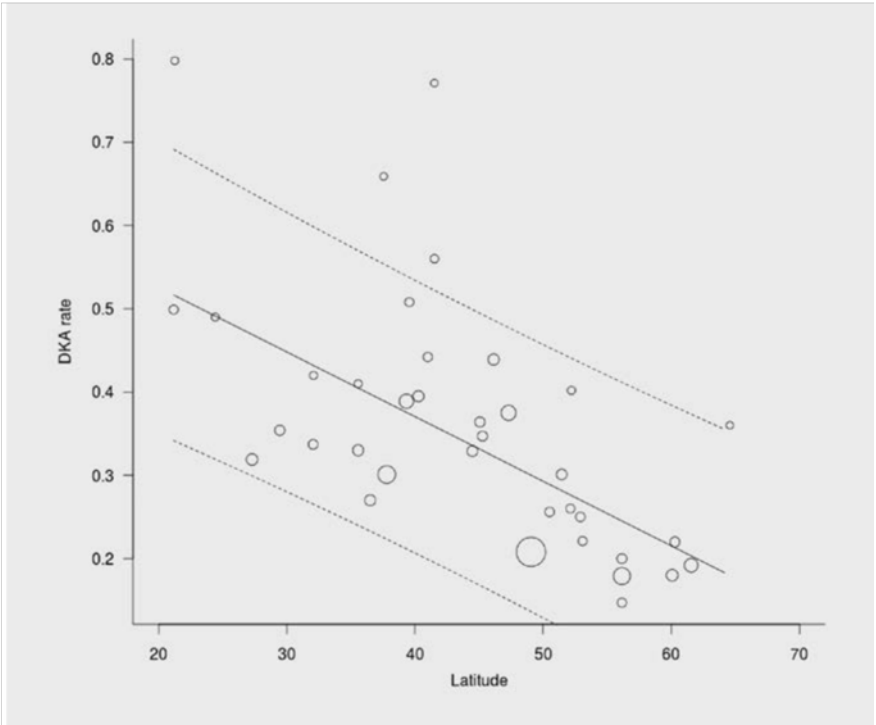


▶ **Fig. 2** Forest plot of DKA frequency (±95 % CI) at diagnosis of type 1 diabetes per study in descending order of sampling variances (Cochran Q = 1200.524, df = 36, p < 0.000; I^2 = 98.34 %, p < 0.000).

Metaregression



▶ **Fig. 3** Meta-regression DKA vs. HDI. Position of the circles indicates observed DKA rate for the human development index (HDI) of a given study, while the diameter refers to their relative weight. Dashed lines indicate 95 % confidence intervals for the univariate linear regression (solid line).



▶ **Fig. 4** Meta-regression DKA vs. Latitude. Position of the circles indicates observed DKA rate for the Latitude of a given study, while the diameter refers to their relative weight. Dashed lines indicate 95 % confidence intervals for the univariate linear regression (solid line).

Metaregression

Würdigung

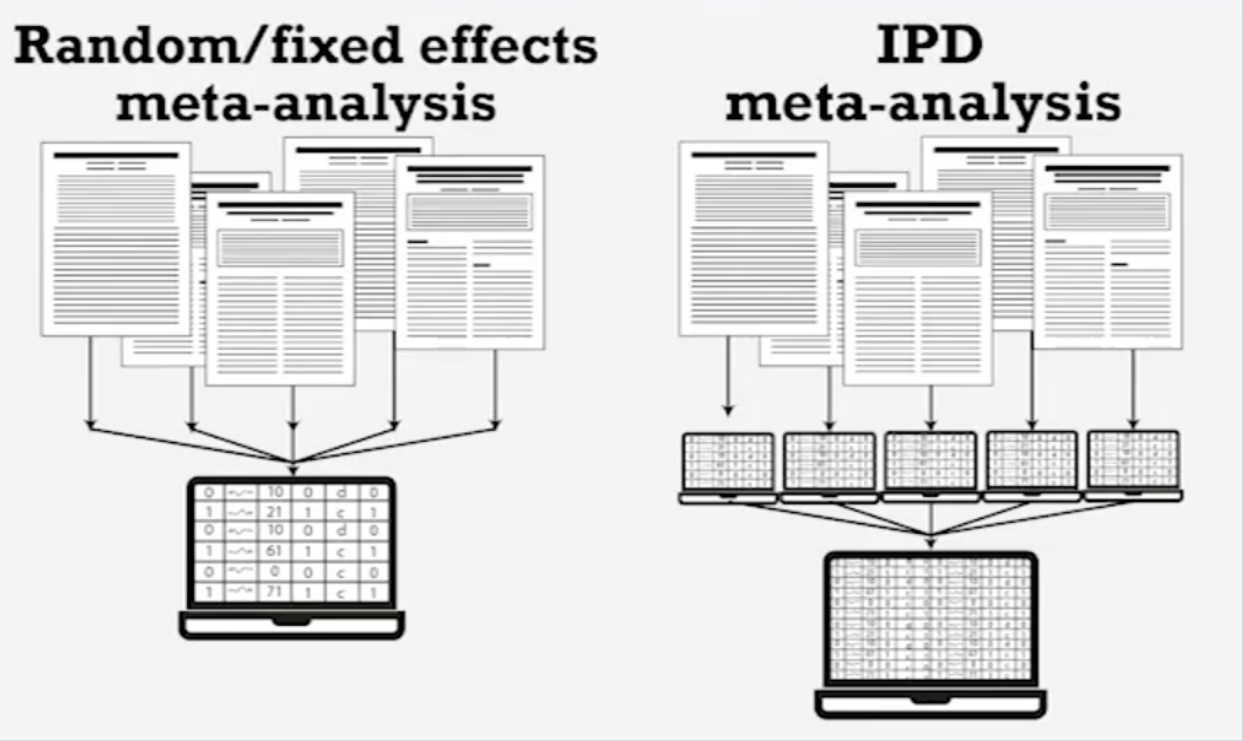
- Metaregressionen sind ein sehr mächtiges Werkzeug zur Beschreibung des Einflusses von Kovariablen auf die Ergebnisse von Studien.
- Gut etablierte Methodik im Bereich der statistischen Regression, die auf den speziellen Fall der Metaanalysen angewendet werden kann.

Individual Patient Data (IPD) Metaanalyse

IPD Metaanalyse

Idee der IPD Metaanalyse

Zugriff auf die individuellen Patientendaten (IPD), die allen Studien zu Grunde liegen



IPD Metaanalyse

Idee der IPD Metaanalyse

Zugriff auf die individuellen Patientendaten (IPD), die allen Studien zu Grunde liegen

- viele Vorteile $\rightarrow$ Goldstandard
- ermöglicht eine Re-analyse der Daten für jede einzelne Studie
- ist ein sehr kollaborativer Ansatz: Einbeziehung eines größeren Personenkreises mit individuellen Erfahrungen
- höhere Qualität: Kontrolle und Korrektur von Datensätzen ist möglich
- nur selten durchführbar

IPD Metaanalyse

Vorteile

IPD MA ist wünschenswert

- unveröffentlichte Studien können einbezogen werden
- besserer Umgang mit einem hohen Anteil ausgeschlossener Untersuchungsdaten
- Festlegung auf eigene, besser vergleichbare Endpunkte ist möglich (z.B. längeres Follow-up, time-to-event data)
- flexiblere Auswertung ist möglich

IPD Metaanalyse

Nachteile

besondere Herausforderungen bei IPD Metaanalysen

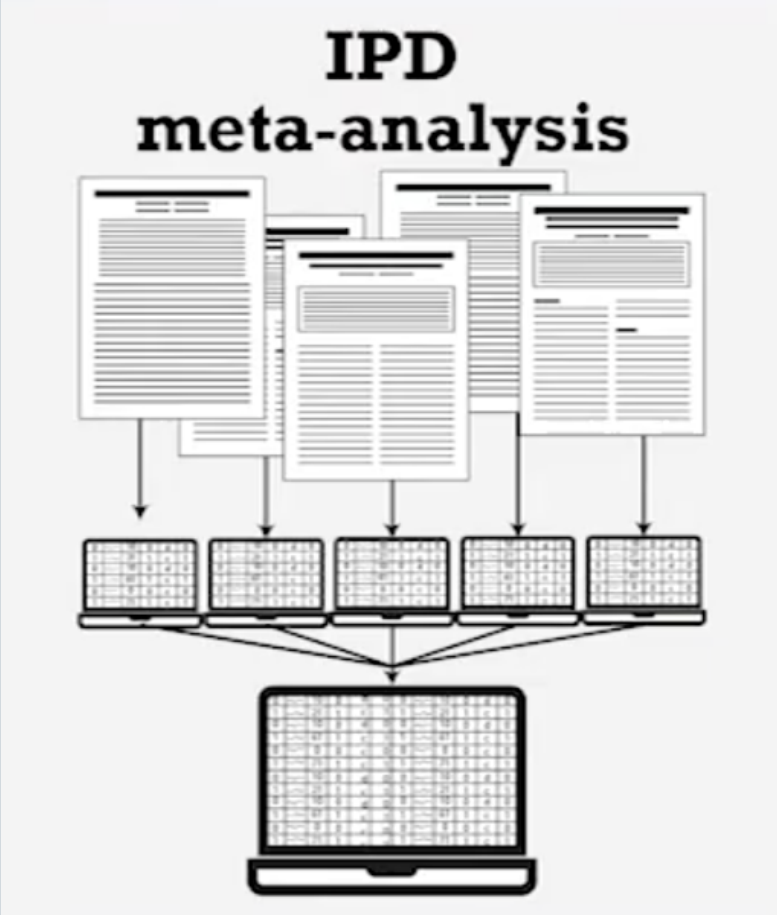
- langer, Ressourcen-intensiver Prozess
- Datenaustausch und Datensicherheit muss geregelt werden
- methodisch anspruchsvolle Auswertung
- Nichteinbeziehung einiger Studien kann einen Bias hervorrufen

IPD Metaanalyse

Auswertestrategie

zweistufige Auswertung:

1. Re-analyse der einzelnen Studien nach einem einheitlichen Protokoll
2. *fixed-* oder *random effects* Metaanalyse



IPD Metaanalyse

Auswertestrategie

- zweistufiges Auswerteverfahren ist gut etabliert und lehnt sich nah an eine klassische Metaanalyse an
- Untersuchung von Untergruppen oder Ereigniszeitanalyse erfordern die Anwendung von fortgeschrittenen Regressionsverfahren (hierarchische Modelle, bei denen die einzelnen Patienten den übergeordneten Studien zugewiesen sind)

Vorlesung "Grundlagen der Versuchsplanung"

innerhalb der VL-Reihe

"Biometrische Prinzipien und Methoden medizinischer Forschung"

Prof. Dr. rer. med. Ingo Röder

(*ingo.roeder@tu-dresden.de*)

Institut für Medizinische Informatik und Biometrie (IMB)
Medizinische Fakultät Carl Gustav Carus
TU Dresden

Outline

Literaturhinweise / Quellen

Kurz und knapp ...

- G. D. Ruxton & N. Colegrave: *Experimental Design for the Life Sciences*, Oxford Univ. Press (2. Auflage), 2003

Details und nachschlagen ...

- G. W. Cobb: *Introduction to Design and Analysis of Experiments*, John Wiley & Sons Ltd, 1998
- A. Dean & D. Voss: *Design and Analysis of Experiments*, Springer Verlag, 1999

Typische Fragen

... des Praktikers

- Wie viele Patienten/Probanden muss ich untersuchen?
- Wie setze ich ein gegebenes Budget (Fallzahl) optimal ein? D.h. *wo / was* messe ich?
- Was kann ich mit meinen verfügbaren Beobachtungen herausbekommen?

... des Biometrikers

- Was ist das Anliegen der Studie?
- Was (genau) ist die zu beantwortende Frage?

Beispiele typischer Zielstellungen

... im Rahmen (medizinischer) Studien

- Gibt es Unterschied zwischen verschiedenen Versuchsbedingungen?
  (z.B. Ist Therapie A besser als Therapie B?)
- Was sind bedeutsame Effektorvariablen
  (z.B. Risiko- oder prognostische Faktoren)?
- Wie groß ist der Effekt einer Einflussvariablen auf eine Zielgröße?
  (z.B. Welchen Einfluss hat das soziales Umfeld auf krankheitsbedingten Arbeitsausfall?)
- Welches sind die optimalen Versuchsbedingungen?
  (z.B. Suche nach Bedingungen für max./min. Effekt)
- Erstellen von Vorhersagen
  (z.B. Prognose der Hospitalisierungsdauer in Abhängigkeit des Gesundheitszustandes bei Einlieferung)

Die Kosten eines schlechten Designs (1)

Zwei weit verbreitete Versuchsplanungs-Mythen:

- Mythos 1: "Für die spätere Auswertung spielt es keine Rolle, wie die Daten im einzelnen erhoben werden."
  - → stat. Methoden haben Voraussetzungen, die erfüllt sein müssen
  - → Verletzung von Voraussetzungen ziehen z.T. Anwendung anderer Verfahren nach sich, die ggf. die eigentliche Frage nicht mehr beantworten lassen
  - → Erfassung der falschen Daten, lässt u.U. die Beantwortung der eigentlichen Frage gar nicht zu
- Mythos 2: "Wenn Du nur fleißig viele Daten sammelst, kommt schon etwas Interessantes dabei heraus."
  - → Quantität kann Qualität nicht ersetzen (*viel* Falsches wird *nicht richtiger*)
  - → Gefahr falsch positiver Resultate (siehe *Problem des multiplen Testens*)

Die Kosten eines schlechten Designs (2)

... bzgl. Zeit

- Landläufige Meinung:
  Nutze die Zeit für die Studie und verschwende keine Zeit mit der Planung

→ Wenn ein Experiment kein Resultat liefert, war alle Zeit, die dafür aufgewandt wurde, umsonst!
(Dies meint nicht: *kein positives Resultat*! Auch negative Resultate enthalten Informationen.)

"Eine klare Antwort auf eine Frage ist besser als vermutete (unklare) Antworten auf drei Fragen" (nach Ruxton & Colegrave, 2003)

Die Kosten eines schlechten Designs (3)

... bzgl. Geld

ähnlich wie bei der Ressource Zeit

- ein zu großes Experiment, kann die Verschwendung von Resourcen bedeuten
- **aber:** auch ein kleines (vermeintlich Ressourcen-sparendes) Experiment kann, wenn es zu keinem Ergebnis führt, Verschwendung von Ressources bedeuten

Die Kosten eines schlechten Designs (4)

... bzgl. ethischer Komponenten

Auch wenn Zeit und Geld "keine Rolle spielen" (sollten), gibt es (z.T. vielschichtige) ethische Gründe, die die Versuchsplanung beeinflussen:

- Tierversuche: Anzahl der Tiere und Art der Maniplutation sollte immer hinsichtlich der Notwendigkeit und Optimalität geprüft werden
- Klinische Studien am Menschen: nachgewiesene Vorteile einer Behandlung schließen bestimmte Designs aus
- Beobachtungsstudien am Menschen: Was wird beobachtet und welche Information wird weitergegeben/gespeichert etc.

Die Fragestellung

Versuchplanung beginnt bereits **vor** der Planung des Experiments!

Die Frage als Grundlage eines Experiments

- Jeder Studie muss eine klare Fragestellung zu Grunde liegen
- Diese Fragestellung muss weiterhin in eine wissenschaftliche Hypothese überführt werden, um sie (statistisch) untersuchen zu können
- Der Biometriker kann zwar bei Planung und Analyse einer Studie helfen, aber **nur** wenn der Anwender eine konkrete Fragestellung bzw. eine zu untersuchenden wissenschaftliche Hypothese formuliert!

"The ideas of experimental design deal with tuning a good question into a good experiment; they won't help you find the good question in the first place."

(George Cobb, 1998)

Wissenschaftliche Hypothese (1)

Definition

Eine **Hypothese** ist eine klare Beschreibung für eine potentielle Erklärung des betrachteten Phänomens, die es zulässt, überprüfbare Vorhersagen abzuleiten. → Bewertungskriterium

Bemerkungen

- Ohne eine Hypothese ist ein systematisches wissenschaftliches Arbeiten nicht möglich
- Eine Hypothese sollte möglichst präzise formuliert werden
- Zu *einem* Phänomen können im Allgemeinen *verschiedene* Hypothesen formuliert werden

Wissenschaftliche Hypothese (2)

Beispiel 1: Fehlende Hypothese

"Der Einfluss gesunder Ernährung auf das Wohlbefinden ist ein interessantes Forschungsgebiet. Es sollte besser untersucht werden."

Bemerkungen

- Warum ist dieses Forschungsgebiet interessant?
- Was ist(sind) die offene(n) Frage(n)?

$\rightarrow$ Obiges Statement (*"Der Einfluss ..."*) weißt ggf. auf Notwendigkeit einer Pilotstudie hin.

Wissenschaftliche Hypothese (3)

Beispiel 2: Unpräzise Formulierung der Hypothese.

"Gesunde Ernährung steigert das Wohlbefinden."

Bemerkungen

- Was ist mit "gesunder Ernährung" gemeint?
- Was ist unter "Wohlbefinden" zu verstehen?
- Auf welchen Personenkreis bezieht sich die Aussage?

$\rightarrow$ Zur wissenschaftlichen Bearbeitung bedarf es einer präzisen Formulierung der Begrifflichkeiten, insbesondere von Einfluss- und Zielgrößen.

Wissenschaftliche Hypothese (4)

Beispiel 3: Formulierung mehrerer möglicher Hypothesen / Auswahl

- *Ausreichendes Trinken verringert die Häufigkeit von Kopfschmerzen bei Jugendlichen im Alter von 14 - 20 Jahren.*
- *Regelmäßiger Genuss von Obst verringert die Infektanfälligkeit bei Rentnern und damit das allgemeine Wohlbefinden.*
- *Ernährung gemäß Empfehlung der Dt. Ges. f. Ernährung e.V. verbessert Wohlbefinden deutscher Arbeitnehmer gemessen anhand Fragebogen zum allgemeinen habituellen Wohlbefinden (FAHW).*

Bemerkungen

- Was ist das wissenschaftliche Anliegen?
- Erlaubt Hypothese Ableitung überprüfbarer Vorhersagen?

$\rightarrow$ Hypothesen sind nicht eindeutig und die Auswahl muss (zunächst) inhaltlich getroffen werden.

Der Weg zur Hypothese - Ein Beispiel

Verhaltensstudie bei Schimpansen

- **Ausgangsbeobachtung** (z.B. aus Pilotstudie): Die Aktivität der Schimpansen variiert sehr stark im Tagesverlauf
- **Fragestellung**: Was ist der Grund dieser Aktivitätsschwankungen?
- **Wissenschaftliche Hypothese** (*eine* mögliche): Das Fütterungsregime beeinflusst die Aktivität.
- **Testbare Vorhersage**: Der Zeitanteil, den ein Schimpanse damit verbringt herumzulaufen / zu klettern, ist in der Stunde vor den Fütterungszeiten höher als in den anderen Zeiten.

Auswahl informativer Vorhersagen - Ein Beispiel

Einfluss sportlicher Betätigung auf die Gesundheit

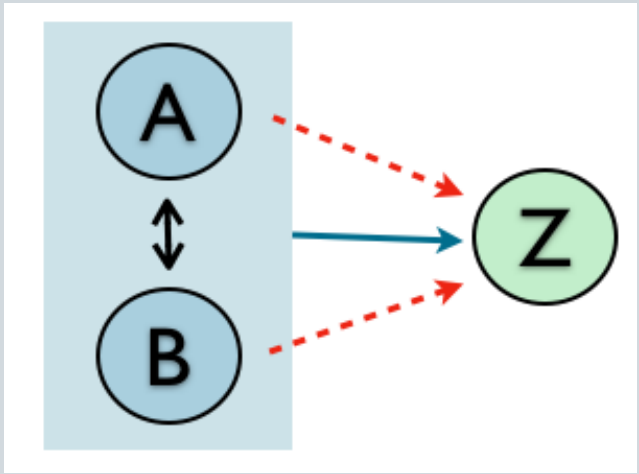
- **Wissenschaftliche Hypothese**: Aktive sportliche Betätigung verbessert den allgemeinen Gesundheitszustand.
- **Vorhersage (1)**: Mitglieder in Sportvereinen gehen weniger häufig zum Arzt als der Durchschnitt der Bevölkerung.
  → "schwache" (unspez.) Vorhersage, da viele Gründe für Zutreffen der Vorhersage in Frage kommen (z.B. Alter, Einstellung zum Sport)
- **Vorhersage (2)**: Teilung der Teilnehmer eines Vorsorgeprogrammes in 2 Gruppen: a) Teilnahme an regelmäßigen Aktivitäten (z.B. Laufen, Schwimmen), b) kein spezifisches Programm. In Gruppe 1 gibt es eine positivere Veränderung individueller Kenngrößen (z.B. BMI)
  → spezifischere Vorhersage: durch Intervention und anzunehmende Homogenität der Gruppen, liefert Test der Vorhersage mehr Informationen bzgl. der Hypothese

Confounding (1)

Begriff

- Durch Komplexität vieler Systeme sind multiple Abhängigkeiten und Wechselwirkungen einzelner Variablen eher die Regel
- Zwei Einflussvariablen (A, B) nennt man *confounded* (miteinander verwoben), wenn das exp. Design es nicht erlaubt, ihre möglichen Einflüsse (auf Zielgröße Z) voneinander zu isolieren.

inner bezogen auf experimentelles Design!

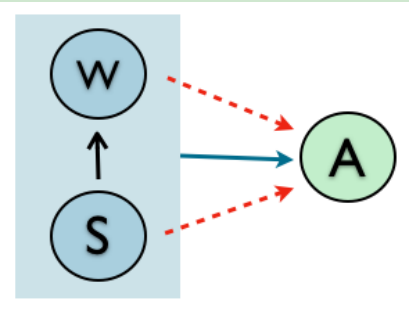


Confounding (2)

Beispiel: Analyse des Algenbefalls eines Naturbades

- **Beobachtungsstudie**: Messung von Algenbefall, Wassertemperatur und Sonneneinstrahlung zu verschiedenen Zeitpunkten
- **Resultat**: Stärkerer Algenwuchs ist sowohl verbunden mit höherer Wassertemperatur als auch mit intensiverer Sonneneinstrahlung
- **Aber**: Wassertemperatur von Sonneneinstrahlung beeinflusst

→ Wassertemperatur (W) und/oder die Sonnenscheindauer (S) sind aus diesen Beobachtungen nicht eindeutig als Grund des Algenwachstums (A) zu identifizieren. Sie sind *confounded*.



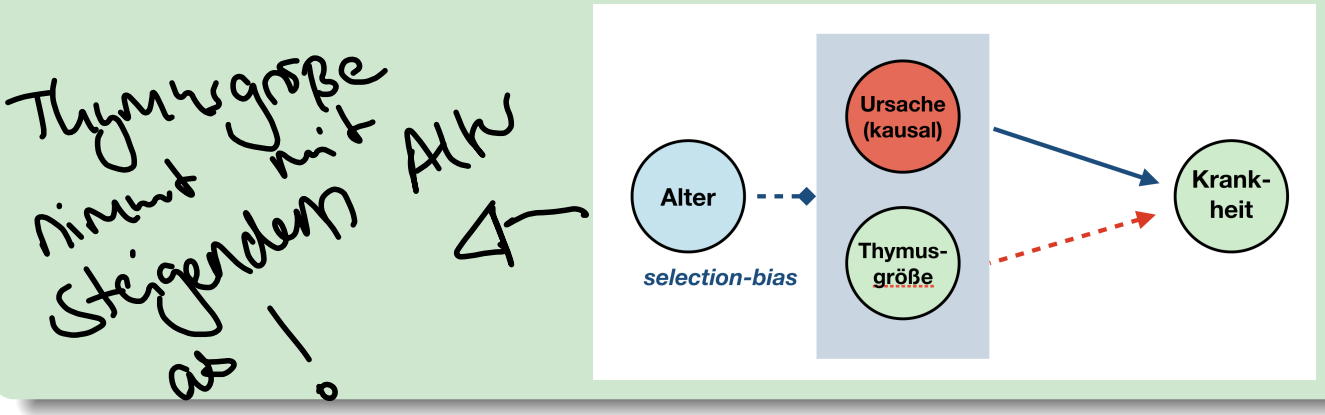
Confounding (3)

Auswahlverzerrung(*selection-bias*)-induziertes Confounding

- Confounding kann auftreten, wenn durch *selections-bias* unabhängige Zielgrößen miteinander verwoben (confounded) werden

Beispiel: Thymusentfernung bei Atemwegsproblemen

- Um 1900 gab es die Empfehlung, bei Atemwegsproblemen von Kleinkindern den Thymus zu entfernen
- Zugrunde liegende Beobachtung: Kleinkinder mit Atemwegsproblemen haben größeren Thymus als gesunde Erwachsene



Exploration vs. Konfirmation (Erkunden vs. Bestätigung)

Exploration

- Aufdeckung wichtiger (unbekannter) Varianzquellen
- Suche nach neuen (interessanten) Effekten

$\rightarrow$ Hypothesengenerierung

Konfirmation

- Bestätigung von postulierten Effekten bzw. deren statistische "Absicherung"

$\rightarrow$ Hypothesenprüfung

Beobachtung vs. Intervention (1)

Beobachtungsstudie

- Keine Intervention
- Eigenschaften werden beobachtet und dokumentiert (z.B. Körpergröße, Ausprägung von Symptomen)

Beispiel: Qualitätskontrolle in Klinikumslaboratorien

- Kontinuierliches Monitoring der Probenqualität in Labor A, B, C zeigt eine klar niedrigere Qualität in Labor A

→ Was ist der Grund?

  - Personal?
  - Probenqualität? (A, B, C erhalten Probem aus verschiedenen Kliniken)
  - Apparative Laborausstattung? (A, B, C nutzen verschiedene Fabrikate)

Beobachtung vs. Intervention (2)

Interventionsstudie

- Gezielte Einflussnahme / Intervention (z.B. neue Therapie)
- Einige Eigenschaften werden zugewiesen (z.B. neue Therapie), andere Eigenschaften werden beobachtet und dokumentiert

Beispiel: Studie zum Vergleich Placebo vs. Medikament

- Anwendung von Medikament und Placebo auf *gleichartige* Patienten (z.B. erreicht durch Randomisierung)
- Durch Ausschluss anderer Varianzquellen kann bei einem Effektunterschied davon ausgegangen werden, dass er vom Medikament verursacht wurde.

Beobachtung vs. Intervention (3)

Beobachtungsstudie

Vorteile

- Einfacher, weniger aufwendig
- Keine externe "Beeinflussung"
- *natürliche* Varianz

Nachteile

- Kausalitäten nicht nachweisbar
- Unkontrollierte Effekte
- Gefahr des Confounings

Interventionsstudie

Vorteile

- gezielte Kontrolle ausgewählter Einflussgrößen
- Kausalitäten u.U. nachweisbar

Nachteile

- Oft aufwendiger
- Gefahr "unnatürlicher" Effekte
- Intervention z.T. nicht möglich

Pilotstudien

... als Spezialfall explorativer Studien

- Beobachtungsstudien zur Hypothesengenerierung
- Interventionsstudien zur Auswahl informativer Interventionen (z.B. mit geringer Fallzahl)

... als Vorbereitung einer konkreten konfirmatorischen Studie

- Überprüfung der Sinnhaftigkeit der Fragestellung
- Überprüfung der experimentellen bzw. der Messtechnik
- (Ab-)Schätzung der Verteilungseigenschaften (z.B. Form, Varianz) der Daten (siehe später: Fallzahlplanung)

Variabilität

Variabilität ...

- ... innerhalb (empirischer) Daten ist eine unvermeidliche Tatsache
- ... ist der Hauptgrund für die Anwendung statistischer Methoden (in Planung und Auswertung von Versuchen)
- ... kann durch geeignete Planung und Analyse von Experimenten/Studien so zerlegt werden, dass interessierende (reale) Effekte und ungewollte (zufällige) Einflüsse anhand ihrer Varianzkomponenten unterschieden werden können

Quellen von Variabilität

Es gibt drei Quellen von Variabilität - eine *gewollte* und zwei *ungewollte*

1. Variabilität durch die zu untersuchenden Bedingungen (*gewollt*)
2. Durch den Messprozess verursachte Variabilität (*ungewollt*)
3. Variabilität im Untersuchungsmaterial bzw. im Untersuchungsprozess (*ungewollt*)

Untersuchung des Einflusses akustischer Reize auf die Aufmerksamkeit von Probanden anhand von EEG Messungen

1. Variabilität der EEG Werte bei verschiedenen Reizen (*gewollt*)
2. (Technische) Variabilität der EEG Messungen (*ungewollt*)
3. Variabilität des Ansprechens individueller Probanden (*ungewollt*)

Arten von Variabilität (1)

Um Strategien zum Umgang mit ungewollter Variabilität und damit zum Design guter Experimente entwickeln zu können, ist es notwendig die verschiedenen "Verhaltensformen" von Variabilität zu kennen ...

Es gibt drei Arten von Variabilität (die u.U. gemeinsam auftreten können)

1. Geplante, systematische Variabilität (*gewollt*)
2. Zufällige, nicht-systematische Variabilität (*ungewollt, aber prinzipiell unkritisch*)
3. Ungeplante, systematische Variabilität (*ungewollt, aber gefährlich, da sie Ergebnisse grundlegend zu verfälschen vermag!*)

Arten von Variabilität (2)

nicht-systematisch

Wissen um diese drei Kategorien (geplant-systematsich / ungeplant-zufällig / ungeplant-systematisch) kann genutzt werden um:

- Ungeplante, systematische Variabilität zu identifizieren und sie ggf. in tolerierbare, zufällige bzw. in geplante, systematische Variabilität zu überführen
- Zufällige Variabilität zu minimieren (und damit genauere Aussagen zu gewinnen)

z.B. durch genauere Zufalls/ statistische Modelle

Geplante, systematische Variabilität

Induziert durch ...

- ... den Experimentator bzw. Planer der Studie, um Unterschiede zwischen interessierenden Bedingungen (Ziel der Studie) aufzudecken.

Wenn systematische Variabilität geplant ist, ...

- ... kann man *vor* Durchführung der Studie Beobachtungen mit erwarteter Ähnlichkeit zuordnen; z.B. durch *gezielte Variation* der Bedingungen oder *Auswahl* der Messungen/ Beobachtungsobjekte

↳ Blockbildung

Beispiel: Überleben bei Bronchialkarzinom unter Therapie A und B

- *Gezielte Zuordnung* von Patienten in Therapiearm A bzw. B

Beispiel: Vergleich einer Konzentrationsübung auf Seh- bzw. Hörvermögen

- *Gezielte Auswahl* der Messmethode bzw. des Sinnesorgans

Zufällige, nicht-systematische Variabilität (1)

Induziert z.B. durch ...

- ... individuelle Unterschiede der (lebenden) Untersuchungsobjekte, oft auch als inter-individuelle Variabilität, "within-treatment" Variabilität oder als *"noise"* bezeichnet
- ... Messfehler

Zufälligkeit dadurch charakterisiert, dass ...

- ... sich diese Art der Variation wie die Ziehung von Zufallszahlen verhält, d.h.
  - ... einzelner Versuchsausgang kann nicht (bzw. nur im Sinne einer Wahrscheinlichkeitsaussage) vorhergesagt werden
  - ... erwartetes Verhalten bei einer Vielzahl von Beobachtungen (z.B. Mittelwert) kann bestimmt / vorhergesagt werden

Zufällige, nicht-systematische Variabilität (2)

Eigenschaften des Zufalls"fehlers"

- Nivellierung des Fehlers durch seinen Zufallscharakter (siehe auch: Gesetz der großen Zahlen)
- Möglichkeit der Schätzung der Größe des Zufallsfehlers (bei korrektem Design der Studie)

Ungeplante, systematische Variabilität

Induziert z.B. durch ...

- ... unbekannte Störgrößen bzw. Selektioneffekte, die eine *systematische Verzerrung* (*Bias*) der Ergebnisse verursachen

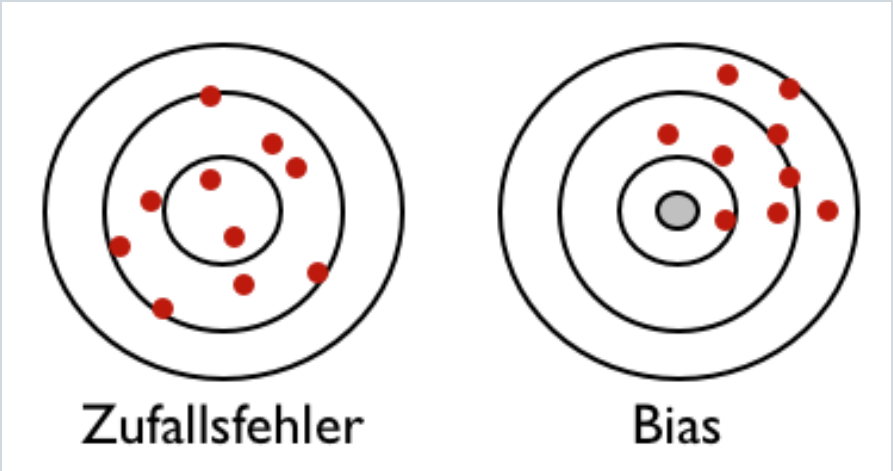
Auswirkungen eines unbekannten Bias ...

- ... kann zu falschen Schlussfolgerungen führen
- ... kann u.U. die gesamte Studie unbrauchbar machen

Zufällige vs. systematische (ungeplante) Variabilität

Unterschiede von Zufallsfehler und Bias

- Sowohl Bias als auch Zufallsfehler führen zu Abweichungen der Beobachtung vom "wahren" Wert
- Allerdings mitteln sich diese Abweichungen im Rahmen vieler Beobachtungen beim Zufallsfehler aus, wohingegen dies beim Bias nicht der Fall ist



Ziele der Versuchsplanung

Primäre Ziele der in dieser VL diskutierten Methoden:

1. Vermeidung/Minimierung ungeplanter, systematischer Variabilität (z.B. Bias, Confounding)
2. Minimierung zufälliger, nicht-systematischer Variabilität

$\rightarrow$ so dass die Variabilität, die durch die interessierenden Faktoren induziert wird, besser identifiziert werden kann

3. Effiziente Schätzung spezifische Effekte, wie z.B. von Wechselwirkungen einzelner Einflussvariablen

Weitere Ziele der optimalen Versuchsplanung

- Verbesserung der Güteeigenschaften von Schätzern bzw. der Erhöhung der Power von Testverfahren durch geeignete Wahl des Studien-/Experiment-Designs

Designprinzipien zur Umsetzung der Ziele

Verringerung des Zufallsfehlers

→ Replikation

Vermeidung / Minimierung ungeplanter systematischer Verzerrungen

→ Randomisierung

→ Blockbildung (blocking)

Effizienter Vergleich von mehr als zwei Versuchsbedingungen

→ Faktorielle Kombination

Replikation (1)

Wiederholte Messungen/Beobachtungen

- ... sind eine Möglichkeit den Zufallsfehler zu kontrollieren
- ... können die Genauigkeit der Schätzung (das "Vertrauen in den Schätzer") erhöhen
- ... ermöglichen den Zufallsfehler zu quantifizieren (siehe Standardfehler als Varianz des Schätzers, VL Biometrie)

Replikation (2)

→ Replikate nicht unabhängig

Beispiel 1: Gibt es eine Abhängigkeit des Kariesrisikos vom Geschlecht?

- Zielstellung: Vergleich des Befalls von Männern und Frauen mit Kariesbakterien (*Steptococcus mutans*)
- Messungen: Erhebung des Bakterienstatus anhand von je 5 separaten Speichelproben eines Mannes und einer Frau (n = 2 x 5), die jeweils mittels unabhängiger PCR untersucht wurden

Ist die gestellte Frage mit diesem Design zu beantworten?

→ Replikation erfolgt immer mit den selben zwei Personen

→ Biologische Variabilität kann nicht abgebildet werden

→ technische Variabilität ist aber gegeben (durch 5-malige Replikation)

Replikation (3)

Beispiel 2: Servicequalität von Krankenkassen

- Zielstellung: Vergleich der Servicequalität in Abhängigkeit vom sozialen Status der Versicherten
- Messungen: Qualität der Antwort auf Beschwerdebriefe verschiedener Personen (ein Student, ein Arbeitsloser, ein Angestellter, ein freiwillig versicherter Selbstständiger) an jeweils 6 (zufällig ausgewählte) Krankenkassen

Ist die gestellte Frage mit diesem Design zu beantworten?

Nein, siehe Beispiel 1 (Folie 36)
- alternatives Studiendesign
  - Verwendung eines Computer gesetzten Standardbriefes der sich nur durch den angegebenen Berufsstatus unterscheidet
  - Damit Vermeidung eines Confounding mit persönl. Infos

Replikation (4)

Bemerkungen:

- Replikation als Mittel zur Berechnung des Zufallsfehlers benötig **echte** Replikate, d.h. Messungen/Beobachtungen müssen (statistisch) **unabhängig** voneinander sein
- Vorsicht vor sogenannten *Pseudoreplikaten* (siehe Beispiele 1 und 2)

→ Verwendung abhängiger Replikate (d.h. von Pseudoreplikaten) zur Schätzung der Varianz liefert fehlerhafte Resultate!

Replikation (5)

(realistischeres) Beispiel zu Pseudoreplikaten:
Bestimmung der Effektivität einer neuen, kostengünstigeren Methode der Versorgung von Knochenbrüchen im Tierversuch

- Zielstellung: Vergleich der residualen Spaltbreite 3 Wochen nach Versorgung des Bruches mit alter und neuer Methode
- Messungen: Erhebung der Spaltbreite in je 3 Tieren pro Methode; pro Bruch (d.h. pro Tier) drei Messungen, d.h. 9 Messungen pro Methode

Kann man die Variabilität der Messungen innerhalb der einzelnen Methoden auf der Basis dieser 9 Replikate korrekt bestimmen?

Nein, da jeweils 3 der Replikate bzgl. des Tiere nicht unabhängig sind und damit die biologische Variabilität nicht geschätzt wird
→ es liegen nur 3 Replikate vor

Replikation (6)

Häufige Ursachen für Pseudoreplikaten

- Verwandschaftsbeziehungen
- Mehrfachmessungen an identischen Objekten
- Zeitverlaufsmessungen
- Gemeinsamer Kontext von Beobachtungen (z.B. soziales Umfeld, Einzugsgebiet von Kliniken)
- Induzierte Pseudoreplikation (z.B. identischer Stimulus für verschiedene, andernfalls unabhängige Probanden innerhalb der Versuchsgruppen $\rightarrow$ mögliches Confounding des Stimulus mit unbekannten Effekten)

Replikation (7)

Umgang mit Pseudoreplikaten

- Mittellung der "inneren" Replikate (einfach, aber nicht optimal)
- Vermeidung von Pseudoreplikaten durch geeignete Versuchsplanung (erfordert u.U. größeren Aufwand)
- Anwendung geeigneter statistischer Modelle, die die Abhängigkeitsstruktur der Messungen korrekt berücksichtigen; erfordert geeigneten Versuchsplan (siehe Blockbildung) und ggf. Unterstützung durch Biometriker

Randomisierung (1)

Begriff

- Randomisierung bedeutet die *zufällige* Zuordnung von Objekten zu experimentellen/ Studiengruppen
- D.h. jedes Objekt hat die gleiche Wahrscheinlichkeit für eine bestimmte Gruppe ausgewählt zu werden

Effekte einer Randomisierung

- Randomisierung konvertiert ungeplante, systematische Variabilität in zufällige, nicht-systematische Variabilität (der man durch Replikation begegnen kann)
- Ausschluss oder Verminderung von Confounding / Bias
- (weitgehende) Gewährleistung der Strukturgleichheit von Vergleichsgruppen

Randomisierung (2)

Beispiele möglicher Fehler bei Randomisierung

- Willkürliche (aber nicht zufällige) Zuweisung
- "Selbstauswahl"

Beispiel ("Willkür"): Zählung von Zellen in unterschiedlichen Kulturen

- Mittels Zufallszahlen werden die Kulturen zugewiesen
- Um effizient zu arbeiten werden zunächst die Zellen in Gruppe 1 gezählt und danach die der Gruppe 2

$\rightarrow$ Vorsicht: Zeit könnte Confounder sein; z.B. Ermüdung

Beispiel ("Selbstauswahl"): Bevölkerungsbefragungen

- Einschluss aller bereitwilligen Passanten

$\rightarrow$ Die Bereitschaft teilzunehmen könnte selbst einen Einfluss haben

Blockbildung (1)

Begriff

- Zusammenfassung *ähnlicher* Objekte zu Blöcken und damit Einführung von Blockvariablen
- Zuweisung der Objekte zu den Versuchsbedingungen separat in den Blöcken

Effekte der Blockbildung

- Blockbildung konvertiert ungeplante, systematische Variabilität in geplante, systematische Variabilität
- Blockbildung erlaubt die Aufteilung der Zwischen-Objekt/Subjekt Varianz in "within-block" und "between-block" Varianz; dies erlaubt effizientere Schätzung des Gruppeneffektes innerhalb der Blöcke
- Blockbildung macht potentielle Confounder/ Biasquellen explizit und kann dadurch Fehlinterpretationen verhindern

Blockbildung (2)

Beispiel: Einfluss eines Rehabilitationsprogrammes auf Hospitalisierungsdauer

- Hypothese: Teilnahme an Reha-Programm senkt Hospitalisierungsdauer
- Studienkollektiv: Alle Patienten mit bestimmter Diagnose/Behandlung
- Es ist bekannt, dass das Alter bei Diagnose einen erheblichen Einfluss auf Hospitalisierungsdauer hat (unabhängig von Reha)

$\rightarrow$ vollständige Randomisierung auf die Gruppen Reha ja / nein ist ggf. nicht optimal, da das Alter eine zusätzliche Variabilität induziert, die einen ggf. vorhanden Effekt verdecken könnte und derer Einfluss durch die Randomisierung u.U. nicht vollständig vermieden werden kann

$\rightarrow$ sinnvoll: Verwendung von Altersgruppen als Blöcke, innerhalb derer dann randomisiert wird

Blockbildung (3)

Nachteile der Blockbildung

- Blockbildung hinsichtlich von Variablen, die keinen Effekt auf Zielgröße (d.h. die kaum Varianz erklären) haben, führen zu einem Verlust an statistischer Power!
  $\rightarrow$ Blockbildung nur dann anwenden, wenn es hinreichend Evidenz für Effekt der Blockvariable gibt

- Blockbildung kann bei geringen Fallzahlen dazu führen, dass einige Blöcke innerhalb der einen oder anderen Behandlungsgruppe leer bleiben. Dies führt zu schwerer handhabbaren "unvollständigen" Versuchsplänen
  $\rightarrow$ Nutzen der Blockbildung steigt im Allgemeinen mit der Fallzahl

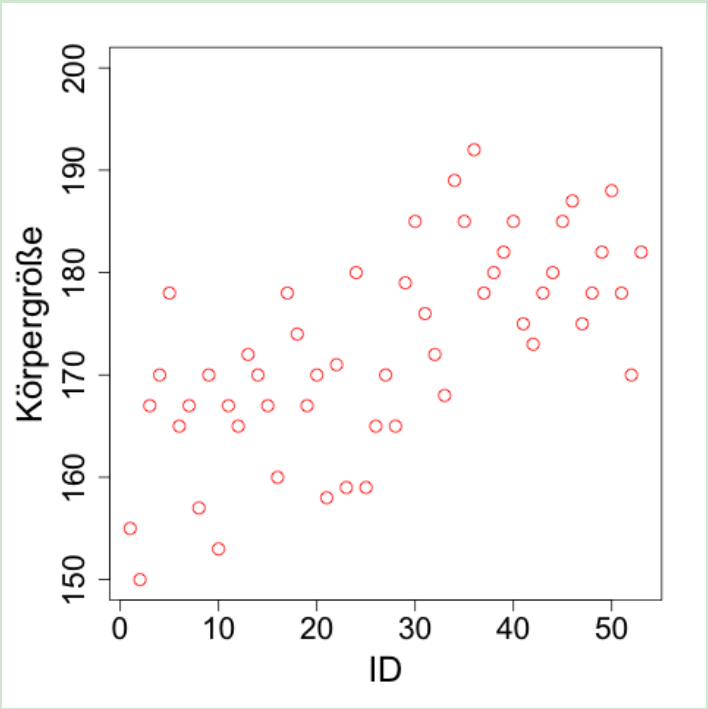
Blockbildung (4)

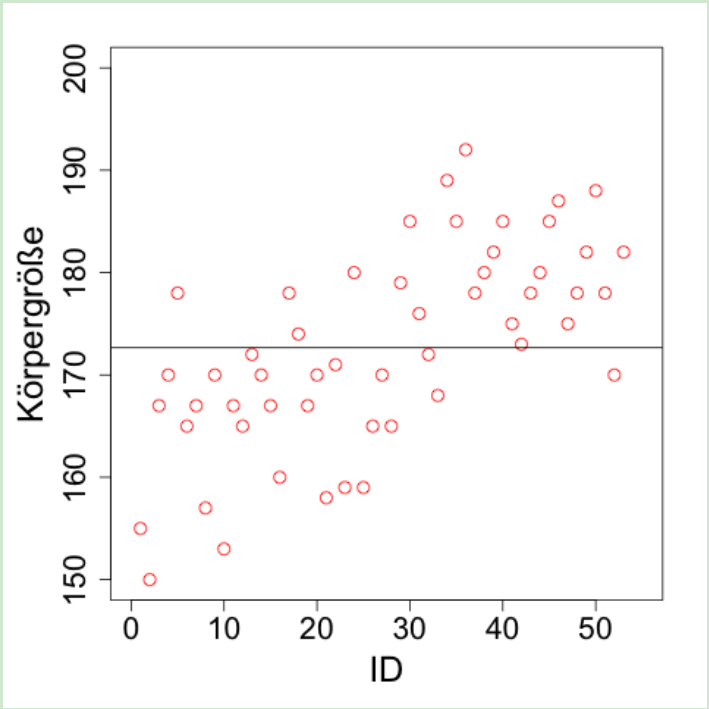
Bemerkungen

- Gepaarte Designs (vgl. gepaarter t-Test, VL Biometrie) sind ein Spezialfall eines Block-Designs
- Wenn eine potentielle Block-Variable als kontinuierlicher Messwert vorliegt (siehe Alter im Reha Bspl.), dann kann eine zur Blockbildung notwendige Kategorisierung u.U. effizienter durch Betrachtung der potentiellen Block-Variablen als Covariable ersetzt werden. Dies erfordert Anwendung anderer Analyseverfahren (z.B. ANCOVA).

Wiederholung des ANOVA-Prinzips am Beispiel: Einfluss des Geschlechts auf die Körpergröße (1)

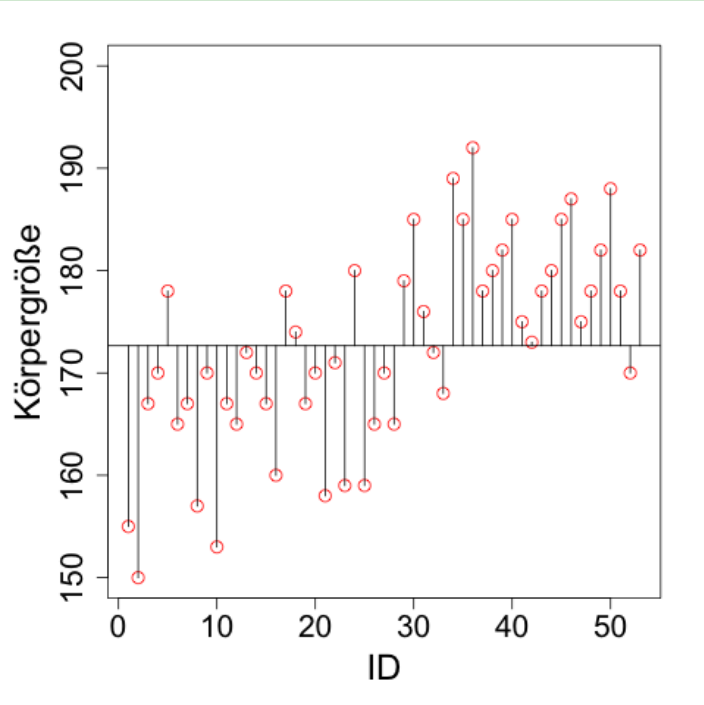
Rohdaten + Gesamtmittel





Wiederholung des ANOVA-Prinzips am Beispiel: Einfluss des Geschlechts auf die Körpergröße (2)

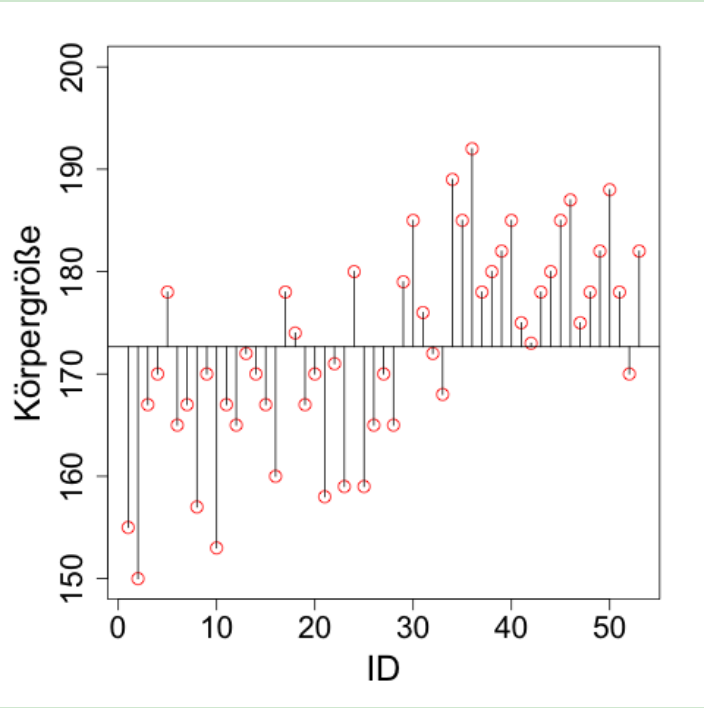
Die Gesamtvarianz (total sum of squares - SS_{total})

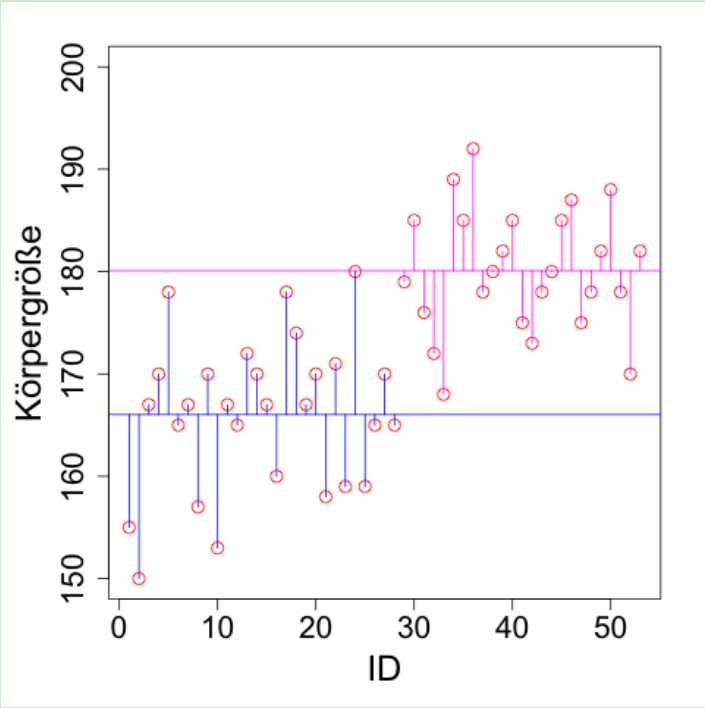


- Betrachte Abstand der einzelnen Messwerte vom Gesamtmittel
- Summiere diese Abstände auf: $SS_{total} = \sum(y - \bar{y})^2$
- Bei Division durch $n - 1$ (Freiheitsgrade) ist dies der bekannte Schätzer der Varianz, d.h. die mittlere quadratische Abweichung, MS_{total}

Wiederholung des ANOVA-Prinzips am Beispiel: Einfluss des Geschlechts auf die Körpergröße (3)

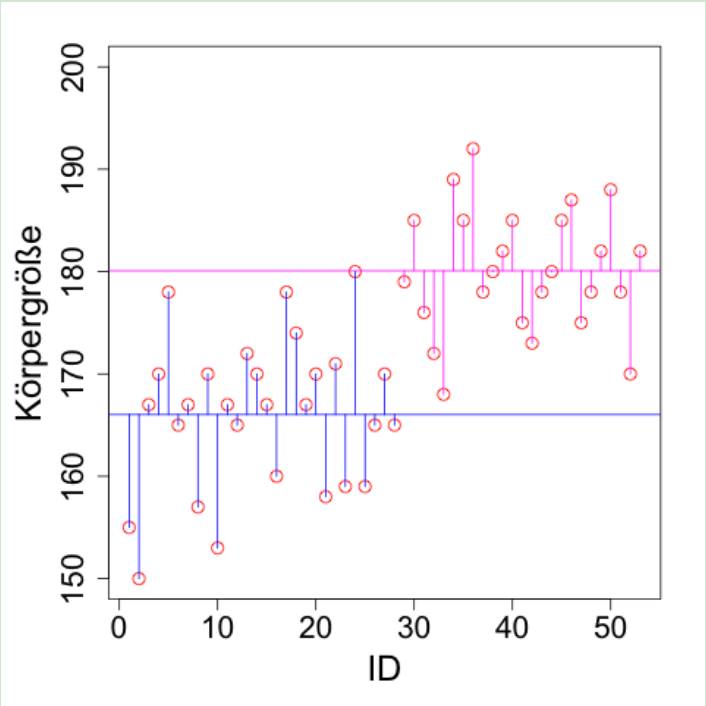
Varianz innerhalb der Faktorstufen





Wiederholung des ANOVA-Prinzips am Beispiel: Einfluss des Geschlechts auf die Körpergröße (4)

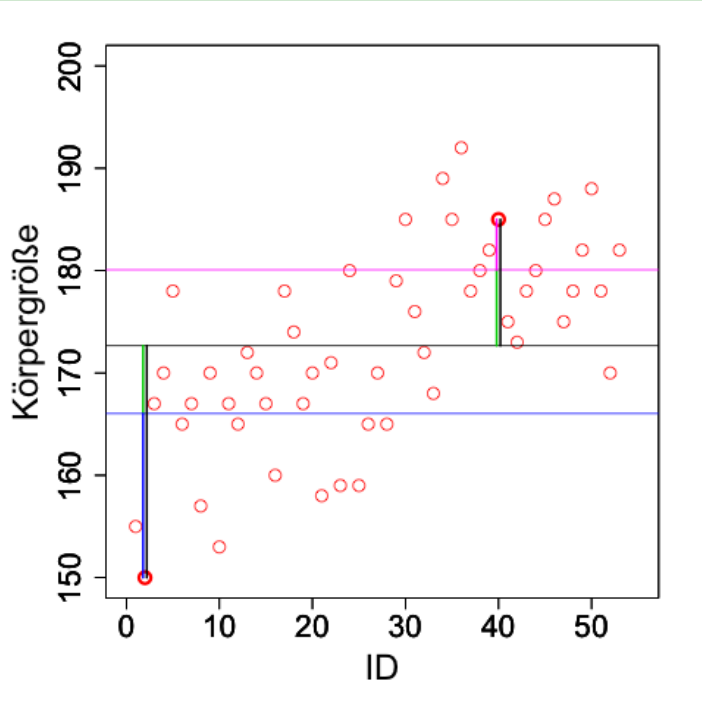
Die Innergruppen-Varianz (within group sum of squares - SS_{within})



- Betrachte Abstand der Messwerte vom jeweiligen Gruppenmittel ($\bar{y}_j,\ j = 1, \ldots, k;$ im Bspl. $k = 2$)
- Summiere Abstände auf: $SS_{within} = \sum_{j=1}^{k} \sum (y - \bar{y}_j)^2$
- Division durch $n - k$ (Freiheitsgrade) ergibt Schätzer der Inner-Gruppen-Varianzen, d.h. der mittleren quadratischen Abweichungen innerhalb der Gruppen MS_{within}

Wiederholung des ANOVA-Prinzips am Beispiel: Einfluss des Geschlechts auf die Körpergröße (5)

Varianz zwischen den Faktorstufen



- Gesamtvarianz wird aufgespalten in Varianz-komponenten: **innerhalb** bzw. **zwischen** den Gruppen
- Die Varianz zwischen den Gruppen ($SS_{between}$) ergibt sich als Differenz der Gesamtvarianz und der Varianz innerhalb der Gruppen, d.h. $SS_{between} = SS_{total} - SS_{within}$
- Division durch $k - 1$ (FG) ergibt Schätzer der Zwischen-Gruppen-Varianz $MS_{between}$

Wiederholung des ANOVA-Prinzips am Beispiel: Einfluss des Geschlechts auf die Körpergröße (6)

Test des Geschlechtseffektes auf Körpergröße

- Dazu Vergleich der geschätzten Varianzen zwischen Faktorstufen (*d.h. durch Geschlecht erklärte Varianz*) mit denen innerhalb der Faktorstufen (*Restvarianz innerhalb der Frauen bzw. Männer*)
- Wenn Varianz zwischen Faktorstufen ($MS_{between}$) deutlich größer ist als Restvarianz (MS_{within}), heißt dies, dass der jeweilige Faktor (hier Geschlecht) einen Einfluss auf die Zielgröße (hier Körpergröße) hat
- Formal Prüfung durch *F-Test* ($H_0 : MS_{between} = MS_{within}$ vs. $H_A : MS_{between} > MS_{within}$):

$$F = \frac{MS_{between}}{MS_{within}} \overset{H_0}{\sim} F(k-1, n-k)$$

Varianzanalyse im Kontext Linearer Modelle (1)

Allgemeine Schreibweise eines varianzanalytischen Modells mit einer Einflussgröße (einfaktorielle bzw. One-way ANOVA)

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

μ ... Gesamtmittel; α ... Faktor; $\epsilon \sim \mathcal{N}(0, \sigma^2)$... Residuum ("Restfehler"), i ... Anzahl Faktorstufen, j ... Anzahl Beobachtungen

Beispiel: Einfluss Geschlecht auf Körpergröße

$$\textit{Körpergröße}_{ij} = \textit{Gesamtmittel} + \textit{Geschlecht}_i + \epsilon_{ij}$$

$i = 2$ (z.B. 1: weiblich, 2: männlich); j ... Anzahl Beobachtungen

Varianzanalyse im Kontext Linearer Modelle (2)

Verallgemeinerung auf 2 und mehr Einflussgrößen

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijkl}$$

...

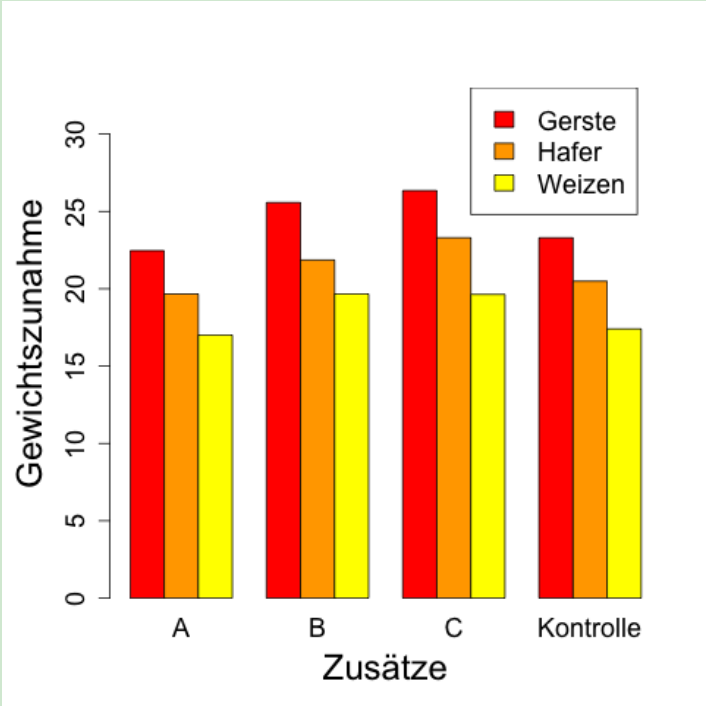
Einbeziehung von Wechselwirkungen

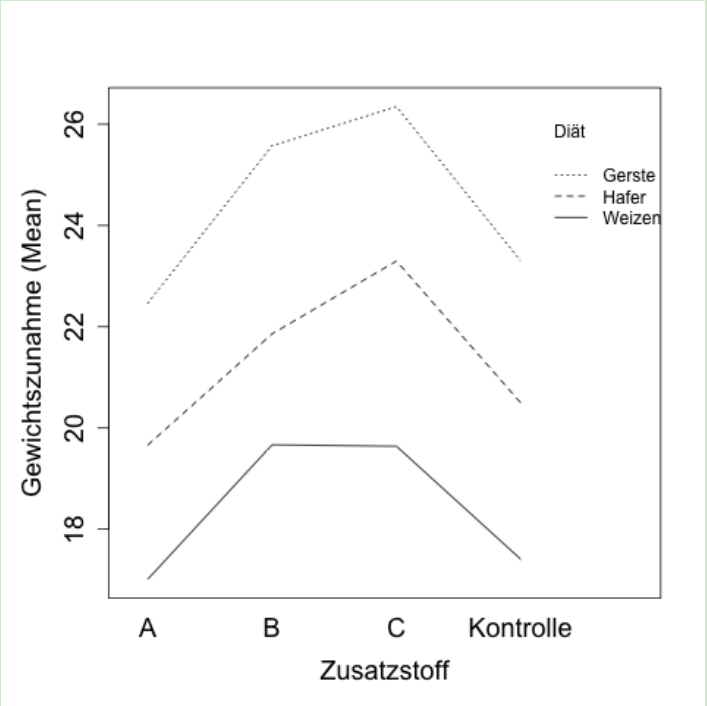
$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

$$y_{ijkl} = \mu + \alpha_i + \beta_j + +\gamma_j + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl}$$

Wiederholung: Interpretation von Wechselwirkungen (1)

Beispiel: Einfluss von Getreidesorte und Zusatzstoffen auf Gewichtszunahme



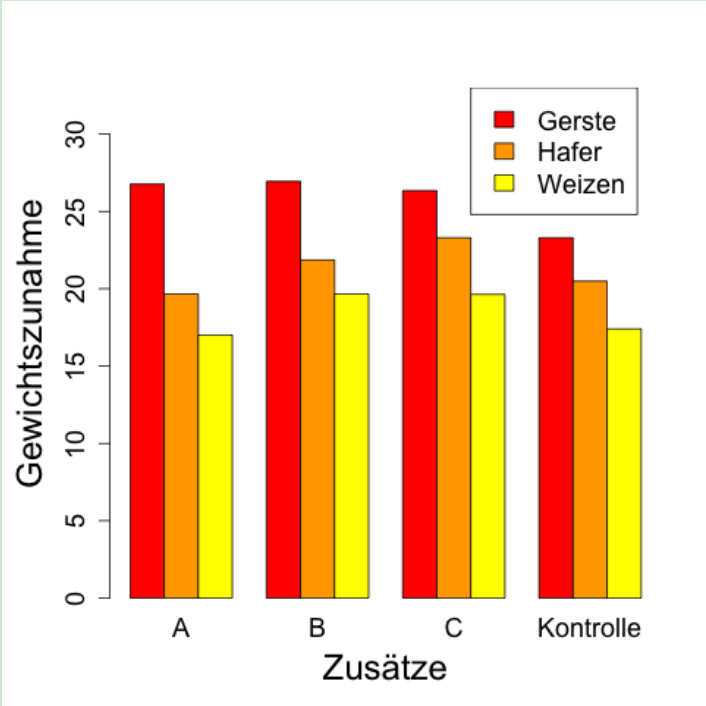


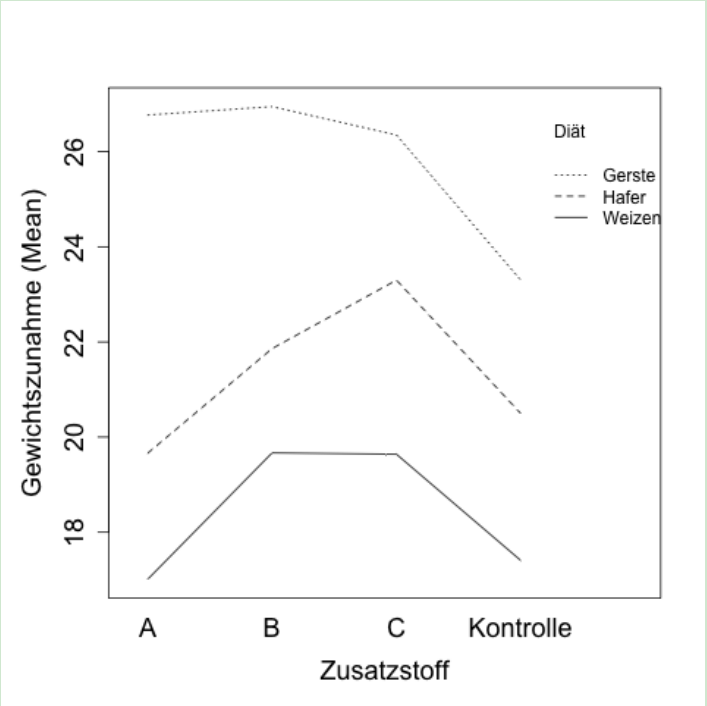

```
> summary(aov(Gewichtszunahme~Diät*Zusatzstoff))
                 Df  Sum Sq Mean Sq F value    Pr(>F)
Diät              2 287.171 143.586 83.5201 2.999e-14 ***
Zusatzstoff       3  91.881  30.627 17.8150 2.952e-07 ***
Diät:Zusatzstoff  6   3.406   0.568  0.3302    0.9166
Residuals        36  61.890   1.719
```

Bemerkung: Schreibweise *Diät*Zusatzstoff* ist Kurzschreibweise für *Diät+Zusatzstoff+(Diät:Zusatzstoff)+*ϵ

Wiederholung: Interpretation von Wechselwirkungen (2)

Beispiel: Einfluss von Getreidesorte und Zusatzstoffen auf Gewichtszunahme






```
> summary(aov(Gewichtszunahme~Diät*Zusatzstoff))
                Df Sum Sq Mean Sq  F value    Pr(>F)
Diät             2 446.04 223.021 113.9110 2.691e-16 ***
Zusatzstoff      3  61.19  20.396  10.4174 4.466e-05 ***
Diät:Zusatzstoff 6  28.82   4.803   2.4531   0.04313 *
Residuals       36  70.48   1.958
```

Bemerkung: Schreibweise *Diät*Zusatzstoff* ist Kurzschreibweise für *Diät+Zusatzstoff+(Diät:Zusatzstoff)+*ϵ

Vollständig randomisiertes 1-Faktor Design (1)

(one-way basic factorial design, BF[1])

Realisierung eines BF[1] Designs

- Im Falle einer interventionellen Studie ordnet es die Beobachtungen zufällig den Stufen des Faktors zu
- Im Falle einer Beobachtungsstudie zieht man zufällige Stichproben aus den jeweiligen Beobachtungspopulationen

Vor- und Nachteile des BF[1] Designs

Vorteile:
- Einfach zu realisieren, analysieren und interpretieren

Nachteile:
- Betrachtet jeweils nur einen Faktor (univariate Analyse)
- U.U. nicht effizient, da es Inhomogenitäten (z.B. durch Counfounder Variablen) nicht spezifisch Rechnung tragen kann

Vollständig randomisiertes 1-Faktor Design (2)

(one-way basic factorial design, BF[1])

Beispiel: BF[1] in interventioneller Studie

- Untersuchung von 2 neuen physiotherapeutischen Behandlungsstrategien (A und B), sowie einer Kontrolle (d.h. Standardbehandlung, C) in der Rehabilitation nach Oberschenkelfrakturen bezüglich der Dauer des stationären Aufenthaltes
- Alle in die Studie eingeschlossenen Patienten werden zufällig auf die 3 Gruppen verteilt:

Gruppe	Datenpunkte (Stationärer Aufenthalt in Tagen)
A	15 17 21 16 13 14 21 18 16 14
B	16 12 18 18 19 15 16 20 17
C	18 21 19 14 24 17 18 20 17 16

Vollständig randomisiertes 1-Faktor Design (3)

(one-way basic factorial design, BF[1])

Beispiel: BF[1] in Beobachtungsstudie

- Erhebung und Vergleich des Hygienestatuts von drei Großküchen (I, II, III) anhand der Anzahl Keim-belasteter (*positiver*) Proben; als *positiv* gilt eine Probe, wenn in ihr eine vorgegebene Keimzahl von Enterobakterien überschritten ist
- Bestimmung der Anzahl *positiver* Proben (aus jeweils 20 Proben) in je 8 zufällig ausgewählten belieferten Einrichtungen:

Gruppe	Datenpunkte (pos. Test)
I	0 1 0 2 0 2 3 0
II	0 0 1 1 0 3 0 0
III	1 1 3 0 2 0 1 0

Balancierung

Begriff

- Ein Design, welches allen Faktorstufen die gleiche Anzahl von Beobachtungen zuweist, nennt man *balanciert*; andernfalls heißt das Design *unbalanciert*

Bemerkungen

- Wenn möglich sollte eine balancierte Form des Designs angestrebt werden, da diese leichter zu analysieren ist bzw. bessere statistische Eigenschaften aufweist. Balancierte Designs ...
  - ... sind weniger anfällig im Bezug auf Verletzungen von Verteilungsvoraussetzungen
  - ... nutzen die in den Daten enthaltenen Information besser aus (z.B. höhere Power)

Faktorstruktur (1)

Begriff: Partitionierung / (Struktur-)Faktor

- Eine **Partition/Partitionierung** der Messwerte/Beobachtungen ist eine Methode diese in Gruppen einzuteilen, so dass *jeder* Wert zu einer Gruppe und *kein* Wert zu mehr als einer Gruppe gehört
- Ein **(Struktur-)Faktor** ist eine sinnvolle Partitionierung der Messwerte/Beobachtungen; als **sinnvoll** bezeichnet man eine Partitionierung in diesem Zusammenhang dann, wenn sie eine Menge von Bedingungen erfüllt, die durch das Design/ den Versuchsplan vorgegeben sind.

Faktorstruktur (2)

Erkennen von (Struktur-)Faktoren in einem Versuchplan

1. Teilen die Messwerte/Beobachtungen in den Gruppen der Partitionierung Eigenschaften, die von den Messwerten/Beobachtungen in anderen Gruppen nicht geteilt werden?
2. Ist es sinnvoll innerhalb der Gruppen Mittelwerte zu bilden und diese zu vergleichen?
3. Führt eine Vertauschung der Gruppenzugehörigkeit zu einer Veränderung des Sinnes der Daten?

Im Falle einer positiven Antwort ("ja") auf alle drei Fragen handelt es sich bei der Partitionierung um einen (Struktur-)Faktor.

Faktorstrukur (3)

Beispiel: "Physiotherapie bei Oberschenkelfraktur"

Gruppe	Datenpunkte									
A	15	17	21	16	13	14	21	18	16	14
B	16	12	18	18	19	15	16	20	17	
C	18	21	19	14	24	17	18	20	17	16

- Die Zeilen generieren einen Faktor (gleiche Behandlung; Vergleich von Zeilenmitteln sinnvoll; Gruppenzugehörigkeit von Daten nicht sinnvoll vertauschbar)
- Die Spalten hingegen generieren keinen Faktor (Werte in gleichen Spalten haben nicht mehr gemein als Werte in verschiedenen Spalten; Vergleich von Spaltenmittel ist wenig sinnvoll; Spaltenzugehörigkeit ist prinzipiell vertauschbar)

Faktorstruktur (4)

Universelle Faktoren

- Diese Faktoren treten in jedem Design auf
- Universelle Faktoren sind:

  Allgemeines Mittel / "Benchmark": Ist der Mittelwert aller Datenpunkte; d.h. dieser Faktor hat nur eine Faktorstufe (Level)

  Residualer "Fehler": Korrespondiert zur Differenz von individuellen Datenpunkten; d.h. jeder Datenpunkt bildet eine Faktorstufe

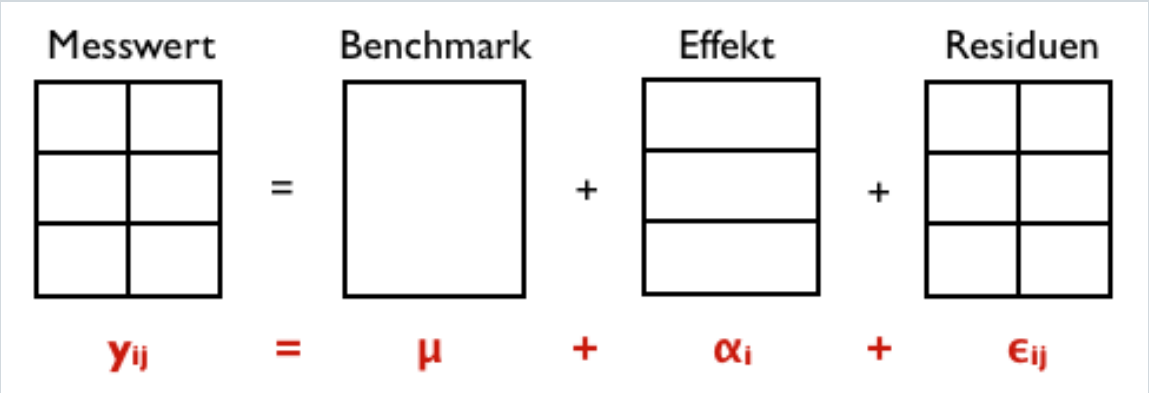
Faktorstruktur im BF[1] Design

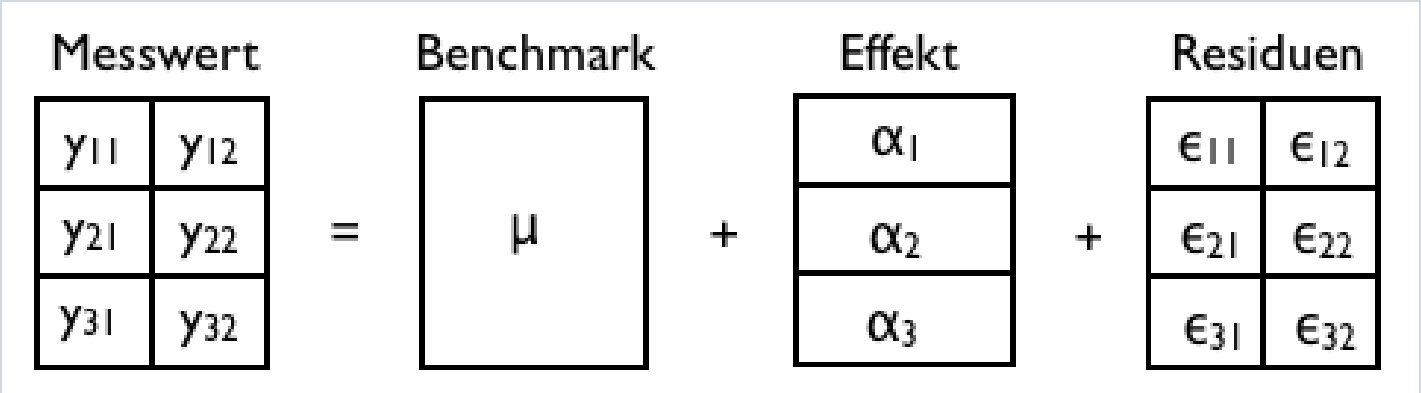
Beispiel: "Physiotherapie bei Oberschenkelfraktur"

Beispiel: "Hygienenstatus"

Design-(Struktur-)Faktoren und ANOVA Faktoren (1)

Zuordnung des ANOVA Modells zu den (Struktur-)Faktoren





Vollständig randomisiertes 2-Faktor Design (1)

(two-way basic factorial design, BF[2])

Das Prinzip der faktoriellen Kombination

Betrachte alle möglichen Kombination der Stufen mehrerer Faktoren

Realisierung eines BF[2] Designs

- Im Falle einer interventionellen Studie ordnet es die Beobachtungen zufällig allen Faktorkombinationen zu
- Faktoren können auch (teilweise) "gegeben", d.h. im Rahmen einer Beobachtungsstudie erhoben sein.
- Balancierte BF[2] Designs weisen in allen Faktorkombinationen die gleiche Beobachtungszahl auf

Vollständig randomisiertes 2-Faktor Design (2)

(two-way basic factorial design, BF[2])

Vor- und Nachteile von faktoriellen Designs

Vorteile:

- Erlaubt effiziente Betrachtung von 2 (oder mehr) Einflussfaktoren gleichzeitig (multivariate Analyse)
- Erlaubt die Schätzung von Wechselwirkungen zwischen den Faktoren
  - $\rightarrow$ Dies setzt das Vorliegen von Replikaten in den Faktorkombinationen (den Zellen des Designs) voraus!

Nachteile:

- Komplexes Design
- Ggf. schwerer zu kommunizieren und/oder zu interpretieren

Vollständig randomisiertes 2-Faktor Design (3)

(two-way basic factorial design, BF[2])

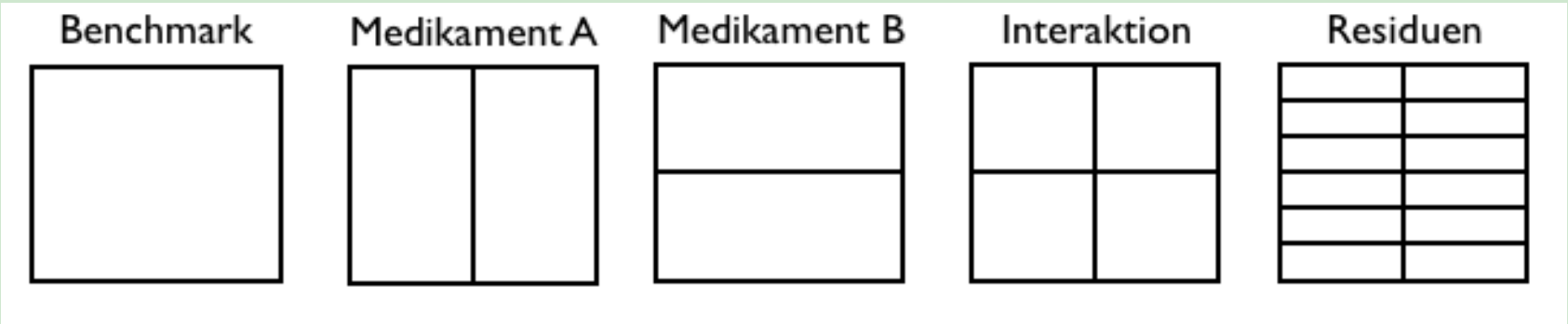
Beispiel: BF[2] zum Studium der Wirkung von zwei Chemotherapeutika

- Anwendung von zwei Medikamenten (A, B), wobei sowohl die isolierte Wirkung der jeweiligen Dosis als auch eine mögliche Interaktion hinsichtlich der Zeit bis zur Remission (in Wochen) geprüft werden soll
- Alle in die Studie eingeschlossenen Patienten werden zufällig auf die 4 Faktorkombinationen verteilt:

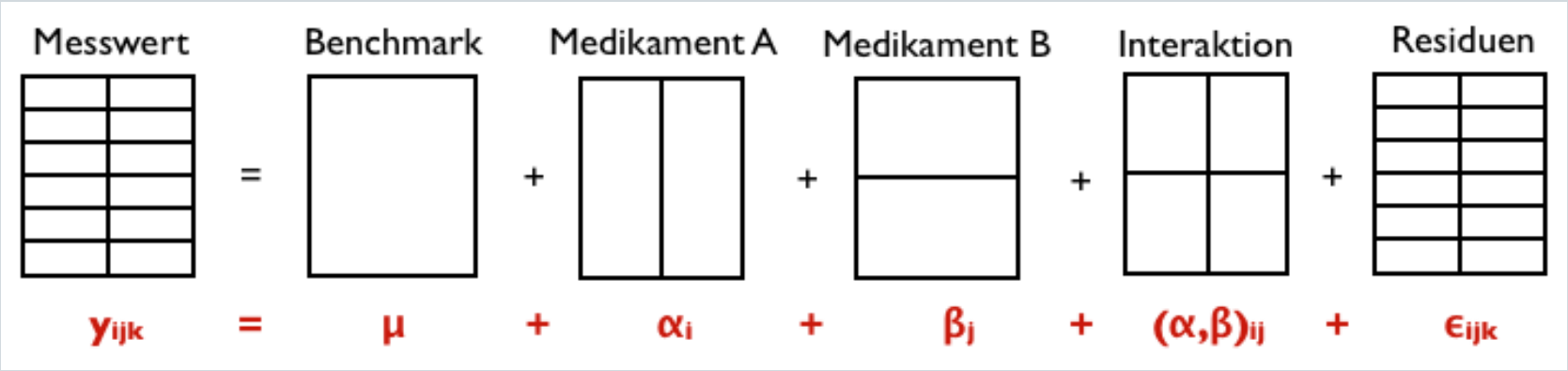
		Medikament A	
		10mg	40mg
Medikament B	100mg	22, 24, 30	25, 35, 22
	200mg	22, 28, 33	30, 35, 38

Faktorstruktur im BF[2] Design

Beispiel: "Kombi-Chemotherapie"



Zuordnung des ANOVA Modells zu den (Struktur-)Faktoren



Vollständig randomisiertes 2-Faktor Design (4)

(two-way basic factorial design, BF[2]

Beispiel: BF[2] im Kontext einer Beobachtungsstudie

- Analyse des Einflusses von natürlicher Radioaktivität und von Feinstaubbelastung auf die Auftretenshäufigkeit von Lungenkrebs in der Bevölkerung[a]
- Dazu Erhebung der Krebsraten in Bevölkerungsstichproben aus vier verschieden exponierten Regionen:
  1. **Weder** mit erhöhter natürlicher Radioaktivität **noch** mit Feinstaubbelastung
  2. **Mit** erhöhter natürlicher Radioaktivität **aber ohne** Feinstaubbelastung
  3. **Ohne** erhöhte natürlicher Radioaktivität **aber mit** Feinstaubbelastung
  4. **Sowohl mit** erhöhter natürlichen Radioaktivität **als auch mit** Feinstaubbelastung

[a] stat. Analyse dieser Fragestellung bedarf anderer Modelle als der klassischen Varianzanalyse; z.B. verallgemeinerte lineare Modell / Poissonregression

Störgrößen / Nuissance-Variablen

Begriff

- Als sogenannte *Nuissance-* oder *Störeffekte* bezeichnet man alle Arten der Variabilität, die nicht durch die interessierenden Versuchsbedingungen verursacht werden

Bemerkungen

- Unkontrollierte Nuissance-Effekte können zu Verzerrung (Bias) der Ergebnisse bzw. zu ihrer Fehlinterpretation führen
- Blockbildung macht Nuissance-Effekte explizit; sie integriert Nuissance-Faktoren als Blockvariablen in das Erklärungsmodell

Vollständig geblocktes Design (1)

(complete block design, CB)

Realisierung eines CB Designs

- Zuordnung von "ähnlichen" Beobachtungseinheiten zu Gruppen (Blöcke)
- Innerhalb der Gruppen werden die Beobachtungseinheiten, wie im Falle des BF Designs, zufällig auf die interessierenden Faktorstufen verteilt (d.h. jede Beobachtungseinheit ist einer Faktorstufe zugeordnet und jeder Block beinhaltet die komplette Menge möglicher Faktorstufen)

Vollständig geblocktes Design (1)

(complete block design, CB)

Vor- und Nachteile von Block-Designs

Vorteile:

- Erlaubt eine bessere Schätzung der Unterschiede zwischen interessierenden Faktoren, da die den Blockfaktoren zurechenbare Varianzkomponente separiert wird
- Besonders effizient bei hoher inter-individueller Variabilität der Gesamtstichprobe

Nachteile:

- CB Design weniger effizient als BF Design wenn die Blockfaktoren keine oder nur wenig Varianz erklären
- Aufwändiger zu planen als ein BF Design

Vollständig geblocktes Design (2)

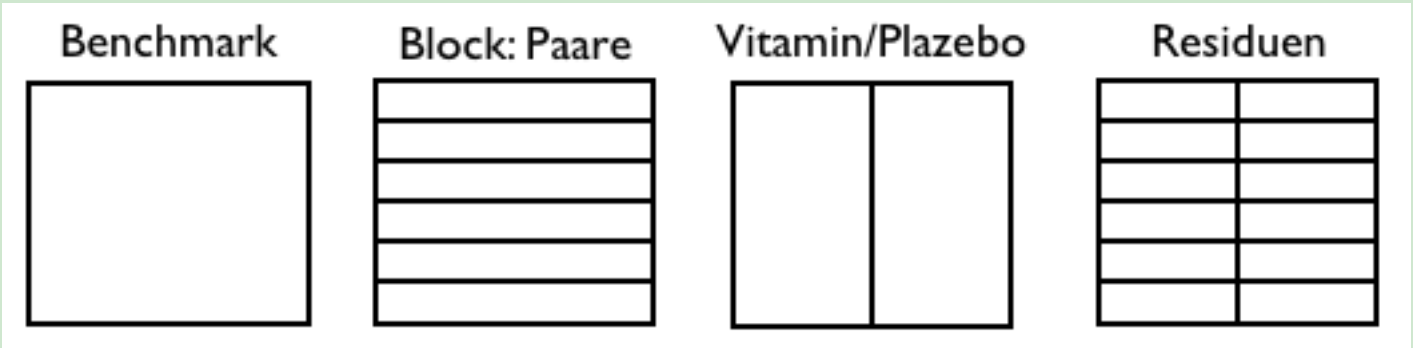
(complete block design, CB)

Beispiel: Untersuchung des Einflusses von Vitamin B6 auf Beschwerden im Zusammenhang mit dem prämenstruellen Syndrom (PMS)

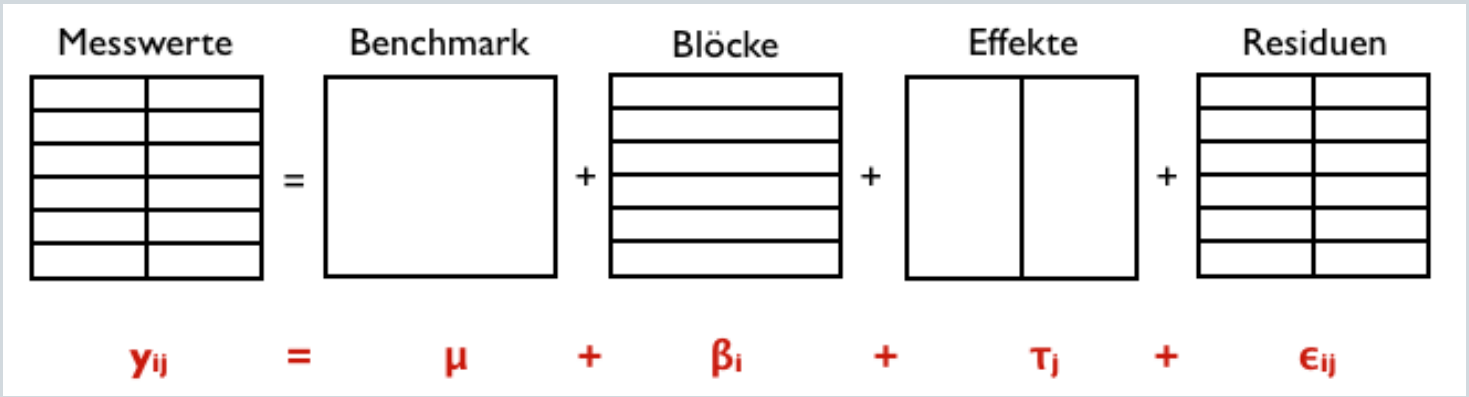
- Da mit einer großen Variabilität der subjektiven Beschwerden zu rechnen ist, sollen Paare von Frauen als Blöcke verwendet werden, wobei innerhalb der Paare jeweils zufällig das Vitamin B6 Präparat bzw. ein Placebo zugewiesen wird
- Wahl der Blockvariablen:
  - Körpergröße?
  - Alter?
  - Schwere/ Art der Symptome (mit Fragebogen ermittelt)?
- Was ist sinnvoller Weise als Blockvariable zu verwenden?

Faktorstruktur im CB Design

Beispiel: "Vitamin B6 Effekt"



Zuordnung des ANOVA Modells zu den (Struktur-)Faktoren



Latin Square (LS) (1)

Anwendung eines LS Designs

- Bei Auftreten von zwei Nuissance-Faktoren und einem interessierenden Faktor
- Wenn Anzahl der Level für alle drei Fakoren identisch
- Wenn alle Paare von Faktoren faktoriell kombiniert (”gekreuzt”) auftreten; speziell: jede Kombination von zwei Faktoren tritt exakt einmal auf

Realisierung eines LS Designs

- Anordnung der Stufen des ersten Nuissance-Faktors als Spalten-Blöcke und der Stufen des zweiten Nuissance-Faktors als Zeilen-Blöcke
- Aufteilung der Stufen des interessierenden Faktors so auf die Zellen, dass jede Stufe genau einmal pro Zeile und pro Spalte auftaucht

Latin Square (LS) (2)

Beispiel: Untersuchung der Wirksamkeit verschiedener Konzentrationsübungen auf Konzentrationsfähigkeit von Patienten mit Aufmerksamkeits Defizit Hyperaktivitäts Störung (ADHS)

- Ausgangssituation:
  - Betrachte 3 Übungen zur Konzentrationssteigerung: Ü1, Ü2, Ü3
  - Es stehen 3 Patienten zur Untersuchung zur Verfügung
  - Die Untersuchungen müssen alle an einem Tag durchgeführt werden
- Zielstellung / Probleme:
  - Ziel: Vergleich der drei Übungen bzgl. der Konzentrationssteigerung
  - Probleme:
    - Hohe Variabilität der Ausgangskonzentration zwischen Patienten
    - Abnahme der Konzentrationsfähigkeit im Verlauf des Tages

Latin Square (LS) (3)

Beispiel: ADHS - Fortsetzung (1)

Designvorschlag 1:

Nichtrandomisiertes CB Design (mit Blockvariablen Patient, Tageszeit)

	Tageszeit		
Patient	vormittags	mittags	nachmittags
1	Ü1	Ü2	Ü3
2	Ü1	Ü2	Ü3
3	Ü1	Ü2	Ü3

Bemerkung:

- Confounding der Tageszeit (Abnahme der Konzentration) mit Intervention (Übung 1 - 3)

Latin Square (LS) (4)

Beispiel: ADHS - Fortsetzung (2)

Designvorschlag 2:

Randomisiertes CB Design (mit Blockvariablen Patient, Tageszeit)

	Tageszeit		
Patient	vormittags	mittags	nachmittags
1	Ü1	Ü2	Ü3
2	Ü2	Ü1	Ü3
3	Ü1	Ü3	Ü2

Bemerkung:

- Speziell aufgrund der geringen Patientenzahl können Unbalanciertheiten zu einem Confounding der Tageszeit mit der Intervention führen

Latin Square (LS) (5)

Beispiel: ADHS - Fortsetzung (3)

Designvorschlag 3:

LS Design (mit Blockvariablen Patient, Tageszeit)

	Tageszeit		
Patient	vormittags	mittags	nachmittags
1	Ü1 (A)	Ü2 (B)	Ü3 (C)
2	Ü2 (B)	Ü3 (C)	Ü1 (A)
3	Ü3 (C)	Ü1 (A)	Ü2 (B)

Bemerkung:

- Nutzt Vorteile eines CB Designs (d.h. Trennung der Varianzkomponenten von interessierender und Nuissance-Variablen) und gewährleistet Balanciertheit / Vermeidung eines Confoundings

Latin Square (LS) (6)

Wann ist ein LS Design sinnvoll?

- Wenn Subjekte/Objekte eine hohe Variabilität aufweisen
  - → Wenn es möglich ist jedes Subjekt/Objekt unter allen Versuchsbedingungen zu messen, dann benutze Subjekte/Objekte als Blockfaktor
- Wenn die Reihenfolge der Bedingungen einen systematischen Effekt induzieren
  - → Benutze die Zeitabfolge als zusätzlichen Blockfaktor und verwende eine LS Struktur um eine Balancierung der Versuchsbedingungen zu garantieren
- Wenn der interessierende Faktor experimentell zu variieren ist, da eine aktive Zuordnung der Faktorstufen innerhalb des LS notwendig ist

Latin Square (LS) (7)

Randomisierung eines LS Designs

1. Zufällige Zuordnung des Subjekte/Objekte zu Zeilennummern
2. Zufällige Zuordnung der Zeitpunkte zu Spaltennummern
3. Zufällige Zuordnung der Faktorstufen zu den LS Buchstaben

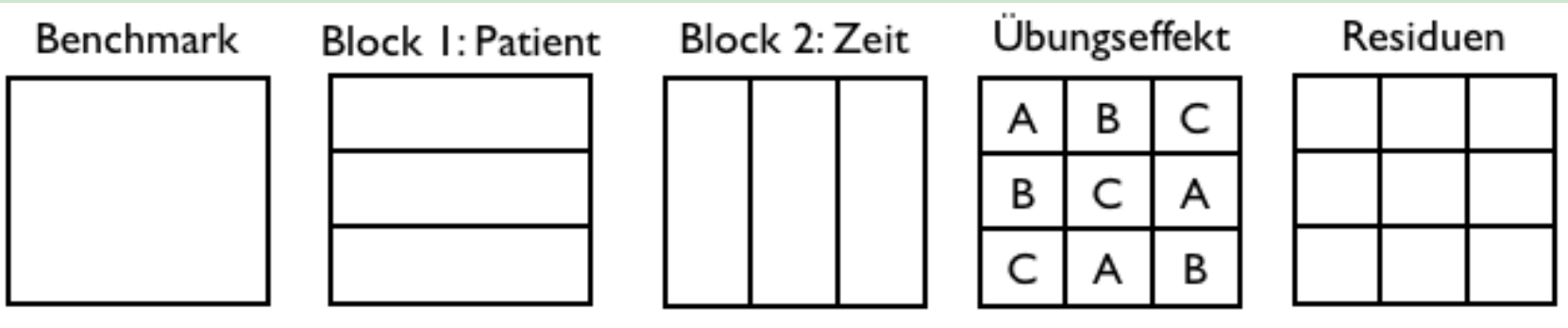
Beispiel: "ADHS"

1. z.B.: 1 - Patient 2; 2 - Patient 1; 3 - Patient 3
2. z.B.: 1 - nachmittags; 2 - vormittags; 3 - mittags
3. z.B.: A - Übung 2; B - Übung 1; 3 - Übung 3

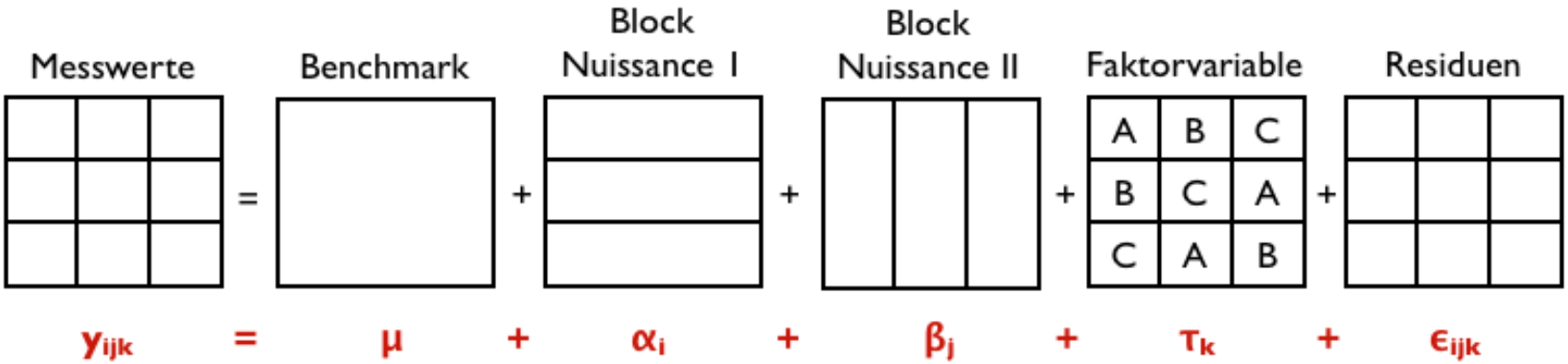
	Tageszeit		
Patient	nachmittags	vormittags	mittags
2	A: Ü2	B: Ü1	C: Ü3
1	B: Ü1	C: Ü3	A: Ü2
3	C: Ü3	A: Ü2	B: Ü1

Faktorstruktur eines LS Designs[1]

Beispiel: "ADHS"



Zuordnung des ANOVA Modells zu den (Struktur-)Faktoren



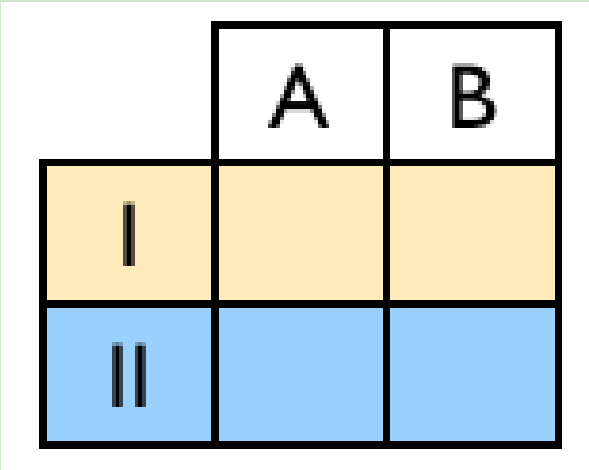
[1] für 3 x 3 LS Design; Verallgmeinerungen zu n x n sind möglich

Split-Plot Design (1)

(split plot design; SP)

Motivations-Beispiel: Bakterienwachstum in Abhängigkeit von Nährmedium und zugesetzten Desinfektionmitteln (1)

- Geprüft werden 2 Nährmedien (I / orange, II / blau) sowie 2 Desinfektionsmittel (A, B)



- Es stehen insgesamt 16 Messungen zur Verfügung; d.h. pro Faktorkombination können 4 Replikate vorgesehen werden

Split-Plot Design (2)

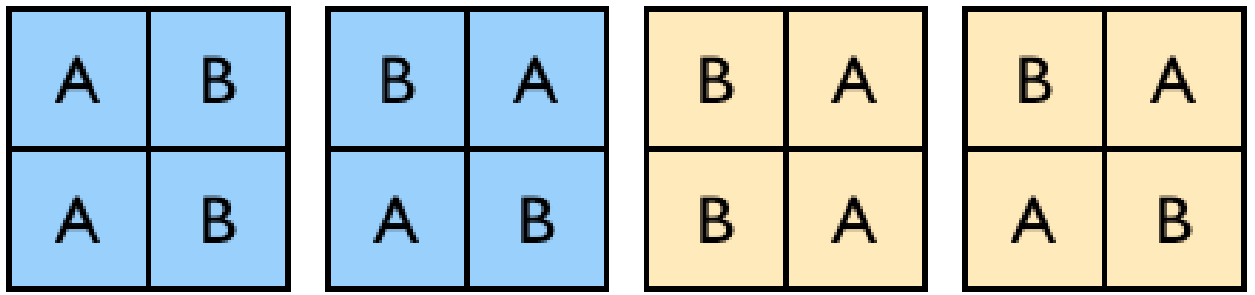
(split plot design; SP)

Motivations-Beispiel: Bakterienwachstum in Abhängigkeit von Nährmedium und zugesetzten Desinfektionmitteln (2)

→ Vollständing randomisiertes faktorielles Design:

A	B	B	A	A	B	B	A
A	A	A	B	B	A	B	B

→ SP Design:



Split-Plot Design (3)

(split plot design; SP)

Warum sollte man auf vollständige faktorielle Randomisierung verzichten?

- Ein Faktor ist vergleichsweise sehr aufwendig zu variieren
  - z.B. verschiedene Bakterienkolonien können in einer Kulturschale ("Multi-well-plate") gezüchtet werden, die aber je nur ein Nährmedium enthält; Variation des Desinfektionsmittel kann einfach innhalb der Kulturschale variiert werden
- Es gibt eine vorgegebene interne Struktur (Hierarchie), die eine vollständige Randomisation ausschließt
  - z.B. Analyse des Einflusses verschiedener Medikamente (Faktor 1)
    - auf Leber und Niere (Faktor 2): d.h. die Organe "variieren" innerhalb der Patienten
    - im zeitlichen Verlauf (Faktor 2); sog. *Repeated-Measurement (RM)* Designs

Split-Plot Design (4)

(split plot design; SP)

Realisierung eines SP Designs

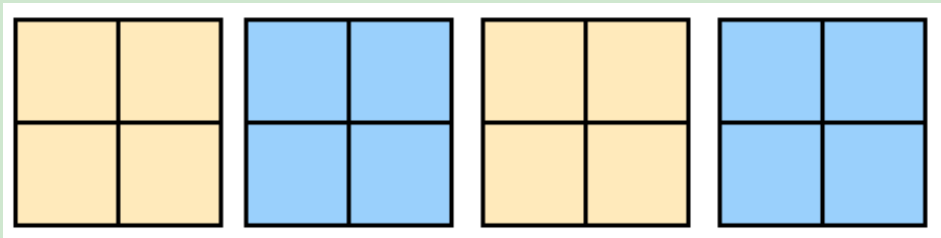
- Wahl des Blockfaktors ("whole plot / between subject factor": z.B. schwer zu variierender Faktor; Individuen)
- Zufällige Zuordnung der Stufen des Blockfaktors zu jeweils gleicher Anzahl von Blöcken ("whole plots")
- Zufällige Zuordnung der Stufen des Inner-Block Faktors ("sub-plot / within subject factor": z.B. leicht zu variierender Faktor; intra-individueller Faktor) innerhalb jedes Blöcke ("subplots")
- Unter Umständen sind die interessierenden Faktoren nicht randomisierbar (z.B. Zeitfaktor bei sequentiellen Messungen innerhalb von Patienten oder Erkrankung/Diagnose als nichtrandomisierbarer "beobachteter" Faktor).

Split-Plot Design (5)

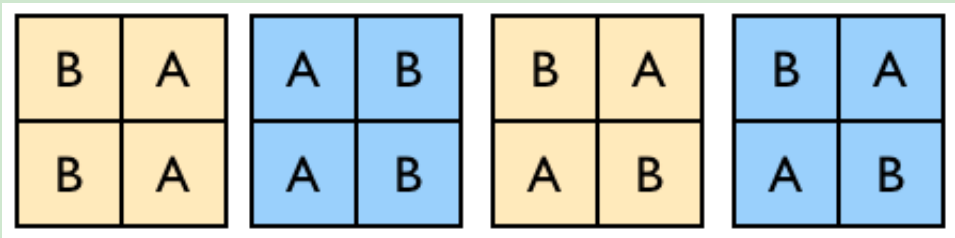
(split plot design; SP)

Beispiel: Realisierung eines SP Designs im "Bakterienkulturen"-Beispiel

- Wahl des Nährmediums als Blockfaktor (effiziente Verwendung von Multi-Wells innerhalb eines Kulturansatztes)
- Zufällige Zuordnung von Nährmedium I (orange) und II (blau) zu Multi-Wells



- Zufällige Zuordnung von zwei Repliaten je eines Desinfektionsmittels (A bzw. B) innerhalb der Blöcke



Split-Plot Design (6)

(split plot design; SP)

Vor- und Nachteile des SP Designs

Vorteile:

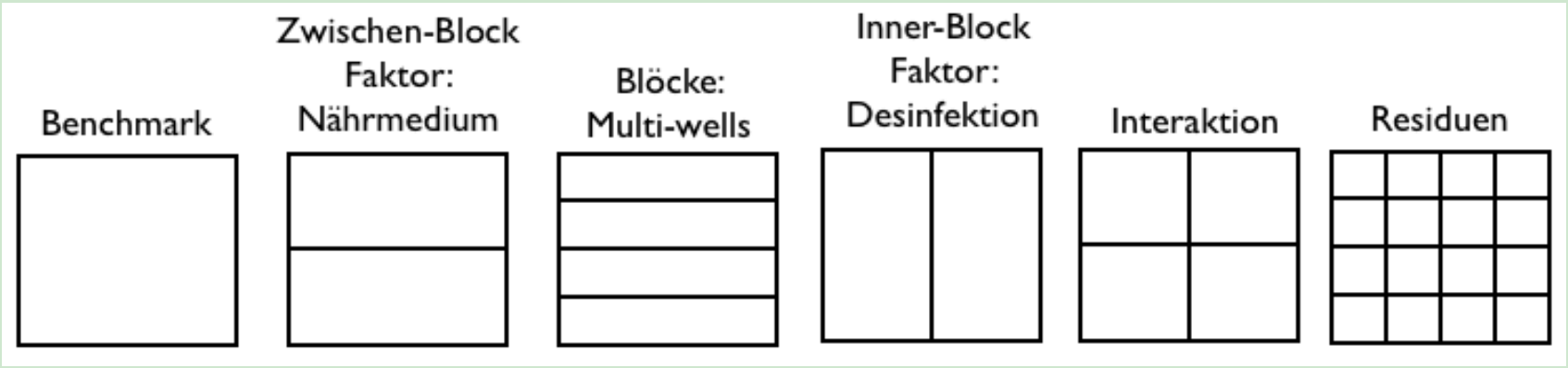
- SP Designs erlauben einen effizienten Umgang
  - mit "schwer variierbaren" Blockfaktoren
  - mit hierarchischen Faktorstrukturen und (intra-individuellen) Mehrfachmessungen
- SP Design erlauben die Komplementierung faktorieller Designs mit Blockstrukturen

Nachteile:

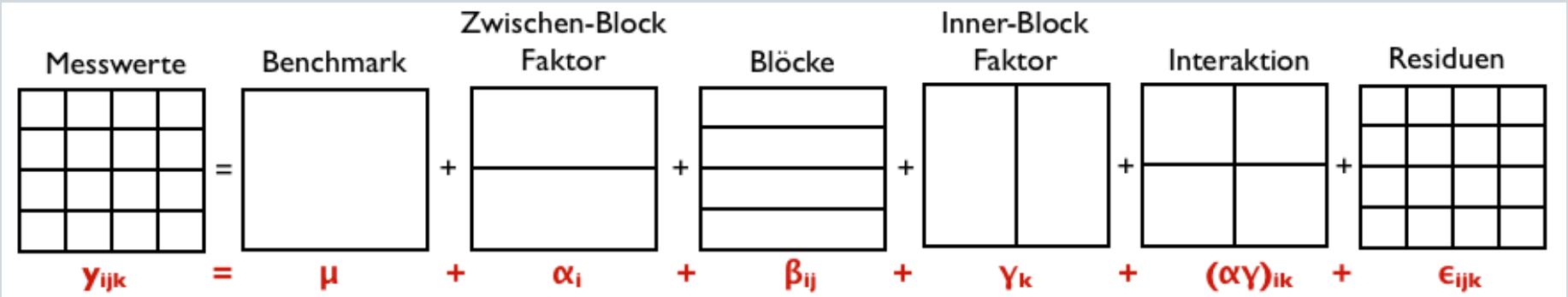
- Komplexe Designstruktur

Faktorstruktur im SP Design

Beispiel: "Bakterienkolonien"



Zuordnung des ANOVA Modells zu den (Struktur-)Faktoren



Faktorielle Kombination vs. Nesting (1)

Faktorielle Kombination

Faktor I	Faktor II		Kombination	
1	A	B	1A	1B
2			2A	2B

- Alle Stufen von Faktor I treten gemeinsam mit allen Stufen von Faktor II auf
- Möglichkeit der Schätzung von Interaktionen

Faktorielle Kombination vs. Nesting (2)

Nesting (Schachtelung)

Faktor I	Faktor II	Kombination
1	A	1A
	B	1B
2	C	2C
	D	2D

- Alle Stufen von Faktor II treten mit genau einer Stufe von Faktor I auf
- Allerdings sind Stufen von Faktor I nur einer Teilmenge der Stufen von Faktor II zugeordnet (Schachtelung der Stufen von II in Stufen von I)
- Keine Schätzung der Interaktion von Faktor I und II möglich

Faktorielle Kombination vs. Nesting (3)

Vollständiges Confounding

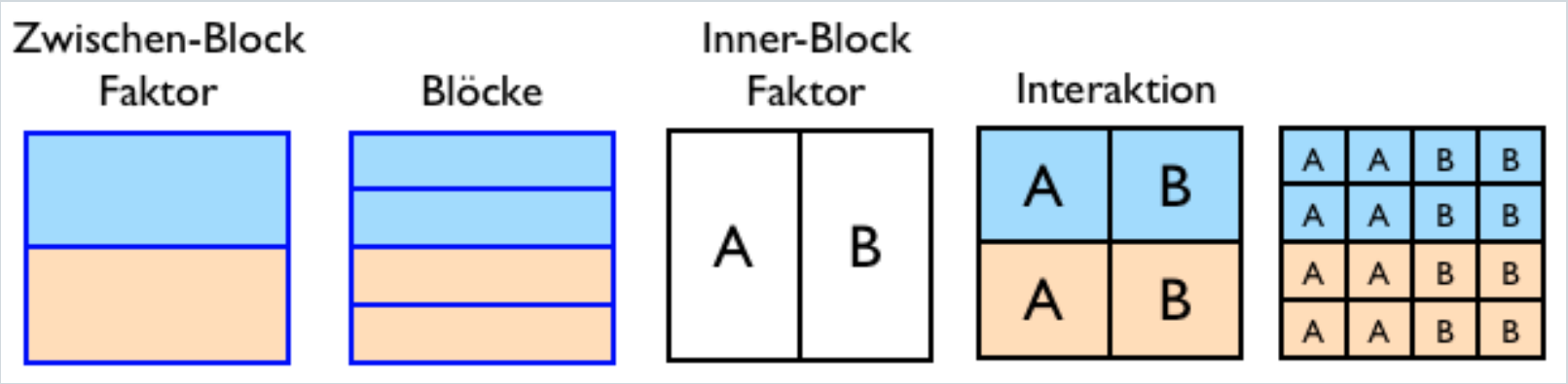
Faktor I	Faktor II	Kombination
1	A	1A
2	B	2B

- Die Stufen von Faktoren I korrespondieren zu denen von Faktor II
- D.h. Faktor I ist bzgl. Faktor II "nested", aber auch umgekehrt
- Effekt von Faktor I und II sind nicht zu trennen bzw. nicht getrennt voneinander zu schätzen

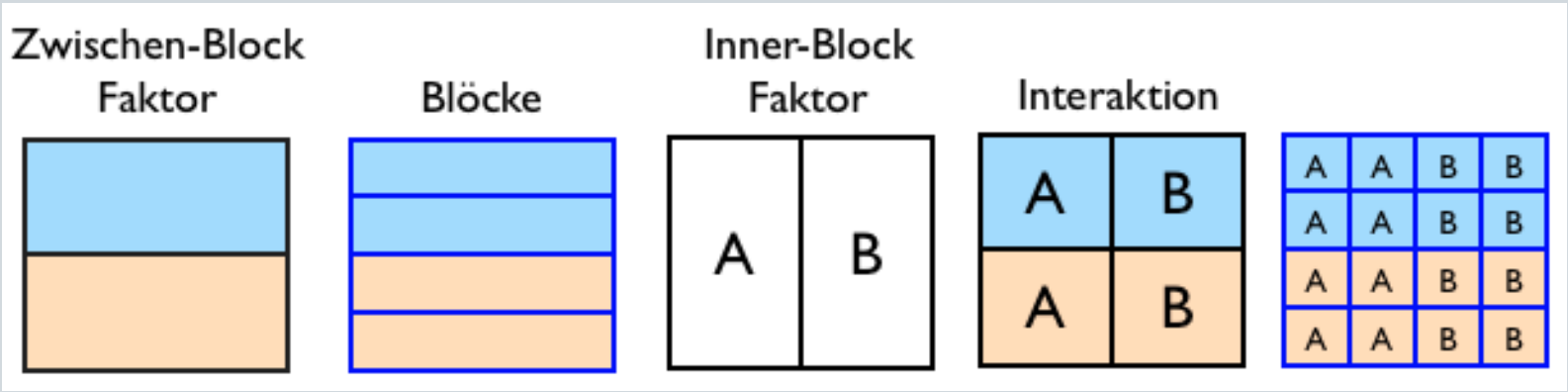
Faktorielle Kombination vs. Nesting (3)

... im SP Design (1)

- Blöcke sind "nested" innerhalb der Stufen des Zwischen-Block-Faktors



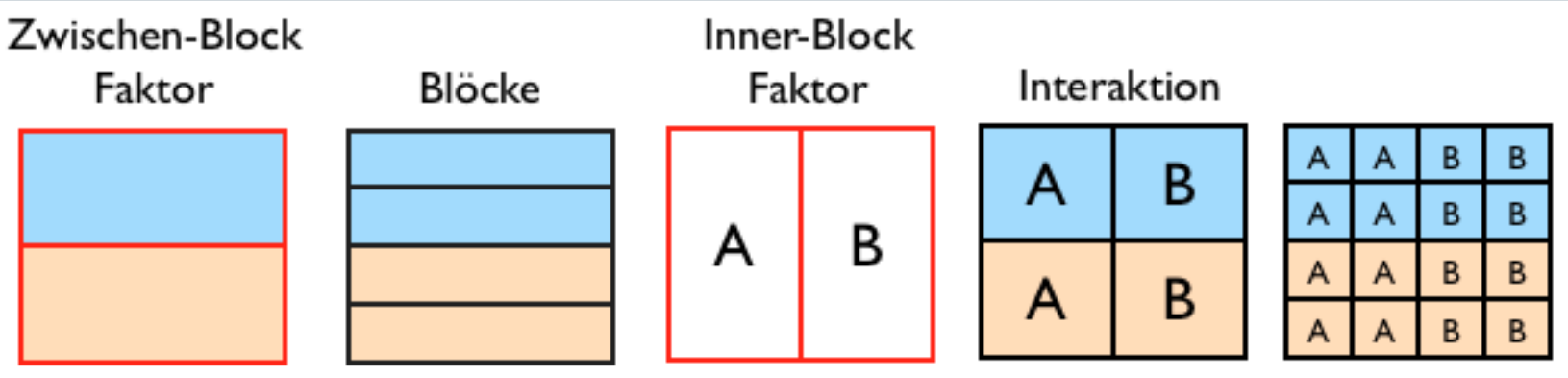
- Subplot sind "nested" innerhalb der Blöcke



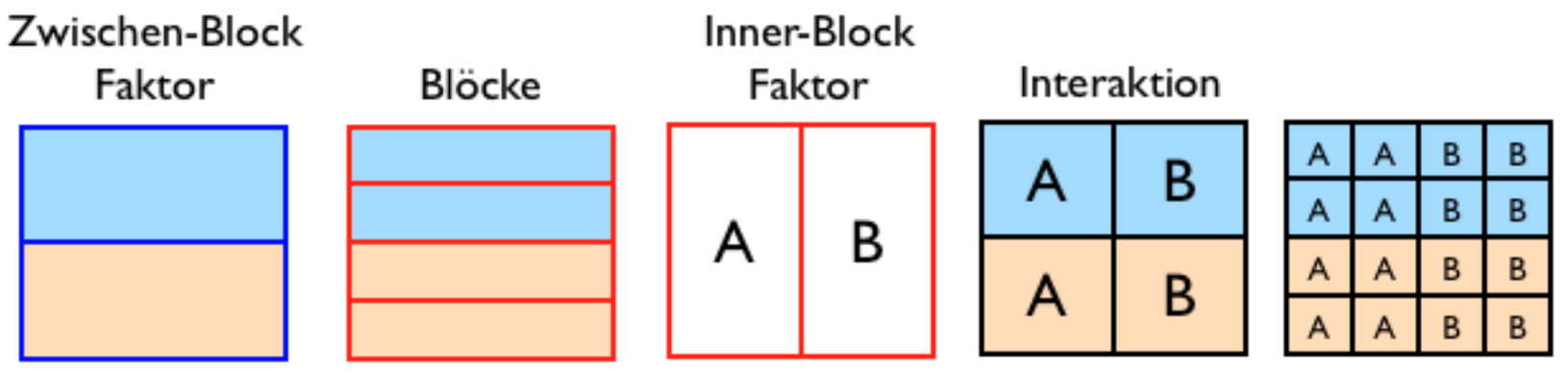
Faktorielle Kombination vs. Nesting (4)

... im SP Design (2)

- Zwischen-Block und Inner-Block Faktoren sind faktoriell kombiniert



- Blöcke und Inner-Block Faktoren sind faktoriell kombiniert



Anwendungsbeispiel: Einfluss eines Intensivkurses auf Prüfungsergebnisse (1)

Analyse mittels BF[1] **Beobachtungs**-Designs

- Beobachtung von je 10 Personen, die vor Ablegen der Prüfung den Intensivkurs entweder *belegt* oder *nicht belegt* hatten
- Vergleich der Prüfungsergebnisse

Bemerkungen

- Confouning des (potentiellen) Effektes des Kurses mit einem (potentiellen) Effekt der Motivation / der Einstellung etc.

Anwendungsbeispiel: Einfluss eines Intensivkurses auf Prüfungsergebnisse (2)

Analyse mittels **CB** Design

- Verwendung der Person als Block-Variable
- Durchführung von je einer Prüfung *vor* und *nach* dem Kurs mit jeder Person (intra-individueller vor/nach-Vergleich)

Bemerkungen

- Vermeidung des Confounings mit Motivation durch intra-individuellen Vergleich
- Aber: Confouning des (potentiellen) Effektes des Kurses mit einem davon unabhängigen Lerneffekt (durch Wiederholung der Prüfung)

Anwendungsbeispiel: Einfluss eines Intensivkurses auf Prüfungsergebnisse (3)

Analyse mittels **randomisierten BF[1]** Designs

- Vermeidung bisheriger Stör-(Nuisance)-Effekte: *Zwischen-Subjekt-Unterschiede bzgl. Motivation* bzw. *Zeit-/ Trainingseffekt* im Rahmen eines vollst. randomisierten one-way BF Designs:
- Randomisierung von freiwilligen Testpersonen in zwei Gruppen:
  (1) Kurs + Prüfung
  (2) Prüfung (außerhalb der Studie: Möglichkeit der Wiederholung der Prüfung inkl. Kursteilnahme)

Bemerkungen

- Vermeidet bisheriges Confounding
- Aber: Verlust der Varianzreduktion (bzgl. Kurs-Effekt) die im CB Design erreicht wurde
- Erhöhter Organisationsaufwand (nachträglicher Kurs + Prüfung)

Anwendungsbeispiel: Einfluss eines Intensivkurses auf Prüfungsergebnisse (4)

Analyse mittels **SP** Designs

- Randomisierung von freiwilligen Testpersonen in zwei Gruppen:
  (1) Prüfung $\rightarrow$ Kurs $\rightarrow$ Prüfung
  (2) Prüfung $\rightarrow \emptyset \rightarrow$ Prüfung (außerhalb der Studie: Anschluss von $\rightarrow$ Kurs $\rightarrow$ Prüfung)

Bemerkungen

- Vermeidet Confounding bzgl. Motivation und Zeit-/ Trainingseffekt
- Reduziert Varianz der Effektschätzung durch Separierung der inter- und intra-individueller Varianz
- Aber: recht hoher Organisationsaufwand + ggf. Akzeptanzprobleme wg. mehrfacher Prüfung

Fallzahlplanung - Einordnung in Versuchplanungskontext

Bisher

Betrachtung genereller Prinzipien zur

- Reduktion bzw. "Kanalisation" von Varianz (*Replikation*, *Randomisierung*, *Blockbildung*)
- Realisierung der Schätzung bestimmter Effekte (*Faktorkombination*)

Nunmehr

- Festlegung der Fallzahl (innerhalb eines gegebenen Designs) zur Begrenzung der Fehlerraten
- Beurteilung der *Power* eines Experiments / einer Studie bei gegebener Fallzahl

  (Im Rahmen der VL hierbei Beschränkung auf einfachste Designs)

Wiederholung: Aufgaben der Fallzahl/Power-Abschätzung

Planungsphase

- Abwägung der möglichen Fehler (1./2. Art) und der Größe des detektierbaren Unterschiedes bei vorgegebener (limitierter) Fallzahl
- Ermittlung der notwendigen Fallzahl zur Abschätzung eines vorgegebenen Unterschieds unter Einhaltung spezifizierter Fehlerniveaus (1./2. Art)
- Abwägung von Fallzahl (Kosten/Aufwand) und Signifikanzniveau + Power (Konfidenz der Aussagen)

Bewertung von Ergebnissen einer Studie:

- Wurden geplante Kenngrößen (z.B. Power, Fehler 1. Art) eingehalten?
- Entspricht die Studiendurchführung der Planung (z.B. Fallzahl erreicht)?

Wiederholung: Fehlerarten (1)

Vorgabe von Fehlerniveaus

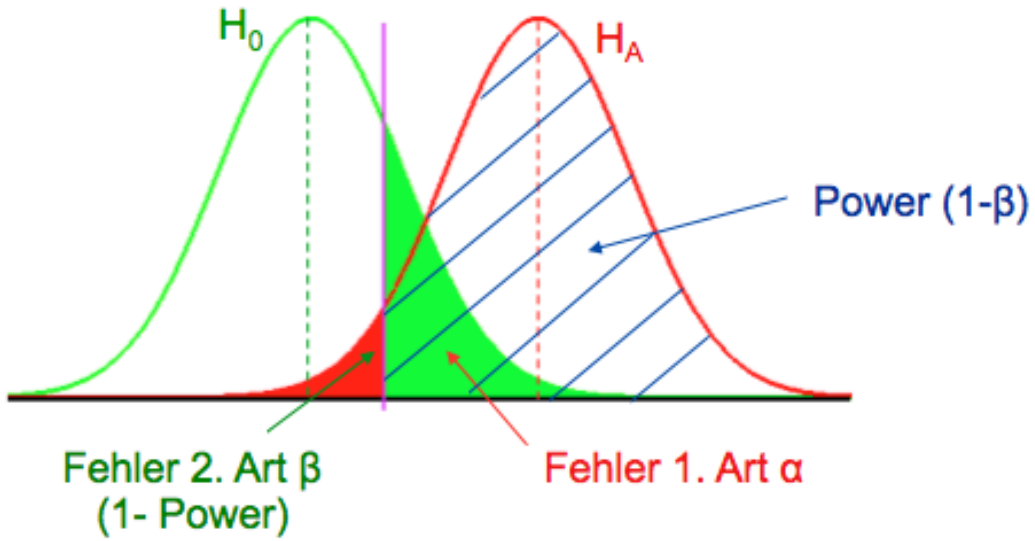
- Unabhängig von der Fallzahl, wird durch den Vergleich des P-Wertes mit einem vorgegebenen Signifikanzniveau der Fehler 1. Art kontrolliert.
- **Aber:** Fehler 2. Art (bzw. damit auch die Power) hängt unter Anderem von der Stichprobengröße ab!

		Realität	
		H_0 falsch	H_0 richtig
Test-entscheidung	H_0 ablehnen	***Power: 1- β***	***Fehler 1. Art: α*** OK
	H_0 beibehalten	***Fehler 2. Art: β*** ?	***1- α***

Wiederholung: Fehlerarten (2)

Zusammenhang Fehler 1. und 2. Art

Angenommen die Verteilungen der Testgröße unter H_0 und H_A wären bekannt:



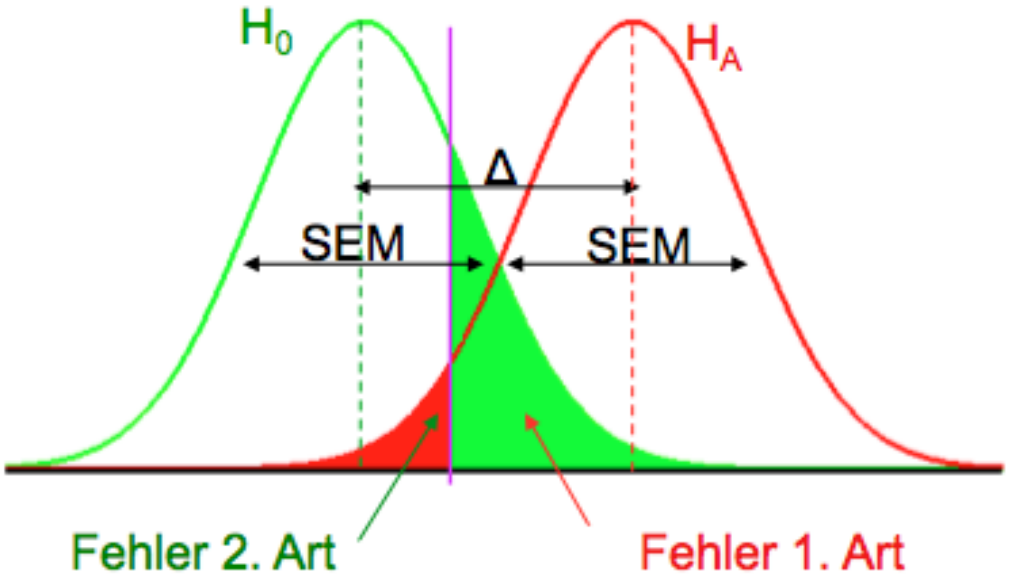
Fehler 1. Art (kontrolliert durch Signifikantniveau α) und Fehler 2. Art (beschrieben durch 1 – Power = β) können im Allgemeinen nicht gleichzeitig minimiert werden.

Abwägung notwendig, welcher Wert für welchen Fehler akzeptabel ist!

→ Dies ist eine inhaltliche Frage, keine statistische!

Wiederholung: Relevante Einflussgrößen (1)

Zusammenhang: Effektgröße, Varianz, Fehler und Fallzahl



- Δ ... relevanter Unterschied
- Genauigkeit der Schätzung der Testgröße:

$$SEM = \frac{\sigma}{\sqrt{n}}$$

→ Fehlerraten sind abhängig von Variabilität der Daten σ, von Fallzahl n (da diese die Genauigkeit der Schätzer beeinflusst) und vom zu entdeckenden Unterschied Δ (d.h. von der Wahl von H_0 und H_A)

Wiederholung: Relevante Einflussgrößen (2)

Ziele der Fallzahlplanung

Plane die Fallzahl eines Experiments oder einer Studie so, dass

- bestimmter Effekt statistisch "gesichert" detektiert werden kann, d.h.
- Fehler 1. Art (Signifikanzniveau) einen vorgeg. Wert nicht übersteigt
- Fehler 2. Art einen vorgegebenen Wert nicht übersteigt, d.h. die Power (1 – Fehler 2. Art) eine Mindestgröße hat

dabei zu beachten:

Die notwendige Fallzahl steigt mit ...

- kleinerem Signifikanzniveau α bzw. höherer Power $1 - \beta$
- kleinerem aufzudeckendem Unterschied Δ (annähernd quadratisch: halber Unterschied, vierfache Fallzahl)
- größerer Varianz σ des Endpunkts
- höherer Drop-out-Rate

"Inkorrekte" Fallzahlen (1)

Probleme *zu kleiner* Studien

- Decken selbst große Unterschiede kaum auf (d.h. zu geringe Power)

$\rightarrow$ Bessere Behandlung wird evtl. nicht etabliert

Beispiel: Vergleich konventionelle vs. innovative Behandlung

- Fallzahl: $n_{konventionell} = n_{innovativ} = 10$
- Annahme eines relevanten Unterschieds von 40 Prozentpunkten:
  - konventionell: 30% 5-Jahres-Überlebenswahrscheinlichkeit vs.
  - innovativ: 70% 5-Jahres-Überlebenswahrscheinlichkeit

$\rightarrow$ Die Studie hat nur eine Chance (Power) von ca. 25%, dass der (extrem) große Unterschied bei geg. Signifikanzniveau von 5% entdeckt wird.

"Inkorrekte" Fallzahlen (2)

Probleme *zu großer* Studien

- "Verbrauchen" mehr Patienten als nötig
- Machen u.U. auch irrelevant kleine Effekte sichtbar

$\rightarrow$ suggerieren ggf. nichtvorhandene (nichtrelevante) Effekte

Beispiel: Vergleich konventionelle vs. innovative Behandlung

- Fallzahl: $n_{konventionell} = n_{innovativ} = 900$
- Relevanter Unterschied bzgl.
  - primären Endpunkts: konventionell, 80% vs. innovativ, 98% 5-Jahres-Überlebenswahrscheinlichkeit
  - sekundären Endpunktes: konventionell, 50% vs. innovativ, 60% der Patienten litten an Übelkeit

$\rightarrow$ Unterschied bei sek. Endpunkt wird mit Power von ca. 99% sichtbar.

$\rightarrow$ Aber schon 115 Patienten je Arm hätten zum Aufzeigen des Unterschieds bzgl. primären Endpunkt mit dieser Sicherheit gereicht.

Zwei Konzepte der Fallzahlberechnung

Präzisions-Analyse

- Kommt bei der Schätzung von Parametern mittels Intervallschätzern zum Einsatz
- Fragestellung: Welche Fallzahl benötigt man, um einen Parameter mit einer vorgegebenen Fehlerwahrscheinlichkeit hinreichend präzise zu schätzen?

Power-Analyse

- Kommt bei der Planung/ Interpretation statischer Tests zum Einsatz
- Fragestellungen z.B.:
  - Welche Fallzahl benötigt man, um einen interessierenden (relevanten) Unterschied bei gegebenem Fehler 1. Art mit einer bestimmten Power zu entdecken?
  - Welche Power zur Detektion eines interessierenden Effektes hat ein Test bei gegebenem Fehler 1. Art für verschiedene Fallzahlen?

Präzisions-Analyse (1)

Fallzahl für KI des Erwartungswerts bei Normalverteilung und bekanntem σ

Wiederholung:

- Das $(1-\alpha)$%-KI wird mittels

$$P\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

bestimmt und lautet

$$\left[\bar{X} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \quad ; \quad \bar{X} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

Präzisions-Analyse (2)

Fallzahl für KI des Erwartungswerts bei Normalverteilung und bekanntem σ (Forts.)

- Die Präzision kann mittels (halber) KI-Breite (E) charakterisiert werden:

$$E = z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

- Durch Umstellung der Beziehung $E = z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ nach n, d.h.:

$$n = (z_{1-\frac{\alpha}{2}})^2 \cdot \frac{\sigma^2}{E^2}$$

kann zu vorgegebener Präzision, d.h. gegebener (halber) KI-Breite, die notwendige Fallzahl n berechnet werden.

Präzisions-Analyse (3)

Beispiel: 95%-KI für den Erwartungswert einer normalverteilten Zufallsvariablen mit bekannter Streuung σ^2

- Gesucht: Fallzahl, die es gestattet den Erwartungswert auf eine Genauigkeit (E) von $\pm 0.1 \cdot \sigma^2 = 0.2$, bei gegebenem Konfidenzniveau $1 - \alpha = 0.95$ zu schätzen. Die Streuung habe den Wert $\sigma^2 = 2$.
- Lösung:

$$E = z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{\sqrt{2}}{\sqrt{n}} = 0.2$$

und damit

$$n = z^2_{1-\frac{\alpha}{2}} \cdot \frac{\sigma^2}{E^2} = 1.96^2 \cdot \frac{2}{0.2^2} = 192.07$$

- D.h. es sind 193 Fälle notwendig.

Power-Analyse (1)

Vergleich zweier Erwartungswerte (μ_1, μ_2) bei bekannter, gemeinsamer Streuung ("2-SP Gauss-Test") und SP-Größe $n_1 = n_2 = n$

- Die Messwerte seien normalverteilt mit Erwartungswert μ_1 bzw. μ_2 und Varianz σ^2
- Die Erwartungswerte werden mittels der arith. Mittel $\bar{x}_1, \bar{x}_2$ geschätzt
- Diese folgen jeweils (für $i = 1, 2$) einer $\mathcal{N}\left(\mu_i, \sigma^2/n_i\right)$ Verteilung
- Die zugehörige Testgröße

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2) \cdot \sqrt{n}}{\sigma \cdot \sqrt{2}}$$

ist $\mathcal{N}\left(\frac{(\mu_1 - \mu_2) \cdot \sqrt{n}}{\sigma \cdot \sqrt{2}}, 1\right)$ verteilt[a]

---

[a] Aus $\bar{X}_i \sim \mathcal{N}\left(\mu_i, \sigma^2/n_i\right)$ folgt $\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$ und damit unter der Annahme $n_1 = n_2 = n$ und nach Division durch SD:
$Z \sim \mathcal{N}\left(\frac{(\mu_1 - \mu_2) \cdot \sqrt{n}}{\sigma \cdot \sqrt{2}}, 1\right)$

Power-Analyse (2)

Vergleich zweier Erwartungswerte (μ_1, μ_2) bei bekannter, gemeinsamer Streuung ("2-SP Gauss-Test") und SP-Größe $n_1 = n_2 = n$ (Forts.)

- Unter H_0 (d.h. $\mu_1 = \mu_2$ bzw. $\mu_1 - \mu_2 = 0$) gilt

$$Z^{(H_0)} \sim \mathcal{N}(0, 1)$$

- Testentscheidung: Lehne H_0 ab, wenn $Z^{(H_0)} > z_{1-\alpha}$ [a]

- Unter der spezifischen, einseitigen Alternativhypothese $H_A : \mu_1 - \mu_2 = \Delta > 0$ gilt

$$Z^{(H_A)} \sim \mathcal{N}(\frac{\Delta\sqrt{n}}{\sigma \cdot \sqrt{2}}, 1)$$

[a] $z_{1-\alpha}$: $(1-\alpha)$ - Quantil der $\mathcal{N}(0,1)$ Verteilung

Power-Analyse (3)

Vergleich zweier Erwartungswerte (μ_1, μ_2) bei bekannter, gemeinsamer Streuung ("2-SP Gauss-Test") und SP-Größe $n_1 = n_2 = n$ (Forts.)

- Unter dieser spezifischen Alternative H_A $(Z^{(H_A)} \sim \mathcal{N}(\frac{\Delta\sqrt{n}}{\sigma\cdot\sqrt{2}}, 1))$ betrachte nunmehr die Power-Bedingung:

$$\begin{aligned} 1-\beta &= P(\text{Ablehnung von } H_0 | H_A) \\ &= P(Z > z_{1-\alpha} | H_A) \\ &= 1 - \Phi(z_{1-\alpha} - \Delta\frac{\sqrt{n}}{\sigma\sqrt{2}}) \\ &= \Phi(\Delta\frac{\sqrt{n}}{\sigma\cdot\sqrt{2}} - z_{1-\alpha}) \end{aligned}$$

Power-Analyse (4)

Vergleich zweier Erwartungswerte (μ_1, μ_2) bei bekannter, gemeinsamer Streuung ("2-SP Gauss-Test") und SP-Größe $n_1 = n_2 = n$ (Forts.)

- Anwendung von Φ^{-1} auf beiden Seiten der Gleichung ergibt

$$\Phi^{-1}(1-\beta) = z_{1-\beta} = \Delta \frac{\sqrt{n}}{\sigma \cdot \sqrt{2}} - z_{1-\alpha}$$

- Umstellung von $z_{1-\beta} = \Delta \frac{\sqrt{n}}{\sigma \cdot \sqrt{2}} - z_{1-\alpha}$ nach n erlaubt Bestimmung des Stichprobenumfanges:

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \cdot 2}{(\frac{\Delta}{\sigma})^2}$$

- Δ/σ nennt man auch *effektiven Unterschied* (d.h. Δ gemessen in Einheiten von σ)

Power-Analyse (5)

Beispiel: Einseitiger Signifikanztest zum Test der Hypothese bzgl. zweier Erwartungswerte ($H_0 : \mu_1 < \mu_2$) bei bekannter, gemeinsamer Streuung

- Gegeben:
  - Gewünschtes Signifikanzniveau $\alpha = 5\%$
  - Angestrebte Power $1 - \beta = 80\%$
  - Zu detektierender Unterschied der Erwartungswerte, $\Delta = 2$
  - Gemeinsame Streuung in beiden Populationen, $\sigma^2 = 9$
- Gesucht: notwendige Fallzahl (gleiche Gruppengröße: $n_1 = n_2 = n$)
- Lösung:
  - $z_{1-\alpha} = z_{0.95} = 1.6449$
  - $z_{1-\beta} = z_{0.80} = 0.8416$

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \cdot 2}{(\Delta/\sigma)^2} = \frac{(1.6449 + 0.8416)^2 \cdot 2}{(2/3)^2} = 27.8$$

- D.h. es sind 28 Beobachtungen pro Gruppe notwendig.

Power-Analyse (6)

Beispiel: Einseitiger Signifikanztest zum Test der Hypothese bzgl. zweier Erwartungswerte ($H_0 : \mu_1 < \mu_2$) bei unbekannter, gemeinsamer Streuung

- Gegeben (wie bisher):
  - $\alpha = 5\%$, $1 - \beta = 80\%$, $\Delta = 2$
  - Gemeinsame Streuung wird geschätzt: $\hat{\sigma}^2 = s^2 = 9$
- Berechnung der notwendigen Fallzahl (gleiche Gruppengröße):

$$n = \frac{(t_{1-\alpha,df} + t_{1-\beta,df})^2 \cdot 2}{(\Delta/\sigma)^2}$$

- Problem: $df = n_1 + n_2 - 2$ hängen von unbekannter Fallzahl ab!
- Damit Bestimmung der Fallzahl nur iterativ möglich

→ Anwendung eines Fallzahlprogrammes:
Hier sind 29 Beobachtungen pro Gruppe notwendig.

Power-Analyse (7)

Bemerkung

Da die Fallzahl sehr sensitiv auf verschiedene Annahmen und Designparameter reagiert, ist u.U. eine systematische Betrachtung mehrere Szenarien angebracht (auch dies wird manchmal als *Poweranalyse* bezeichnet). D.h.:

- Wie ändert sich Power bei geplanter Fallzahl und variierenden Planungsparametern?
- Bei der Bewertung einer Studie: Prüfung der Planungsparameter; ggf. Nutzung einer Poweranalyse (welche idealer Weise im Studienprotokoll vorliegen sollte)

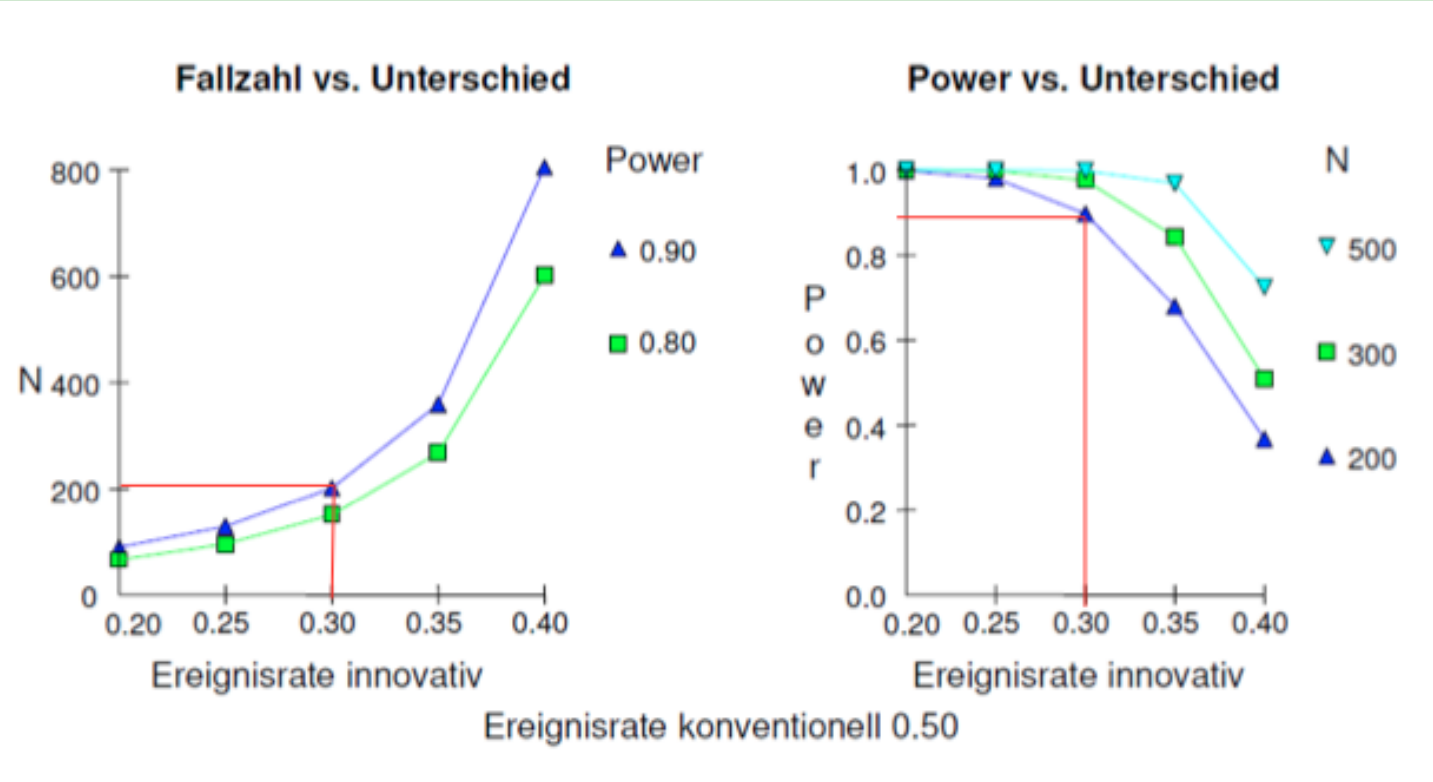
Power-Analyse (8)

Beispiel: systematische *Poweranalyse* in Studie zu Hypoglykämien bei Typ-1-Diabetikern

- Fragestellung: Kann die Zahl der Hypoglykämien bei Typ-1-Diabetikern durch eine innovative Schulung gesenkt werden?
- Endpunkt: Wartezeit auf schwere Hypoglykämie
- Methode: Kaplan-Meier Überlebenszeit, 2-seitig
- Beobachtungszeit je Patient: 2 Jahre
- Effektgröße: Ereignisrate bei Standardschulung: 0.5/Jahr; angestrebt/erwartet bei innovativer Schulung: 0.3/Jahr
- Akzeptierte Fehlerraten: α=5%, β=10%
- erwartete Drop-out-Rate: 0.2/Jahr

Power-Analyse (8)

Beispiel: systematische *Poweranalyse* in Studie zu Hypoglykämien bei Typ-1-Diabetikern (Forts.)



Fallzahlplanung - Zusammenfassung

Bei Fallzahlplanung festzulegen:

- Welcher primärer Endpunkt soll betrachtet werden?
  $\rightarrow$ für versch. Endpunkte ergeben sich u.U. verschiedene Fallzahlen!
- Mit welcher statistischen Methode wird ausgewertet?
  $\rightarrow$ für verschiedene Teststatistiken ergeben sich u.U. verschiedene Fallzahlen!
- Welcher Unterschied soll aufgedeckt werden?
  $\rightarrow$ unterschiedliche große Effekte benötigen versch. Fallzahlen zu ihrer Detektion!
- Mit welcher Varianz der Daten muss man rechnen?
  $\rightarrow$ bei größerer Varianz der Daten benötigt man größere Fallzahlen um einen gegebenen Effekt "nachzuweisen"!
- Wie groß dürfen die Fehlerraten der statistischen Diagnostik sein?
  $\rightarrow$ Welcher Fehler (1. / 2. Art) ist für die konkrete Studie wichtiger!

Software zur Fallzahlanalyse

Validierte (Profi-)Software

- nQuery (www.statsols.com/nquery)
- PASS (www.ncss.com/software/pass/)

Freie Software

- G*Power (www.gpower.hhu.de)
- ADDPLAN (www.berryconsultants.com/software/addplan/)
- R / z.B. Pakete: pwr, samplesize (www.r-project.org)

Auswahl freier Web-Tools (Web-tools meist nicht validiert)

- `http://www.stat.uiowa.edu/~rlenth/Power/index.html`
- `http://statpages.org/#Power`
- `http://www.stat.ubc.ca/~rollin/stats/ssize/index.html`
- `http://hedwig.mgh.harvard.edu/sample_size/size.html`

Vorlesung „Sampling-Strategien“

innerhalb

VL „Prinzipien und Methoden medizinischer Forschung“

Dipl.-Ing. Gabriele Müller
(*gabriele.mueller@tu-dresden.de*)

Institut für medizinische Informatik und Biometrie (IMB)
Medizinische Fakultät Carl Gustav Carus
TU Dresden

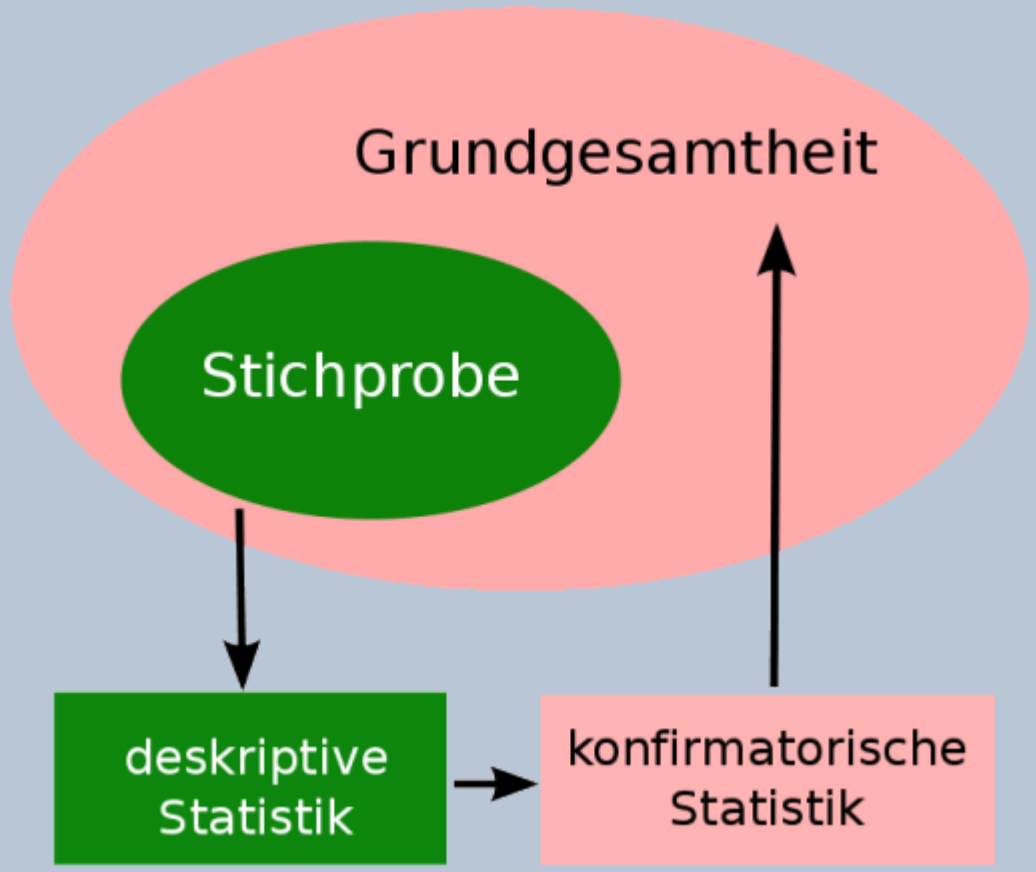
TECHNISCHE UNIVERSITÄT DRESDEN

Outline

1. Einführung
2. Eigenschaften von Stichproben
3. Sampling-Strategien
4. Daten vom Einwohnermeldeamt

Wozu brauchen wir Stichproben?

… siehe Biometrie- / Epidemiologie-Vorlesung

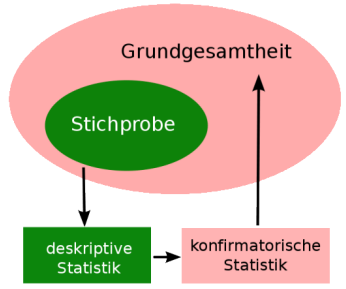


Welche Begriffe sind in Bezug auf Stichproben bekannt?

… siehe Biometrie- / Epidemiologie-Vorlesung

- RCT:
  - Randomisation
  - Verblindung
- Fall-Kontroll-Studie:
  - Matching
- Stichprobengröße (Fehler 1. + 2. Art)
- Bias
- Stratifizierung
- Design-Typen: Paralleldesign, Faktorielles Design, Cross-over-Design …
- Informed Consert → Einverständnis des Probanden / Patienten
- …

Die Fragestellung



Die Frage als Grundlage eines Experiments

… und als Grundvoraussetzung jeglicher Stichprobenüberlegung!

- Auf **welche** Grundgesamtheit bezieht / beziehen sich unsere Fragestellung(en)? ➔ Ein- + Ausschlusskriterien!

Beispiele

- Test eines neuen Antidiabetikas
  - alle Diabetiker / Typ 2 Diabetiker / Typ 2 Diabetiker zu Krankheitsbeginn
  - Für welche Patientengruppe kann das Präparat einen Nutzen bringen? Marktchancen ➔ Sicht der Krankenkassen – Kosten-Nutzen-Relation

---

- Häufigkeit von Tumorerkrankungen im 20-km-Radius von Kernkraftwerken?
  - Wer ist exponiert?
    - Derjenige, der dort wohnt? / arbeitet? / wohnt + arbeitet?

Die Machbarkeit

Welche weiteren Fragen tauchen auf?

- Kenne ich die Grundgesamtheit?
  - Habe ich eine Chance, die Grundgesamtheit in die Stichprobenziehung einzubeziehen?
  - Wie kann ich eine ausreichende Anzahl von Probanden / Patienten erreichen?
- Wie groß ist der Anteil der Teilnehmenden an der Zielpopulation?
  - Wie groß ist die Wahrscheinlichkeit des Drop out?
- Gibt es eventuell so viele Ausschlusskriterien (ethische, methodische…), dass die verbleibende Stichprobe nur noch eine Teilpopulation darstellt?
- Bei berechneter Stichprobengröße:
  - Gibt es überhaupt so viele Patienten? (seltene Erkrankungen)
  - Kann ich die Studie überhaupt finanzieren?

Die Machbarkeit

Beispiele

- Wenn ich alle Einwohner Dresdens befragen will, gehe ich zum Einwohnermeldeamt.
  - ➢ Wohin gehe ich, wenn ich alle Diabetiker Dresdens befragen will?

- Medikamentenstudien
  - Kinder und Schwangere werden häufig aus ethischen Gründen ausgeschlossen.
    - ➢ Was tun, wenn diese erkranken?
  - Multimorbide Patienten werden ebenfalls oft aus ethischen oder methodischen Gründen ausgeschlossen (z.B. wegen Wechselwirkungen mit bestehenden Medikamenten).
    - ➢ Was, wenn Nebenwirkungen des neuen Medikaments gerade die bestehenden Erkrankungen verstärkt (Nieren- / Herzerkrankungen)?

Definitionen

- **Grundgesamtheit (GG)**

  Gesamtheit aller Elemente (Personen, Tiere, Moleküle...), die aufgrund ihrer Eigenschaften als „Merkmalsträger“ für die Studie / Untersuchung in Frage kommen

- **Stichprobe (SP)**

  Repräsentanten der Grundgesamtheit

Repräsentative Stichprobe

Theoretische Zielvorgabe:
Die Stichprobe repräsentiert die Grundgesamtheit in allen Merkmalen

Ad-hoc-Stichprobe

passiert sehr häufig

Stichprobe, die Personen umfasst, die umständehalber gerade zur Verfügung stehen (engl. „samples of convenience“ → „Bequemlichkeitsstichprobe“)

globale Repräsentativität

nahezu alle Merkmale stimmen überein

spezifische Repräsentativität

Merkmale, die im Sinne der Studie relevant sind

Bortz J, Döring N 2002: Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler; 3. Aufl.; Springer; Berlin, Heidelberg, New York, S. 400-406

Definitionen

Konsequenzen (1)

- Fast alle klinischen Studien basieren auf Ad-hoc-Stichproben!
  - SP umfasst die zufällig in den beteiligten Einrichtungen behandelten Patienten
    → umso spezieller die Einrichtungen, umso weniger repräsentativ die Stichprobe!
  - Randomisation ≠ Repräsentativität bezogen auf Grundgesamtheit!
    → Randomisation schafft nur Strukturgleichheit zwischen Gruppen
    → in Randomisation gehen nur bekannte Personen ein, d.h. i.d.R. diejenigen, die die eigentliche Stichprobe bilden!
  - Studien sind somit nur repräsentativ für eine „theoretische“ Grundgesamtheit, die es so in Wirklichkeit oftmals nicht gibt

Beispiele -- Studie zur…

- Wirkung eines Blutdrucksenkers bei Hypertonikern durchgeführt in Herzzentren
- Betreuungsqualität von Hypertonikern durchgeführt bei Mitgliedern des SHÄV

Definitionen

Konsequenzen (2)

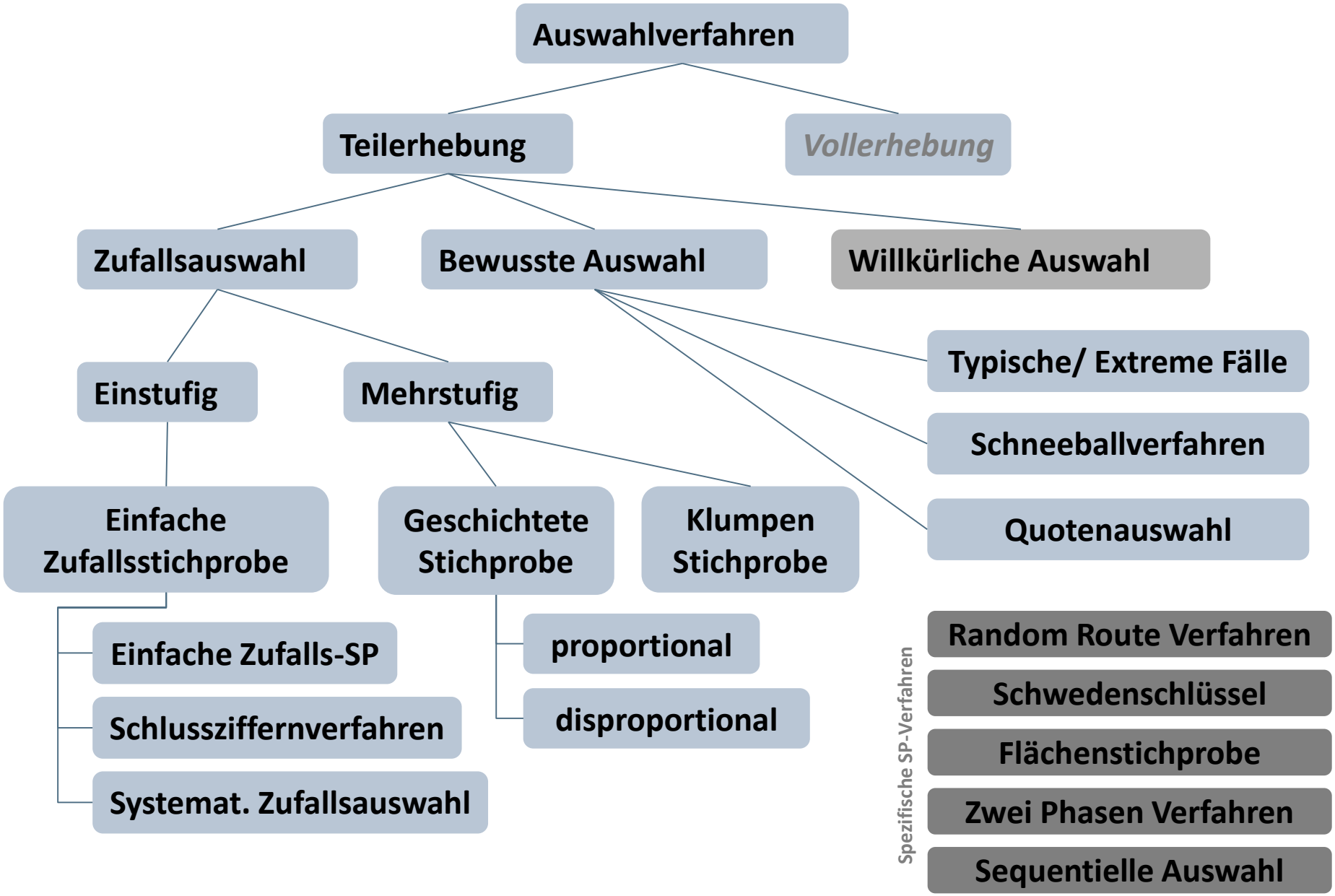
- Große oder sehr große Teilnehmerzahl ≠ Repräsentativität
  - Große Teilnehmerzahlen vergrößern u.U. nur den systematischen Fehler!

Beispiel Wahlforschung

- Prognosen und Hochrechnungen der Wahlforschungsinstitute aufgrund der Angaben von 1.000 – 1.500 Wahlberechtigter stimmen nicht aufs Prozent, sind relativ genau
- Online-Barometer von Zeitungen wie z.B. „Spiegel“ mit mehreren 10.000 Teilnehmern spiegeln die Präferenzen der Leserschaft
- Grund: Wahlforschungsinstitute haben durch jahrzehntelange Erfahrung genaueste Kenntnisse der Sozialstruktur und des Wahlverhaltens und wählen ihre Stichprobe aufgrund dieser Kenntnisse aus
- Vorteil: Der „wahre“ Wert liegt nach der Wahl vor!

Repräsentativität setzt genaueste Kenntnisse über Merkmalsverteilungen in der Grundgesamtheit in Bezug zur Fragestellung voraus!

Zusammenfassung



Auswahlprinzipien

Wie wähle ich aus?

Voraussetzung: Zugriff auf jedes Element muss möglich sein

- **Zufällige Auswahl**

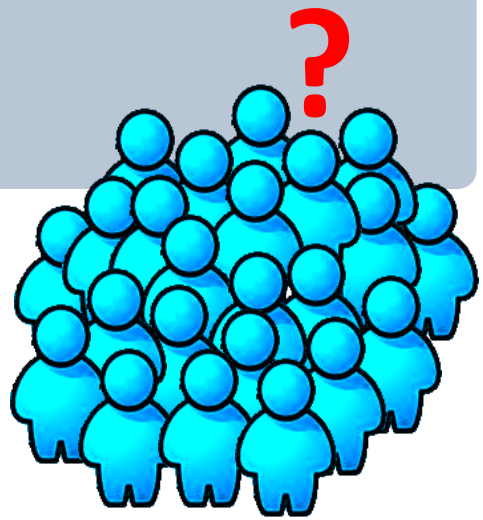
  jedes Element der Grundgesamtheit hat eine identische Wahrscheinlichkeit $p > 0$ Teil der Stichprobe zu werden

- **Bewusste Auswahl**

  Auswahl der Stichprobenelemente basiert auf festen Regeln, aber Stichprobe selbst basiert nicht auf dem Zufallsprinzip

- **Willkürliche Auswahl**

  für die Auswahl der Stichprobenelemente existieren keine festen Regeln



Sampling-Strategien für „Zufällige Auswahl“

Vorbemerkungen

- **Voraussetzung:** Zugriff auf jedes Element der Grundgesamtheit
- **Achtung:** Bei Teilnahmeverweigerung kann eine zufällig ausgewählte Person nicht einfach durch eine andere ersetzt werden, ohne das Prinzip der zufälligen Auswahl zu verletzen!

Welche Verfahren sind möglich?

- **Einstufige Stichprobenauswahl :** Die Auswahlwahrscheinlichkeit ist für jedes Element gleich groß. Die Auswahl erfolgt in einem Schritt.
- **Mehrstufige Stichprobenauswahl aus Schichten:** Vor der Stichprobenauswahl wird die Grundgesamtheit in homogene Gruppen eingeteilt.
- **Mehrstufige Stichprobenauswahl aus Klumpen** (engl. Cluster): Vor der Stichprobenauswahl wird die Grundgesamtheit in homogene Klumpen eingeteilt, die in ihrer Zusammensetzung der Grundgesamtheit entsprechen. Als Stichprobe werden dann ein oder mehrere Klumpen zufällig ausgewählt.
- **Kombinierte Stichprobenverfahren**

Einstufige Stichprobenauswahl (1)

Verfahren

- Einfache Zufallsstichprobe
- Schlussziffernverfahren
- Systematische Zufallsauswahl

Einstufige Stichprobenauswahl (2)

Einfache Zufallsstichprobe (simple random sampling)

- Auch als Urnenprinzip / Lotterieverfahren bekannt
- alle Elemente der Grundgesamtheit werden „durchmischt“ und anschließend „durchnummeriert“ → Gewährleistung der zufälligen Zuteilung
- Auswahl der Stichprobe erfolgt per Zufallszahl

Beispiele

- Einwohnermeldedateien, Epidemiologische Register, Karteien...
- Problem der Aktualität
  - Bsp.: Personen im Alter zwischen 16 und 25 Jahre
    → bei Eltern gemeldet aber eigentlich wohnhaft am Ausbildungsort
  - Umzug, Todesfälle, Namensänderung infolge Heirat oft erst mit (erheblichem) Zeitverzug registriert

Einstufige Stichprobenauswahl (3)

Schlussziffernverfahren

- alle Elemente der Grundgesamtheit werden „durchmischt“ und anschließend „durchnummeriert“
- ausgewählt werden alle Elemente, die mit einer zufällig gewählten Endziffer x enden
- Ziel: einen bestimmten Prozentsatz aus Grundgesamtheit auswählen

Beispiele

- x = 1-stellig = 6 → SP = {6, 16, 26, 36 ....} : SP umfasst 10% der GG
- x = 2-stellig = 58 → SP = {58, 158, 258...} : SP umfasst 1% der GG
- x = 6, 58, 11, 35 und 99: SP umfasst 14% der GG
- **Mikrozensus:** Bei Gemeinschaftsunterkünften erhält jede Person eine eigene fortlaufende "Wohnungs"-Nummer. Alle „Wohnungen“, deren Einerstelle nicht mit drei zufällig gezogenen einstelligen Zahlen übereinstimmt, gelangen in die Stichprobe (70% der Personen).

http://www.gesis.org/unser-angebot/daten-analysieren/amtliche-mikrodaten/mikrozensus/grundfile/mz2006/allg-informationen-2006/

Einstufige Stichprobenauswahl (4)

Systematische Zufallsauswahl (systematic sampling)

- Bei größeren Grundgesamtheiten (z.B. Einwohner eines Bundeslandes) ist „Durchnummerieren“ nicht möglich / zu aufwendig.
- Zufällige Bestimmung eines ersten Elementes
- Ausgehend von diesem Element: Auswahl jedes k-ten Elementes
- „Schrittweite“ $k = N_{GG} / N_{SP}$

Beispiel

- Ziehen einer SP von 500 Einwohnern einer Stadt mit 10.000 Einwohnern
  - k = 10.000 / 500 = 20 → jeder 20. Einwohner wird ausgewählt
  - Beginn: Zufallszahl unter den ersten 20 Einwohnern

Mehrstufige Stichprobenauswahl aus **Schichten** (1)

Hintergrund

- Stichprobe soll Grundgesamtheit möglichst genau repräsentieren
- Um sicherzustellen, dass auch seltene Merkmale in ausreichender Menge in der Stichprobe vorhanden sind, sind meist große Stichproben notwendig.
- Aus Kosten- und ethischen Gründen sind aber möglichst kleine Stichproben gefordert.
- Sicherstellung einer Balanciertheit bezüglich der gewählten Gruppen
- Voraussetzung: Kenntnisse über Zusammensetzung GG / Einflussfaktoren
- Strata sollten eng mit dem Untersuchungsziel zusammenhängen

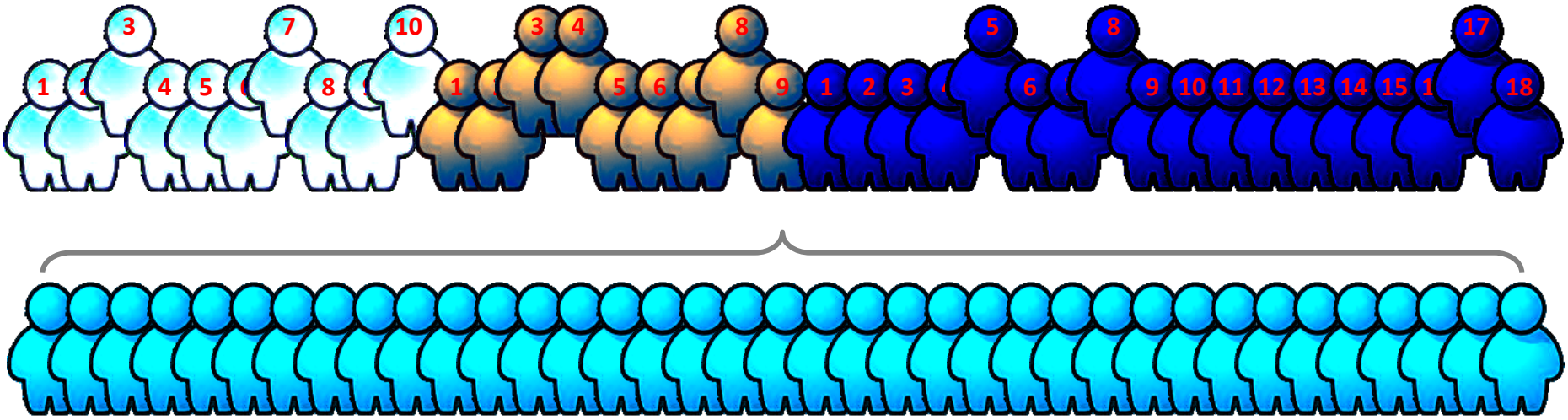
Typische Schichten

- Alter, Geschlecht, Risikofaktoren, Schweregrade der Erkrankung
- Soziale, religiöse, ethnische Faktoren
- Wohnort, Gruppenzugehörigkeit (Partei, Berufsgruppe, Mieter/Eigentümer ...)

Mehrstufige Stichprobenauswahl aus Schichten (2)

Verfahren

- Vor der Stichprobenauswahl wird die Grundgesamtheit in homogene, disjunkte Gruppen (so genannte Strata) eingeteilt. →
  Synonyme Bezeichnungen: Geschichtete, quotierte oder stratifizierte Stichprobenauswahl, engl. stratified sampling
  - Proportionale Stichproben
  - Disproportionale Stichproben



Mehrstufige Stichprobenauswahl aus Schichten (3)

Proportionale Stichproben (proportionate stratification)

- Für jede Gruppe wird ein Anteil gezogen, der dem Anteil der Gruppe in der Grundgesamtheit entspricht.
- Auswahl der Stichprobenelemente innerhalb der Schichten entsprechend „Einstufige Stichprobenauswahl“

Beispiele

- Umfrage zu Ernährungsgewohnheiten
  - Sicherstellung durch Stratifizierung, dass alle Alters-, religiöse und ethnische Gruppen entsprechend ihres Anteils an der deutschen Bevölkerung in der SP enthalten sind
- Telefonumfrage
  - Stratifizierung in Ein- und Mehr-Personen-Haushalte, um Ein-Personen-Haushalte nicht gegenüber Mehr-Personen-Haushalten zu privilegieren

Mehrstufige Stichprobenauswahl aus Schichten (4)

Disproportionale Stichproben (disproportionate stratification)

- Für jede Gruppe wird ein anderer Anteil gezogen (z.B. um für Gruppenvergleiche gleich bzw. genügend große Fallzahlen zu erhalten)
- innerhalb der Schichten: „Einstufige Stichprobenauswahl“
- Vorteil: Aussagen über „Randgruppen“ möglich
- Nachteil: Repräsentativität geht ein Stück weit verloren

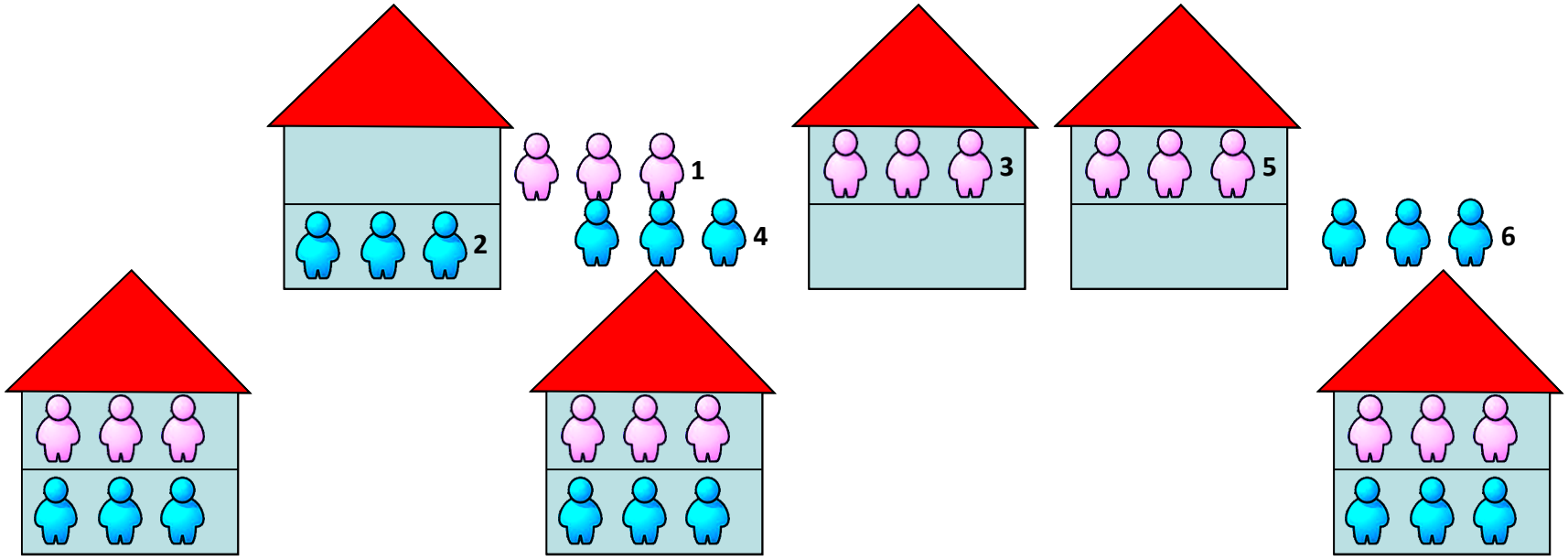
Beispiele

- Einfluss eines Zytostatika auf das Tumorwachstum
  - Patienten werden entsprechend Tumorstadium in Gruppen geteilt
    → je Tumorstadium wird gleich große Anzahl Patienten ausgewählt
- Befragung von Wöchnerinnen zur Wohnsituation
  - Der Großteil der Wöchnerinnen ist verheiratet oder lebt in einer festen Partnerschaft. Um belastbare Aussagen zu allein lebenden Wöchnerinnen zu erhalten, wird eine größere Anzahl von ihnen befragt, als dies ihrem Anteil in der Bevölkerung entspricht.

Mehrstufige Stichprobenauswahl aus **Clustern** (1)

Verfahren

- Vor der Stichprobenauswahl wird die Grundgesamtheit in homogene „Klumpen“ (eng. Cluster) eingeteilt, die in ihrer Zusammensetzung der Grundgesamtheit entsprechen. Als Stichprobe werden dann ein oder mehrere Cluster zufällig ausgewählt. (eng. cluster sampling)
- Die Cluster-Bildung kann auch mehrfach hintereinander durchgeführt werden (multistage sampling).



Mehrstufige Stichprobenauswahl aus Clustern (2)

Beispiel: Wahlprognosen

- 1. Stufe: Cluster = Stimmbezirk → Auswahl von N_1 Stimmbezirken
- 2. Stufe: Cluster = Haushalte im Stimmbezirk → Auswahl von N_2 Haushalten
- 3. Stufe: zufällige Auswahl der zu befragenden Person im Haushalt

Kumulationsverfahren (probability proportional to size sampling)

- Um bei Bevölkerungsstichproben auf Clusterbasis die Einwohnerverteilung zu berücksichtigen, wird ein sogenanntes Kumulationsverfahren angewandt.

Beispiel: Wahlprognosen (ADM-Verfahren Verfahren des Arbeitskreises Deutscher Marktforschungs-Institute)

- Wahlbezirke werden nach Kreisen und Gemeindegrößenklassen geschichtet. Dabei entstehen in Westdeutschland 3.280 Zellen, in Ostdeutschland 1.120.
- Diese werden in möglichst überschneidungsfreie und kumulierbare Teilstichproben (sog. Netze) zerlegt. Ein Netz umfasst 210 Sample Points in Westdeutschland und 48 Sample Points in Ostdeutschland.
- Nun werden die Haushalte gesucht, in denen eine Zielperson zu ermitteln ist.

Einschub: **Spezifische SP-Verfahren** (1)

Random-Route-Verfahren

- Häufiges Auswahlverfahren der Umfrageforschung
- Nach Zufallsprinzip wird für Interviewer ein Startpunkt für Route gewählt.
- Von dort aus muss er eine vorgegebene Laufroute gemäß einfacher Zufallsregeln einhalten, wie
  - geradeaus: Befragung 3. Haus rechts
  - nächste Querstraße links: Befragung 2. Haus links und 5. Haus rechts
  - nächste Querstraße rechts....
- Z.T. werden Angaben um Stockwerke ergänzt (z.B. oberste Klingel links)
- **Problem**: Oft werden diejenigen im Haus angetroffen, die keiner Berufstätigkeit nachgehen (Rentner, Hausfrauen, Arbeitslose) und somit entsteht ein Befragungs-Bias!
- Bei Wahlprognosen: 1. Person ermittelt anhand Random-Route-Verfahren → Adressen → 2. anschließend Person befragt

Einschub: Spezifische SP-Verfahren (2)

Beispiel: Random-Route-Verfahren

Listen Sie zunächst fortlaufend 23 Privathaushalte auf. Sie beginnen in der vorgegebenen Straße (Startstr.) bei der vorgegebenen Hausnummer. Achten Sie darauf, die Hausnummern wie angegeben zu begehen (gerade oder ungerade). Sollte die Anzahl der Haushalte nicht ausreichen, listen Sie auf der gegenüberliegenden Straßenseite weiter auf. Fehlen Ihnen dann immer noch Haushalte, listen Sie die fehlenden Adressen in der ersten Querstraße rechts auf.

Bei kleineren Orten ist es möglich, daß wir Ihnen keine Straße vorgegeben haben, sondern einen Buchstaben. Suchen Sie sich in diesem Fall eine Startstraße, die mit dem angegebenen Buchstaben beginnt. Ist dies nicht möglich, wählen Sie eine Straße mit dem im Alphabet folgenden Buchstaben.

Es gibt auch Ortschaften, in denen es keine Straßennamen gibt, beginnen Sie Ihre Auflistung dann bei einem öffentlichen Gebäude, z. B. bei der Post oder bei der Kirche.

Übertragen Sie die Adressen, die mit einem Kreuz versehen sind auf die beigelegte große weiße Adressenliste; in diesen Haushalten sollen Sie die Interviews durchführen. Leben zwei oder mehr Personen der Zielgruppe in einem Haushalt, muß die Befragungsperson nach dem auf der Adressenliste angegebenen Auswahlschlüssel bestimmt werden.

WICHTIG !!! ⇒ DIE ADRESSENLISTE IST IN GUT LESERLICHER DRUCKSCHRIFT AUSZUFÜLLEN !

LFD-NR	FAMILIENNAME	STRASSE	H-NR	ETAGE	TÜR
01 X	> <				

Einschub: Spezifische SP-Verfahren (3)

Schwedenschlüssel

- ein von Leslie Kish entwickeltes Verfahren zur Zufallsauswahl von Befragungspersonen in Haushalten mit mehreren Personen
- **Ziel:** in ausgewählten Wohnungen/Haushalten soll jede Person die gleiche Chance haben, befragt zu werden
- **Auswahlkriterien:** Haushaltgröße sowie Geschlecht und Alter der einzelnen Personen
- Haushaltsmitglieder werden nach Alter und Geschlecht geordnet in eine Tabelle eingetragen
- Interviewer besitzt Schlüsselliste, welche Person er bei welcher Haushaltsgröße befragen soll

Einschub: Spezifische SP-Verfahren (4)

Beispiel: Schwedenschlüssel

Auswahl der Zielperson (= zu befragende Person)

- Stellen Sie fest, wie viele Personen im **Befragungsalter ab 14 Jahren** im Haushalt leben.
- Tragen Sie das Alter dieser Personen in die Kästchen der Rubrik "Personen im Befragungsalter" ein, und zwar getrennt nach:
  männlichen Personen = **M**, Kästchen-Nr. 1-4
  weiblichen Personen = **W**, Kästchen-Nr. 5-8

  Beginnen Sie jeweils mit der ältesten Person (Kästchen-Nr. 1 = älteste männliche, Kästchen-Nr. 5 = älteste weibliche Person), dann die zweitälteste usw. bis zur jüngsten Person eines jeden Geschlechts.

  (Sollte es einmal vorkommen, daß in einem Haushalt mehr als 4 männliche oder mehr als 4 weibliche Personen leben, so beginnen Sie mit der zweitältesten Person dieses Geschlechts.)
- Danach gehen Sie **von links nach rechts** die Reihe der **Zufallszahlen** durch: ausgewählt (zu befragen) ist diejenige Person, deren Kästchen-Nr. **als erste** in der Reihe der Zufallszahlen erscheint.
- Kringeln Sie diese Zufallszahl und tragen Sie schließlich den Vornamen der Zielperson in die dafür vorgesehene Zeile ein.

Beispiel: Im Haushalt leben 4 Personen im Befragungsalter:
der 49jährige Vater, die 46jährige Mutter sowie ein 19jähriger und ein 16jähriger Sohn.

Personen im Befragungsalter: **ab 14 Jahren**

	1	2	3	4
M	49	19	16	
W	5 46	6	7	8

Zufallszahlen:

8	6	(3)	2	5	1	4	7

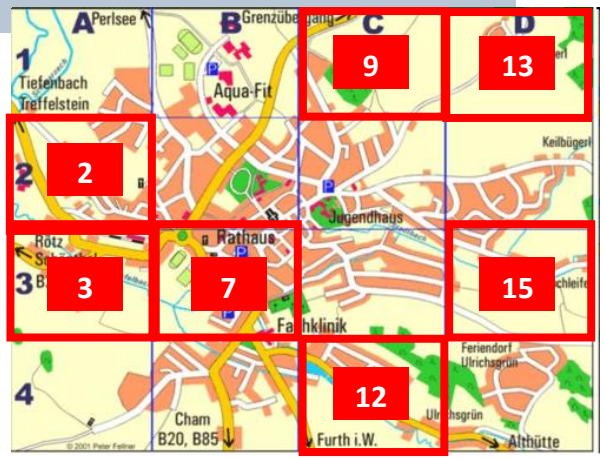
Vorname der Zielperson: Björn
(unbedingt eintragen)

Für das Interview wurde der 16jährige Sohn ausgewählt, weil dessen Kästchen-Nr. als erste in der Reihe der Zufallszahlen erscheint.

Einschub: Spezifische SP-Verfahren (5)

Flächenstichprobe

- In Ländern ohne verwertbare Bevölkerungskarteien oder Einwohnerregister (z.B. USA, Großbritannien) wird ein spezifisches Clusterverfahren angewandt: **Flächenstichprobe**
- Auf Landkarte oder Stadtplan wird das entsprechende Gebiet in Einzelflächen oder Quadrate aufgeteilt.
- Per Zufallszahl wird eine bestimmte Anzahl Quadrate ausgewählt, in denen alle Personen oder eine Stichprobe befragt werden.
- Verfahren wird auch in Land- und Forstwirtschaft angewandt, z.B. zur Ermittlung von Bodenqualität oder Schädlingsbefall.



Einschub: Spezifische SP-Verfahren (6)

Zwei-Phasen-Verfahren (two-phase sampling)

- Existiert keine geeignete Grundlage für Personenauswahl oder ist unklar, wie groß die Grundgesamtheit der infrage kommenden Personen ist, können die bisherigen Sampling-Verfahren nicht angewandt werden.
- Zwei-Phasen-Verfahren
- 1. Phase: große Zufallsstichprobe → Eruieren der für Definition der Grundgesamtheit notwendigen Merkmale
- 2. Phase: Gewünschte Merkmalsträger der 1. Stichprobe bilden Grundgesamtheit für letztendliche Stichprobe.

Beispiel

- Es sollen 300 Personen befragt werden, die ehrenamtlich im Sozialbereich tätig sind.
- 1. Phase: Befragung von 30.000 - 50.000 Personen nach ehrenamtlicher Tätigkeit
- 2.Phase: aus ehrenamtlich Tätigen wird Stichprobe für Befragung gezogen

Einschub: Spezifische SP-Verfahren (7)

Sequentielle Auswahl (sequential sampling)

- Stichprobenumfang wird vor Beginn des Auswahlverfahrens nicht festgelegt
- Beginn mit kleiner Stichprobe → nach Analyse wird festgelegt, ob mit dieser SP die Fragestellung beantwortet werden kann (z.B. Ablehnung von H_0)
- Wenn nicht: Ziehung einer weiteren Stichprobe + Analyse der Gesamtstichprobe
- Vorgang wird so lange wiederholt, bis hinreichender Informationsstand erreicht ist
- **Vorteil:** Minimierung des Stichprobenumfanges
- **Nachteil:** Wahrscheinlichkeit eines falsch-positiven Ergebnisses steigt

Kombinierte Stichprobenverfahren

Verfahren

- Mehrstufiges Sampling-Verfahren, welches verschiedene Sampling-Methoden (je Stufe eines) miteinander kombiniert

Beispiel

- Multicenterstudie zur Behandlungsqualität bei Mammakarzinom
  - 1. Stufe: Clusterverfahren zur Ermittlung der einzubeziehenden Regionen
  - 2. Stufe: Stratifizierte Auswahl der einzubeziehenden Behandlungseinrichtungen (Mammazentren vs. Sonstige)
  - 3. Stufe: Einfache Zufallsstichprobe aus den Patientinnen der letzten 12 Monate

Sampling-Strategien für „Bewusste Auswahl“

Vorbemerkungen

- Nicht alle Elemente der Grundgesamtheit haben gleiche Chance p > 0 in die Stichprobe aufgenommen zu werden
- Bewusste Auswahlverfahren erfolgen nach einem Auswahlplan, dessen Regeln üblicherweise angegeben werden können und die überprüfbar sind.
- Die inferenzstatistischen Techniken sind nicht anwendbar.
- Wenn überhaupt: Aussagen nur über **Inferenzpopulation** möglich
- **Inferenzpopulation** = die Grundgesamtheit, über die auf der Basis der vorliegenden Stichprobe tatsächlich Aussagen gemacht werden können.

Welche Verfahren sind möglich?

- Auswahl typischer Fälle
- Auswahl extremer Fälle
- Auswahl dominierender Fälle
- Schneeballverfahren
- Quotenauswahl

Spezifische Fallauswahl

Spezifische Auswahlkriterien

- **Auswahl typischer Fälle**:
  subjektive Kriterien → spezifische Merkmalsausprägungen
- **Auswahl extremer Fälle**:
  Kriterium → Art der Merkmalsausprägung
- **Auswahl dominierender Fälle** (Konzentrationsprinzip):
  Kriterium → Konzentration bzw. Häufigkeit

  Es wird eine bewusste Konzentration auf einen Teil der Grundgesamtheit vorgenommen, der als wesentlich oder typisch in Bezug auf den Erhebungsgegenstand angesehen wird.

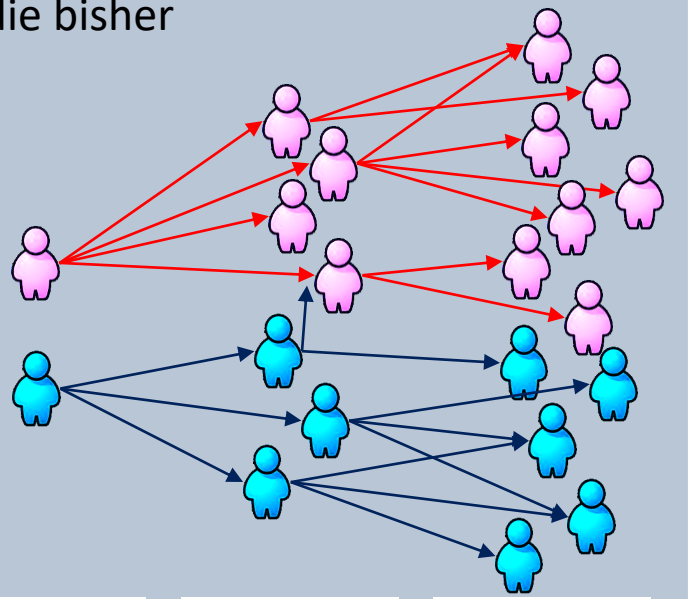
Beispiel Konzentrationsprinzip

- In die Beurteilung von Zusatzleistungen der gesetzlichen Krankenkassen werden nur Kassen einbezogen, die einen Marktanteil von mindestens 10% haben.
  Kleine Kassen bleiben unberücksichtigt.

Schneeballverfahren (1)

Fallauswahl

- **Ausgangspunkt (1.Welle):** Auswahl (ggf. per Zufallsauswahl) von Individuen einer GG oder mit definierten Merkmalen
- Befragung der Probanden
- Probanden werden gebeten, weitere Personen mit definierten Merkmalen zu benennen
- **2. Welle**: Befragung aller genannten Probanden, die bisher noch nicht befragt wurden
- Probanden werden gebeten, weitere Personen mit definierten Merkmalen zu benennen
- **3. Welle**: Befragung aller genannten Probanden, die bisher noch nicht befragt wurden
- Probanden werden gebeten, weitere Personen mit definierten Merkmalen zu benennen usw.



Schneeballverfahren (2)

Vor- / Nachteile

- **Vorteile:**
  - Z.T. nur mit diesem Verfahren bestimmte Personengruppen untersuchbar (Drogenkonsumenten, illegale Einwanderer usw.)
  - Eignet sich zur Untersuchung sozialer Netzwerke
- **Nachteil:**
  - Keinerlei Aussagen über Grundgesamtheit möglich, da diese nicht ermittelbar

Quotenauswahl (1)

Verfahren (quota sampling)

- Wird oft in Marktforschung und anderen Befragungen genutzt, wenn keine geeignete Probanden-Kartei vorhanden ist
- Personen werden so ausgewählt, dass bestimmte Merkmale (Geschlecht, Alter, Beruf, Wohnort etc.) in der Stichprobe genau so häufig vorkommen wie in der Grundgesamtheit
  → **genaue Kenntnis der GG notwendig!**
- Jeder Interviewer bekommt festgelegte Quote zu befragender Merkmalsträger sowie festgelegtes Erhebungsgebiet
- Jedem Interviewer steht es frei, wen er im Rahmen der Quote befragt
- **Hoffnung:** aufgrund Vielzahl eingesetzter Interviewer gleichen sich die willkürlichen Präferenzen des Einzelnen für bestimmte Personentypen insgesamt aus

Quotenauswahl (2)

Es gibt zwei Arten von Quotenplänen (QP)

Nicht ineinander greifender QP

- Quoten der einzelnen Merkmale sind unabhängig voneinander

Ineinander greifender QP

- Quoten bei einer Variable müssen sich auch bei den anderen Merkmalen widerspiegeln

Beispiel

- Bei einer Befragung mit 100 Personen zu Ernährungsgewohnheiten sollen 50 Fitnesscentermitglieder und 50 „Sportmuffel“ befragt werden, davon 50 Männer und 50 Frauen
- **Nicht ineinander greifender QP**: im Extremfall könnten alle Fitnesscenter-mitglieder Männer und alle „Sportmuffel“ Frauen sein
- **Ineinander greifender QP:** Sowohl bei den Fitnesscentermitgliedern als auch den „Sportmuffeln“ müssen jeweils 25 männlich und 25 weiblich sein

Quotenauswahl (3)

Beispiel: Quotenanweisung

Quotenanweisung des Instituts für Demoskopie Allensbach

Name des Interviewers: *L. Mahler*
Wohnort: *Berlin*
Insgesamt: *5* Interviews
im Wohnort/in: *Berlin*

Umfrage 1767

Fragebogen Nr.: *51-55*

Gemeindegröße:										
Gemeinden unter	2000 Einwohner•	1	2	3	4	5	6	7	8	9	10
2000 - unter	5000 Einwohner•	1	2	3	4	5	6	7	8	9	10
5000 - unter	20000 Einwohner•	1	2	3	4	5	6	7	8	9	10
20000 - unter	100000 Einwohner•	1	2	3	4	5	6	7	8	9	10
100000 - unter	500000 Einwohner•	1	2	3	4	5	6	7	8	9	10
500000 und	mehr Einwohner•	1	2	3	4	5	●	7	8	9	10
Alter:			2 männlich					3 weiblich			
16 - 29 Jahre		1	2	3	4	5	1	●	3	4	5
30 - 44 Jahre		1	●	3	4	5	1	2	3	4	5
45 - 59 Jahre		1	2	3	4	5	1	●	3	4	5
60 Jahre und älter		1	●	3	4	5	1	●	3	4	5
Berufstätige:											
Landwirte und mithelfende Familienangehörige in der Land- und Forstwirtschaft (auch Gartenbau und Tierhaltung)		1	2	3	4	5	1	2	3	4	5
Arbeiter (auch Landarbeiter, Facharbeiter, nicht-selbständige Handwerker und Auszubildende)		1	2	3	4	5	1	●	3	4	5
Angestellte und Beamte (auch Auszubildende und Soldaten)		1	●	3	4	5	1	2	3	4	5

Quotenauswahl (4)

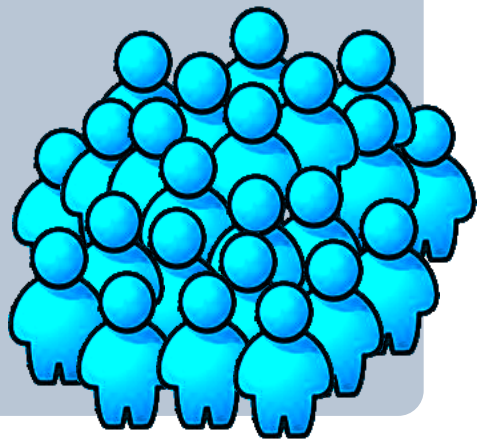
Probleme

- Je mehr Interviews bereits geführt wurden, desto schwieriger wird es, Interviewpartner für die verbliebenen Lücken zu finden
  → Gefahr des Schummelns! → zu komplexe Quotenpläne vermeiden
- Interviewer ist schwächstes Glied: Bildung von privaten Panels wird begünstigt (Herausbildung eines befragungswilligen Personenkreises von Freunden, Verwandten, Nachbarn und Arbeitskollegen)
- Informationen über GG liegen häufig nicht vor bzw. sind nicht aktuell; nur leicht erkennbare demographische Merkmale können verwendet werden.
- Mit der Kontrolle einiger ausgewählter Quotenmerkmale (Alter, Bildung) ist noch nichts über andere soziale Merkmale, z.B. Einstellungen und Verhaltensweisen, gesagt.
- Ausschöpfungsquote und Informationen über Nonresponse fehlen

Willkürliche Auswahl

Verfahren

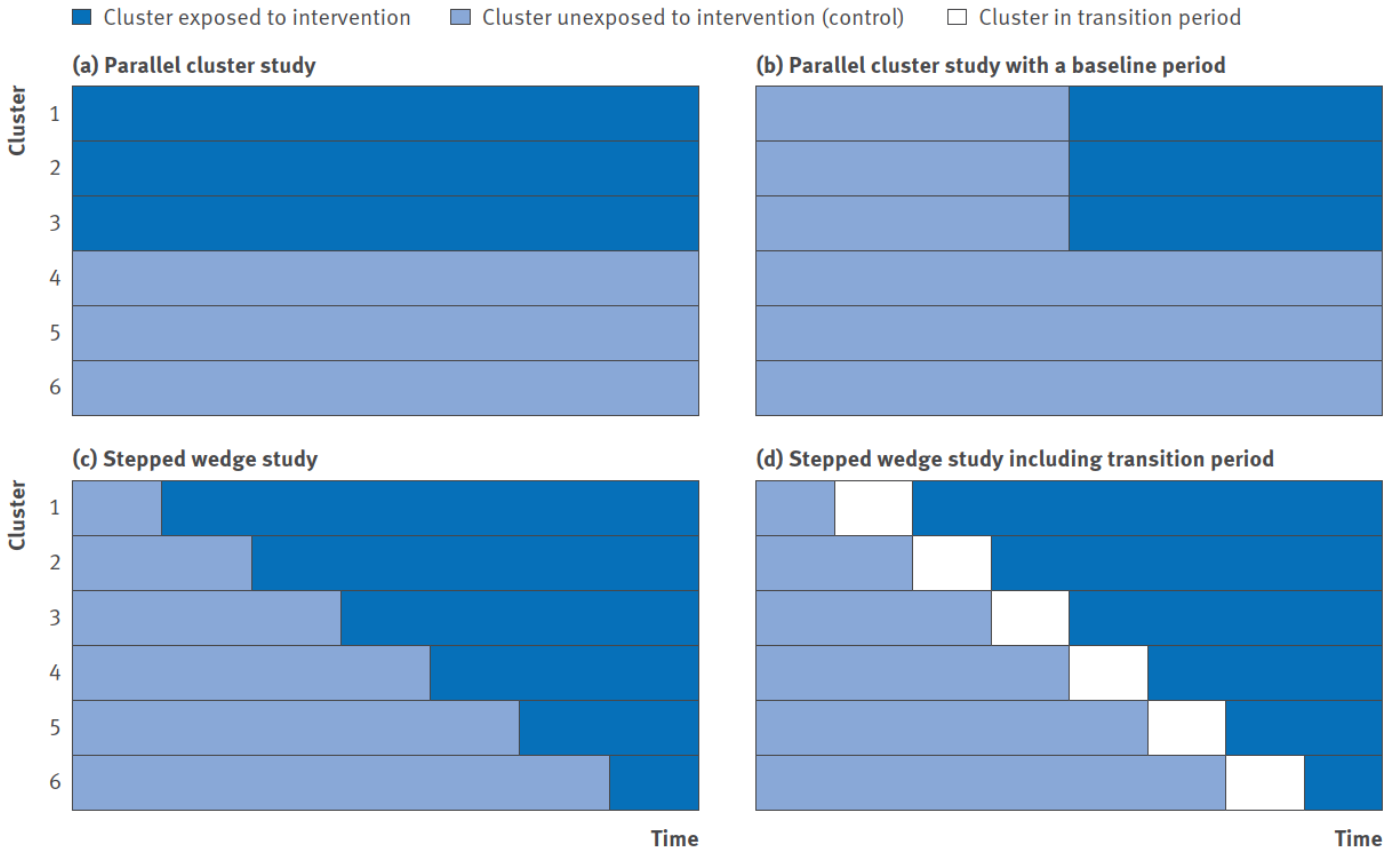
- Als willkürlich werden Auswahlverfahren bezeichnet, bei denen die Entscheidung über die Aufnahme eines Elements der GG in die Stichprobe im Ermessen des Auswählenden liegt.
- I.d.R. werden diejenigen in die Studie / Befragung aufgenommen, die gerade zur Verfügung stehen (Ad-hoc-Stichprobe)
- Bei fast allen klinischen Studien ist dies der Fall, denn i.d.R. entscheidet der behandelnde Arzt darüber, ob er einem Patienten die Teilnahme anbietet oder nicht! Auch kommen nur „eigene“ Patienten infrage.
- **Abgeleitete Designs:** Auswahl einzelner Patienten für RCT, Fall-Kontrollstudien
- Weniger geeignet für Längs- und Querschnittsstudien
- Auch geschichtete Auswahl möglich (z.B. Randomisierungslisten nach Geschlechter / Alter / Krankheitsschwere…)



Abgeleitete Designs für Stichprobenauswahl aus Clustern

Design Randomisierung auf Clusterebene z.B. Krankhäuser, Arztpraxen

- Parallelgruppen Design
- Crossover Design
- Cluster-randomisierte kontrollierte Studie: z.B. Stepped Wedge Design



Stepped Wedge Design (1)

Hintergrund

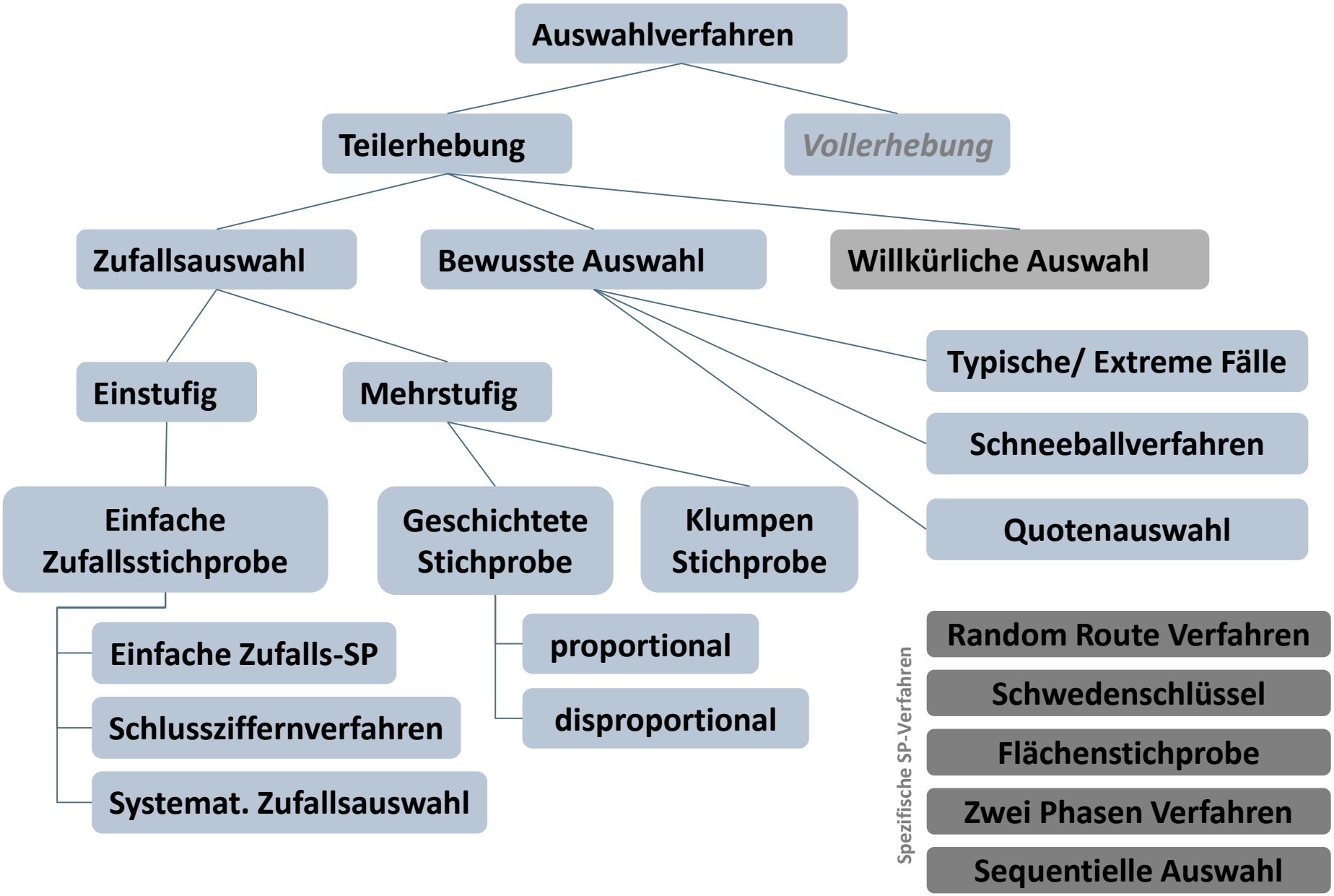
- Komplexe Intervention, die Versorgungsprozesse und -strukturen adressieren, sind häufig schwer nur für einzelne Patient:innen umsetzbar, da Interventionen Auswirkungen auf die Organisationseinheiten (z.B. Arztpraxis, Krankenhaus oder Pflegeeinrichtung) haben
- Randomisierung auf individueller Ebene ist hier aus methodischer Sicht häufig inadäquat → hohes Risiko für Kontaminationseffekte zwischen Kontrolle und Intervention
- Randomisierung auf Ebene Organisationseinheiten (Cluster) → Stepped Wedge Design
- Jedes Cluster startet in Kontrollphase → Intervention startet zeitversetzt
- Cluster werden zufällig den festgelegten Wechselzeitpunkten zugeordnet
- Wechsel kann sich direkt an Kontrollphase anschließen oder nach einem Rollout-Zeitraum

Stepped Wedge Design (2)

Methodische Herausforderungen

- Planung Studienablauf
  - Anzahl Wechselzeitpunkte
  - Anzahl Cluster / Wechselzeitpunkt
  - Länge Zeiträume zwischen Wechseln
- Fallzahlplanung
  - Basis: Fallzahlplanung analog RCT
  - Korrektur bezüglich Designeffekt (Probanden/Cluster, Intra-Cluster-Korrelation, Anzahl Steps)
- Komplexe Auswertungsmodelle
  - Berücksichtigung von Zeiteffekten
  - Mehrebenenmodelle

Zusammenfassung



Einwohnermeldeamt (1)

Stichprobenziehung über Einwohnermeldeämter...

- Einwohnermeldeämter dürfen zu Forschungszwecken so genannte „Bevölkerungsstichproben“ ziehen und Adressen zur Verfügung stellen
- Achtung: Einwohnermeldeämter sind kommunale Behörden, d.h. sie können nur SP für ihren Einzugsbereich ziehen!
- Kosten werden durch Kommunen festgelegt, z.T. sind Auskünfte an öffentliche Einrichtungen kostenlos

Einwohnermeldeamt (2)

Datenumfang...

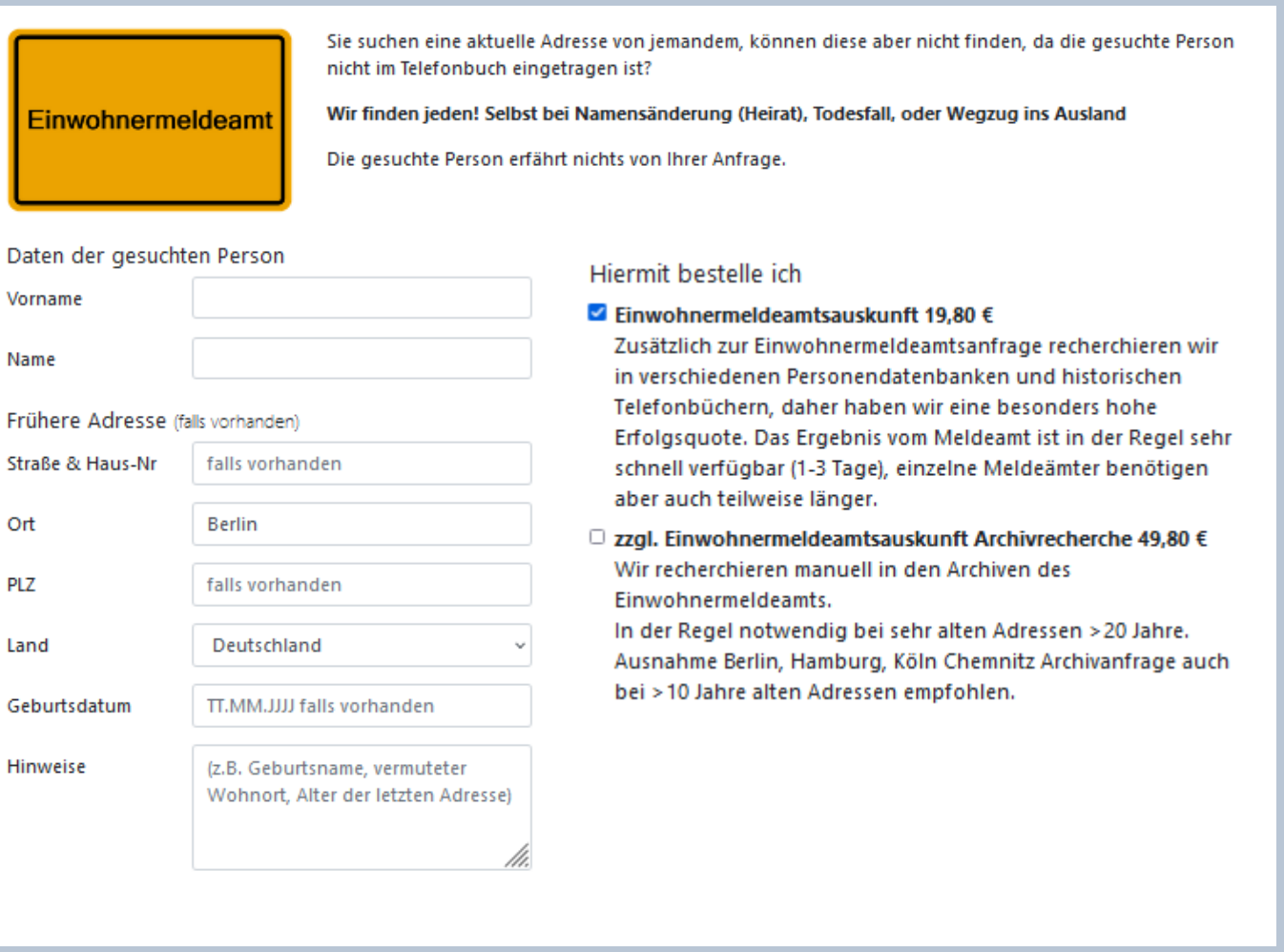
- **Name, Vorname, Doktorgrad, Adresse** → *einfache Melderegisterauskunft*
- Frühere Vor- und Familiennamen
- Tag und Ort der Geburt
- Gesetzliche(n) Vertreter
- Staatsangehörigkeiten → *erweiterte Melderegisterauskunft*
- Frühere Anschriften
- Tag des Ein- und Auszugs
- Familienstand, beschränkt auf die Angabe, ob verheiratet oder eine Lebenspartnerschaft führend oder nicht
- Vor- und Familiennamen sowie Anschrift des Ehegatten oder Lebenspartners
- Sterbetag und –ort

- Religion
- waffenrechtliche Erlaubnis

Einwohnermeldeamt (3)

Wenn ein Proband für eine Nachbeobachtung gesucht wird...

- https://einwohnermeldeamt24.de (privater Anbieter)



Alternative zum Einwohnermeldeamt

Professionelle Adressrecherche

- **z.B.: Adress Research**, eine Marke der Deutschen Post Adress GmbH & Co. KG
  - Kosten für Ermittlung einer Privat- oder Firmenadresse inkl. Überprüfung auf Zustellbarkeit: 8 Euro*
  - Kosten für telefonische Ermittlung einer Adresse: 4,70 Euro*
- Andere Anbieter über Internet zu finden - Kosten ähnlich

* *Seit 2019 keine Kostenangaben gefunden*