

Klassische Test Theory

Jalynskij et al.

Einstieg

Stellen Sie sich vor, Sie dürfen völlig frei ein Thema für eine Forschungsarbeit wählen. Was würden Sie untersuchen?

Viele Sozialwissenschaftlern/innen gemeint ist, ist ein aufgeprägtes Interesse an komplexen (oft unbeobachtbaren) sozialwissenschaftlichen Konstrukten (z.B. Intelligenz, Narzissmus, Ehre, Liebe, Depression. . .).

→ “The burden of complex measurement”

The burden of complex measurement

“Psychological measurement can be a difficult task. We aim to measure not directly observable variables like cognitive abilities, personality traits, motivation, quality of life, diseases in psychopathology, etc. Measuring such latent constructs is much more challenging than determining body height or weight (for which we have a measuring tape and a scale as measurement instruments). Obviously we cannot simply ask: “How depressed are you?” or “How intelligent are you?”. We need a measurement instrument such as a test or questionnaire to assess a participant’s location on the underlying latent variable.” Mair, 2018, S. 1

Die Bürde der Messung komplexer Phänomene

Die Bürde der Messung komplexer Phänomene

...being (almost always) doomed to measure with error when accessing latent constructs.

... zur Kontrolle benötigen wir eine Theorie über die Entstehung des Messfehlers (Messfehlertheorie).

- Beispieltheorie: Klassische Test Theorie (KTT)
- Erklärung: True Score \sim Observed Score \sim Measurement Error.

Lösungsvorschlag: Das True Score Approach

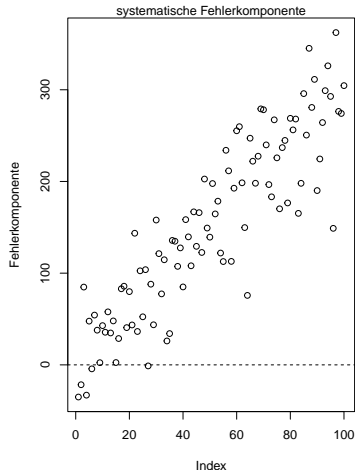
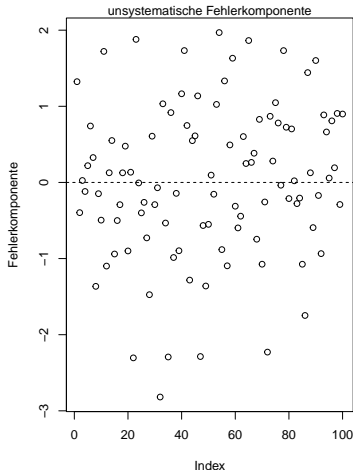
$$\tau = X + \epsilon \Leftrightarrow X = \tau - \epsilon \quad (1)$$

- X : observed score (known)
- τ : true score (unknown)
- ϵ : error component (unknown)

CTT als Messfehlertheorie

Grundannahme

$E(\epsilon) = 0$ (die Fehler sind unsystematisch)



Warum “unsystematisch”?

Fluktuationsphänomen

..Adding small fluctuations tend to dampen one another!

“Consider a simple case of asking students their ages but to insure privacy, asking them to flip a coin before giving their answer. If the coin comes up heads, they should add 1 to their real age, if it comes up tails, they should subtract 1. Clearly no observed score corresponds to the true scores. But if we repeat this exercise 10 times, then the mean for each student will be quite close to their true age”. (Revelle, in prep., S. 207)

Let's do it! (..in R)

```
true_age <- 28 # Verändere mich
add_privacy <- function(true_age){
  # heads or tails
  hot <- sample(c(0,1), 1, replace = TRUE)
  ifelse(hot==1, true_age+1, true_age-1)
}
M <- 10 # Verändere mich!
# Messung duzende (M) Male wiederholen
reps <- replicate(M, add_privacy(true_age))
mean(reps)

## [1] 27.8
```

Prognose & Selbstexperiment

- Wie verändert sich der Wert mit steigender Wiederholungszahl?
- Ab wann sind die Veränderungen konstant?

Example

Versuchen Sie es selbst! Nutzen Sie den Code auf der vorherigen Folie. Geben Sie ihr eigenes Alter ein und verändern Sie den Wert für n . (siehe: Übungsaufgabe 1)

Anmerkung: Sie sollen die Konzepte nicht jede Zeile des Codes verstehen!

Die True Score Logik: ein Gedankenexperiment

Gedankenexperiment

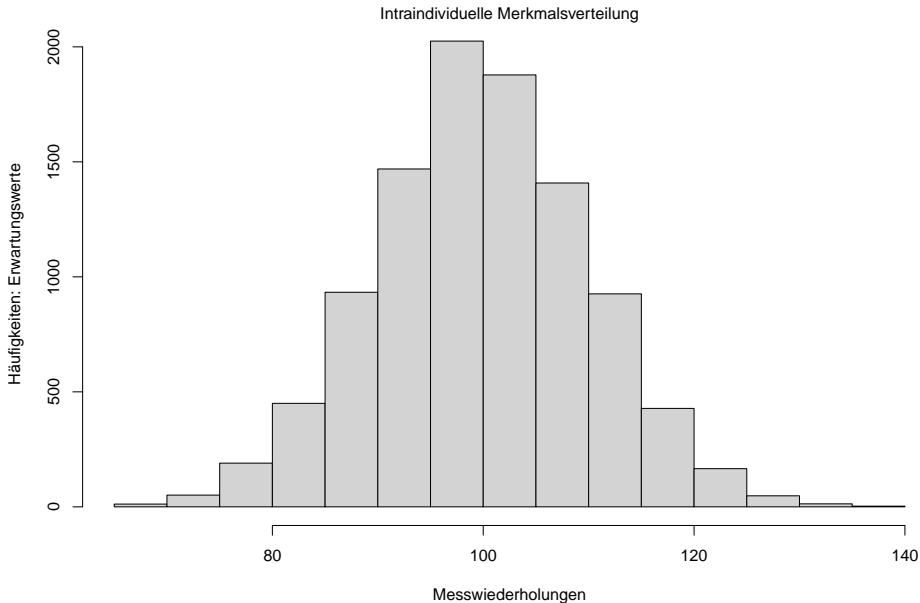
Hätten wir die Möglichkeit eine Person i unzählige Male zu unter den gleichen Bedingungen zu messen ($m=1,\dots,M$), dann erhalten wir eine (Normal-)verteilung seiner Messwerte (unter der Annahme $E(\epsilon) = 0$), wobei im Mittel über alle intraindividuellen Messwerte ($X_j = 1, \dots, M$) die Beobachtungswerte dem wahren Wert der Person entsprechen

$$E(X_{j=1,\dots,M}) = \tau \quad | \quad \epsilon \sim \text{Normal}(0, \sigma) \quad (2)$$

Let's do it! (..in R)

```
set.seed(123)
# unzählige Male
M <- 1e4 # Verändere mich!
# True Score: i (z.B. Intelligenztest)
T <- 100
# random error; Schwankungsbreite +/- 10
E <- rnorm(M, 0, 10)
# Measurements
X <- T + E
# Imaginäre Wdh der Messung
reps <- replicate(M, sample(X, 1))
```

Ergebnisse: Intraindividuelle Merkmalsverteilung



Ergebnisse: Beobachtungs- & Erwartungswert

```
set.seed(123)  
# Ein tatsächlicher Beobachtungswert (Zufallszug)  
(X_i <- sample(X, 1))
```

```
## [1] 88.09457
```

```
# Erwartungswert des Individuums:  $T = E(X)$   
(E_X <- mean(reps))
```

```
## [1] 99.89027
```

Prognose & Selbstexperiment

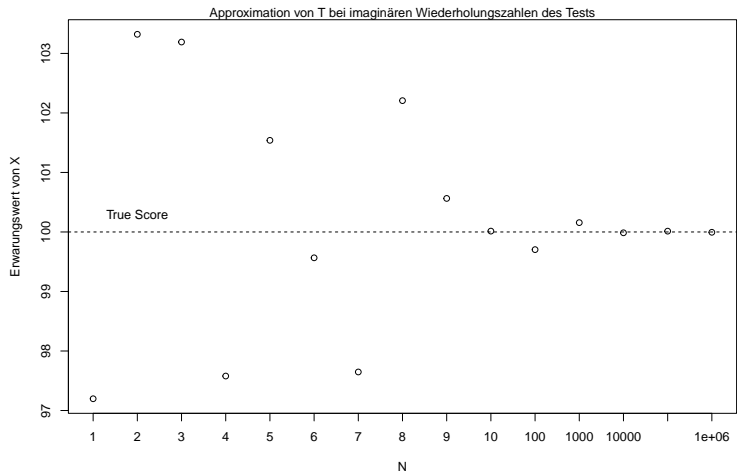
- Wie verändert sich der Wert mit steigender Wiederholungszahl?
- Ab wann sind die Veränderungen konstant?

Example

Versuchen Sie es selbst! Nutzen Sie den Code auf der vorherigen Folie. Verändern Sie den Wert für M und überprüfen Sie Ihre Prognose. (siehe: Übungsaufgabe 2)

Anmerkungen: Wollen Sie unsere Ergebnisse exakt reproduzieren setzen Sie folgendes Snippet vor jedes Codebeispiel

```
# Set RNG state  
set.seed(123)
```

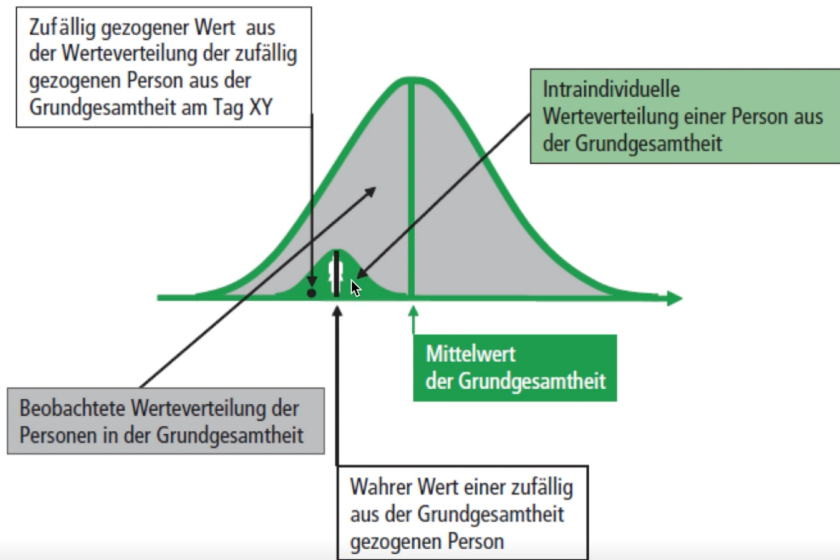


True Score Logik

Um die True Score Logik zu verinnerlichen, wollen wir die einzelnen Konzepte in R umsetzen und mit den Kennwerten “spielen” um eine Intuition für sie zu bekommen; v.a.:

- Interindividuelle Merkmalsverteilung
- Intraindividuelle Merkmalsverteilung
- True, Observed & Error Score – $\tau.X, \epsilon$

Zur Verortung: Kernkonzepte



Anwendungsbeispiel: Intelligenztest

Gedankenexperiment

Stellen Sie sich vor wir würden mit 10.000 Probanden einen Intelligenztest durchführen, indem jeder Proband 5000 mal den gleichen Intelligenztest wiederholt.

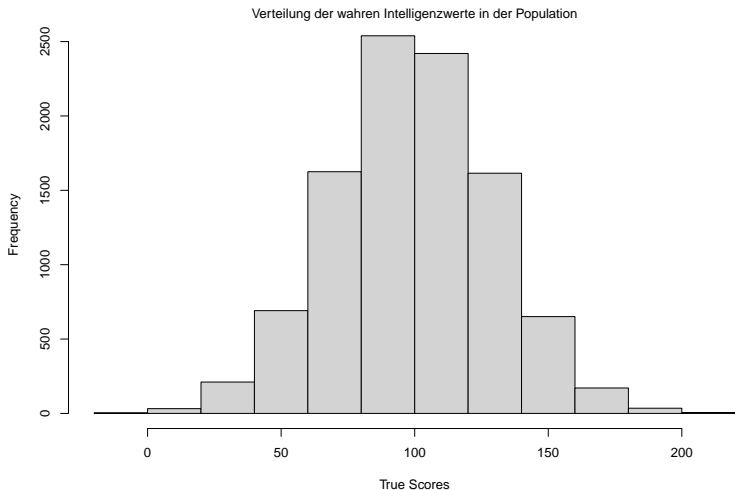
Populationen: Individuen & Messwerte

..dazu ziehen wir (zufällig) aus einer normalerweise unbekannten Population $N=10.000$ Personen mit den True Scores (τ)

```
# Populationsgröße (N)  
N <- 1e4 # (1e4): Verändere mich!  
# Generierung der True Scores  
# 100: Mittlere Intelligenz in der Population  
# 30: Abweichungen vom Populationnsmittelwert  
T <- round(rnorm(N, 100, 30), digits = 0)  
# Anzahl der (sonst imaginären) Testwiederholungen  
M <- 5000 # (5000): Verändere mich!  
X <- lapply(T, function(T) rnorm(M, T , 5))  
names(X) <- T
```

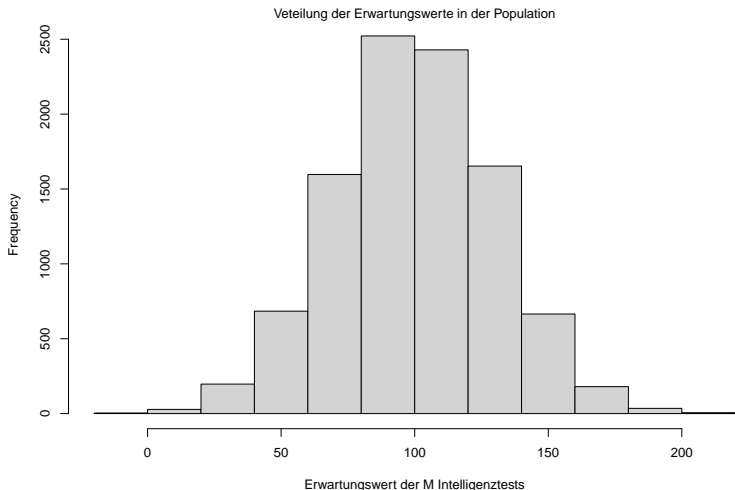
Verteilung der True Scores in der Population

...die uns unbekannten True Scores sind gemäß der Simulation wie folgt verteilt



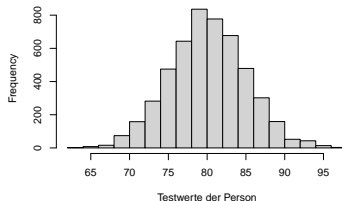
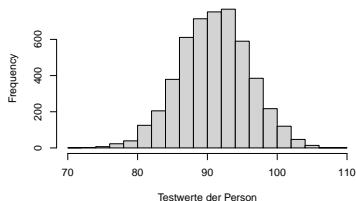
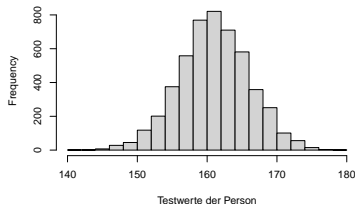
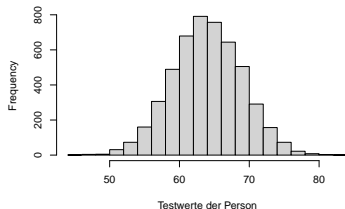
Interindividuelle Merkmalsverteilung

... mitteln wir die 5000 durchgeführten Testergebnisse ($X_{j=1,\dots,5000}$) für jede Person ($i = 1, \dots, 10000$) und plotten diese, dann erhalten wir die interindividuelle Merkmalsverteilung ($E(X) \approx T$)



Intraindividuelle Merkmalsverteilungen

... aus dieser Population (interindividueller Merkmalsverteilung) ziehen wir zufällig 4 Probanden/Probandinnen und betrachten ihre intraindividuellen Merkmalsverteilung über die 1000 Testwiederholungen.



Intraindividuelle Merkmalsverteilungen

..da wir nur ein einziges Mal messen, ziehen wir (zufällig) einen Wert aus der (Normal-)verteilung der intraindividuellen Messwerte. Dieser intraindividuelle Messwert (X_{ij}) ist ein Schätzer für den True Score (τ_i) der Person der wegen des Messfehlers unter oder über dem Messwert liegt.

```
X_i <- sample(X, 1)
# Sample randomly a testvalue
(X_ij <- sample(X_i[[1]], 1)) ; names(X_i)
```

```
## [1] 63.09449
```

```
## [1] "64"
```


Example

Versuchen Sie es selbst! Nutzen Sie nachfolgende Funktion und führen sie sie immer wieder aus. Damit ziehen Sie aus unserer Population immer wieder zufällig neue Probanden. Achten sie darauf, wie sich True Score, Observed Score und Messfehler verändern. (siehe: Übungsaufgabe 3)

- Funktion einlesen: Makieren + Str/Ctrl/Cmd-Enter
- Mehrfach ausführen: Makieren + Str/Ctrl/Cmd-Enter
- Ergebnis auf Folie wiederholen:

Zufallsziehung

```
rsample_i <- function(){  
  # Sample (randomly) an individual  
  (X_i <- sample(X, 1))  
  # Sample randomly a testvalue  
  (X_ij <- sample(X_i[[1]], 1))  
  cat(" True Score (T):", names(X_i), "\n",  
      "Testwert (X):", round(X_ij), "\n",  
      "Messfehler (E):",  
      abs(round(X_ij) - as.numeric(names(X_i))))}  
  # Sample an individual randomly!  
  rsample_i()
```

```
## True Score (T): 64  
## Testwert (X): 63  
## Messfehler (E): 1
```

Prognose & Selbsttest

- Wie wirken sich Veränderungen in der Wiederholungszahl des Tests (M) aus?
- Was geschieht bei Veränderungen der Populationsgröße (N)?
- Wie wirken sich die Veränderungen auf den Zusammenhang zwischen True Score, Observed Score und Messfehler aus (Tipp: $\tau = X + \epsilon$)

Example

Versuchen Sie es nun selbst! In vorherigem Code finden Sie die Anmerkung `Verändere mich!` (siehe v.a. N , M). Erstellen Sie ihre eigenen Populationen und variieren Sie systematisch die Werte M und N . (siehe: Übungsaufgabe 4)

Lokale Unabhängigkeit

Lokale Unabhängigkeit

Lokale Unabhängigkeit \Leftrightarrow die Itemantworten sind unter Kontrolle der Traitausprägung unabhängig voneinander

- Die Logik latenter Variablen (2.0)

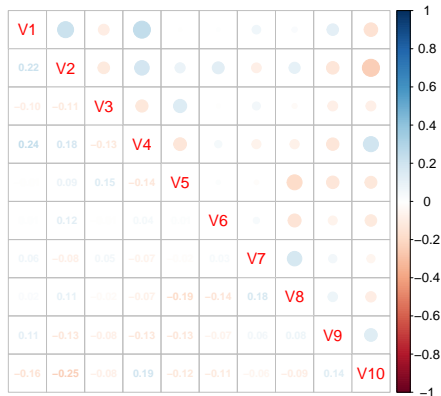
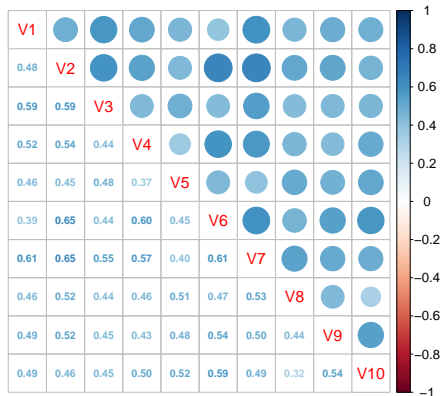
$$\zeta : \text{gen.process} \rightarrow \text{VAR}(j = 1, \dots, M) : \text{cor}(j, k) > 0 \quad (3)$$

- Die Logik lokaler Unabhängigkeit

$$\text{cor}(j, k) | \zeta = 0 \quad (4)$$

- ζ : Konstrukt (latente Variable)
- j, k : Items in einem Test (beobachtete Variablen)
- M : Anzahl der Items in einem Test

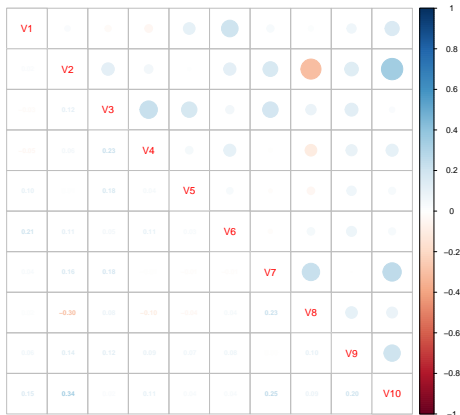
Lokale Unabhängigkeit \sim Korrelationsmatrizen



Example

Hier sehen Sie zwei Korrelationsmatrizen. Besprechen Sie sich mit Ihren Nachbarn. Wenden Sie das Prinzip lokaler Unabhängigkeit auf die 3 Korrelationsmatrizen an. Zu welchem Ergebnis kommen Sie?

Lokale Unabhängigkeit \sim Korrelationsmatrizen II



Example

Stellen Sie sich vor, nach Kontrolle der Traitausprägung sieht Ihre Korrelationsmatrix wie folgt aus. Zu welchen Schluss kommen Sie bezüglich der Annahme lokaler statistischer Unabhängigkeit? Warum?

..von der additive Varianzzerlegung zu Definition der Reliabilität

Definition: Reliabilität

$$Rel(Y_j) = \frac{Var(\tau_j)}{Var(Y_j)} = \frac{Var(\tau_j)}{Var(\tau_j) + Var(\epsilon_j)} \quad (5)$$

Die Reliabilität wird definiert als Anteil ($Var(\tau_j)$) der wahren Varianz an der Gesamtvarianz ($Var(Y_j)$).

Prognose & Selbsttest

Wie wirken sich eine steigende Varianz im Fehlerterm auf die Reliabilität der Messung aus? (Tipp: $Rel(Y_j) = \frac{Var(\tau_j)}{Var(\tau_j) + Var(\epsilon_j)}$)

Example

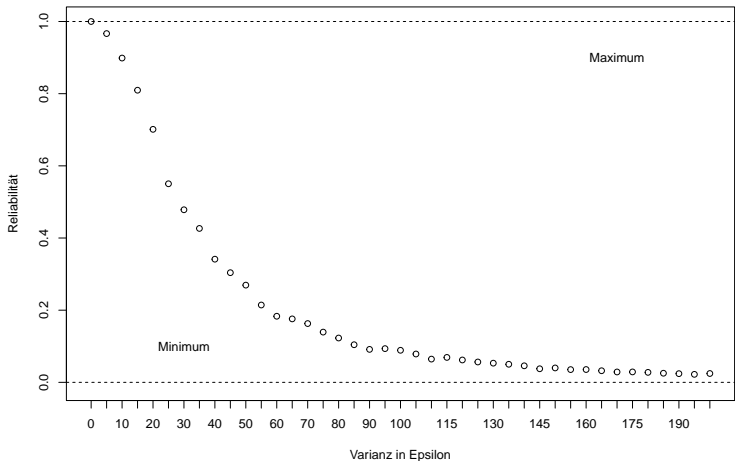
Versuchen Sie es nun selbst! In folgendem Code finden Sie die Anmerkung `Verändere mich!`. Variieren Sie die Werte und überprüfen Sie ihre Prognose. (siehe: Übungsaufgabe 7)

Reliabilität \sim Messfehler

```
N <- 1000 # Verändere mich
tau <- rnorm(N, mean = 100, 30)
var_epsilon <- 25 # (25): Verändere mich!
epsilon <- rnorm(N, 0, var_epsilon)
reliab <- function(tau, epsilon){
  rel <- var(tau) / var(tau + epsilon)
  cat("Reliabilität der Messung:", rel)
}
reliab(tau, epsilon)
```

```
## Reliabilität der Messung: 0.5357339
```

Auflösung



... Die Reliabilität ist also ein Maß für die Messfehlerfreiheit einer Messung

Exkurs: Zum Problem(?) systematischer Störeinflüsse

Ausgangssituation

Da die KTT nur unsystematische Störeinflüsse berücksichtigt, bleiben systematische Einflüsse unberücksichtigt.

z.B.: soziale Erwünschtheit, politische Korrektheit

Frage: Wie wirken sich systematische Störeinflüsse aus?

- z.B. auf Korrelationen?
- z.B. auf Regressionsgewichte?

Auswirkungen system. Störeinflüsse: Korrelationen

Szenario

Systematischer Einfluss über alle Fälle hinweg

```
n <- 100
x <- rnorm(n)
y <- rnorm(x)
cor(x, y) ; cor(x+3, y) ; cor(x, y + 3) ; cor(x + 3, y + 3)

## [1] -0.04953215
## [1] -0.04953215
## [1] -0.04953215
## [1] -0.04953215
```

Auswirkungen system. Störeinflüsse: Korrelationen

Szenario

Systematischer Störeinfluss auf einige (die ersten 50) Probanden

```
x[1:50] <- x[1:50] + 3  
cor(x, y)
```

```
## [1] -0.1602125
```

```
# Additional increase MME
```

```
x[1:50] <- x[1:50] + 6  
cor(x, y)
```

```
## [1] -0.1608405
```

Auswirkungen system. Störeinflüsse: Regressionsgewichte

Szenario

Systematischer Einfluss über alle Fälle hinweg

```
x <- rnorm(100) ; y <- rnorm(x)
```

```
# Original result
```

```
lm(y ~ x)$coef[[2]]
```

```
## [1] -0.05247161
```

```
# Manipulated predictor
```

```
beta_ast <- lm(y ~ I(x+3))$coef[[2]]
```

```
# Manipulated outcome
```

```
lm(I(y+3) ~ I(x))$coef[[2]]
```

```
## [1] -0.05247161
```

Auswirkungen system. Störeinflüsse: Regressionsgewichte

Szenario

Systematischer Störeinfluss auf einige (die ersten 50) Probanden

```
x <- rnorm(100) ; y <- rnorm(x)
x[1:50] <- x[1:50] + 3
beta <- lm(y ~ x)$coef[[2]]
# Manipulated predictor
x[1:50] <- x[1:50] + 6
# Manipulated coefficient
lm(y ~ x)$coef[[2]]

## [1] -0.03411985
```

Messmodelle (..to be continued!)

Messmodelle & Messäquivalenz

Inwieweit messen wir mit unterschiedlichen Tests dasselbe Konstrukt?

Anmerkung: unterschiedliche Messmodelle bedingen unterschiedliche Reliabilitätskoeffizienten.

..Fortsetzung in der Übung zur Reliabilität folgt.

“The classical test theory model is the theory of psychological testing that is most often used in empirical applications. The central concept in classical test theory is the true score. True scores are related to the observations through the use of the expectation operator: the true score is the expected value of the observed score. Thus, a researcher who sees intelligence as a true score on an intelligence test supposes that somebody’s level of intelligence is his expected score on an IQ-test.” (Borseboom 2005, S. 3)

Abschließende Anmerkung

“However, although various authors have warned against it, the platonic true score interpretation is like an alien in a B-movie: no matter how hard you beat it up, it keeps coming back.”
(Borseboom 2005, S. 32)

Probleme der Klassischen Test Theory

- z.B.: die Grundgleichung bleibt empirisch ungeprüft ($\tau = X + \epsilon$)
- z.B.: Annahme zu Messfehlern oft problematisch ($E(\epsilon) = 0$)
- z.B.: Kohli et al. (2015) Nur moderne Erweiterungen der KTT (v.a. Underlying Normal Variable Approach) können annähernd mit modernen Item Response Theory Model (IRT) mithalten.

⇒ Einblick in die IRT