Klassische Testtheorie

Bißantz, Jalynskij, Kupffer & Prestele

BF3 Testtheorie

Section 1

Einstieg: Forschungsthemen-Wunschliste

Einstieg: Forschungsthemen-Wunschliste

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

Auflösung: das Interesse an komplexen Phänomenen

Viele Sozialwissenschaftler/innen haben ein ausgeprägtes Interesse an komplexen (oft unbeobachtbaren) sozialwissenschaftlichen Konstrukten (z.B. Intelligenz, Narzissmus, Ehre, Liebe, Depression...).

$\longrightarrow$ die Bürde der Messung komplexer Phänomene

Zentrale Konzepte der KTT sind hierbei:

1. Messfehler
2. Wahrer Wert
3. Reliabilität

Section 2

Die Bürde komplexer Messungen

Die Bürde komplexer Messungen: Messfehler

"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)

Zusammenhang: komplexe Messung & Testtheorie

- Ausgangspunkt

Die Bürde der Messung komplexer Phänomene

Als Sozialwissenschaftler*inn ist man häufig dazu "verdammt", latente Konstrukte mit *Messfehler* zu erfassen.

$\Rightarrow$ *Theorie über die Entstehung des Messfehlers* (Messfehlertheorie).

- Beispieltheorie: Klassische Testtheorie (KTT)
  - Erklärung: Wahrer Wert = Beobachtungswert + Messfehler
- (...demnächst: Item Response Theory (IRT))

Section 3

Messfehler & True Scores in der KTT

Zusammenhang: Messfehler & True Score

Aus der formalen Darstellung der Bürde der Messung komplexer Phänomene..

$$X = \tau - \epsilon \tag{1}$$

..folgt logisch äquivalent..

Definition: True Score

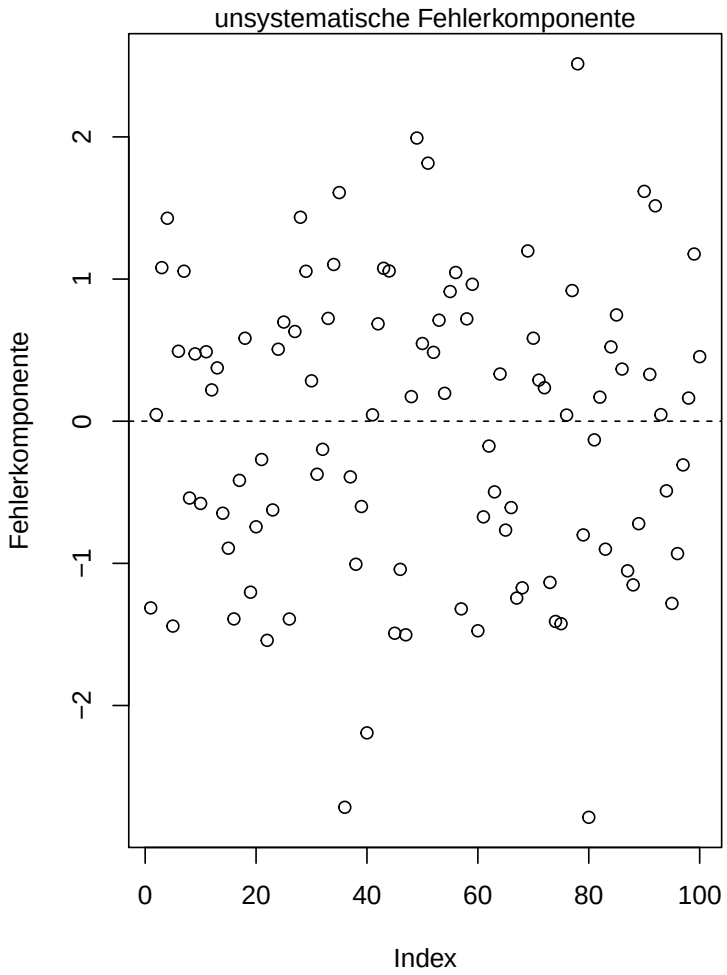
$$\tau = X + \epsilon \tag{2}$$

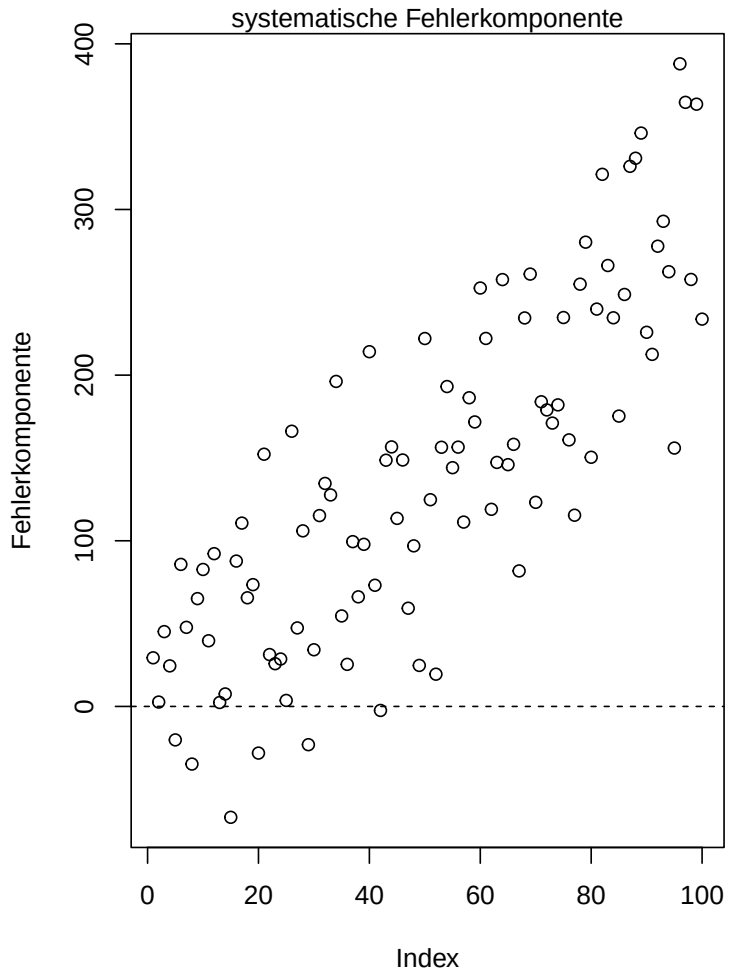
- X: Beobachtungswert (bekannt)
- τ: Wahrer Wert (unbekannt)
- ϵ: Fehlerkomponente, Messfehler (unbekannt)

Grundannahme der Messfehlertheorie (KTT)

Grundannahme: unsystematische Fehlerkomponente

$$E(\epsilon) = 0 \tag{3}$$





Warum eine unsystematische Fehlerkomponente?

Fluktuationsphänomen

"Adding small fluctuations tend to dampen one another!" (McElreath, 2020, S.73)

Kniff: Unter der KTT-Annahme "unsystematischer Störeinflüsse" (=Fluktuationen), gleichen sich die Messfehler im Mittel aus, sodass der Messwert (X) dem wahren Wert (τ) entspricht.

$$E(X_{j=1,\ldots,M}) = \tau \quad | \quad \epsilon \sim Normal(0, \sigma) \tag{4}$$

1. Beispiel: Revelle's Münzwurf
2. Beispiel: die verflixte Waage

Beispiel 1: Fluktuationsphänomen (Revelle's Münzwurf)

"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
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)
}
# Anzahl der Messwiederholungen
M <- 10
reps <- replicate(M, add_privacy(true_age))
mean(reps)
```

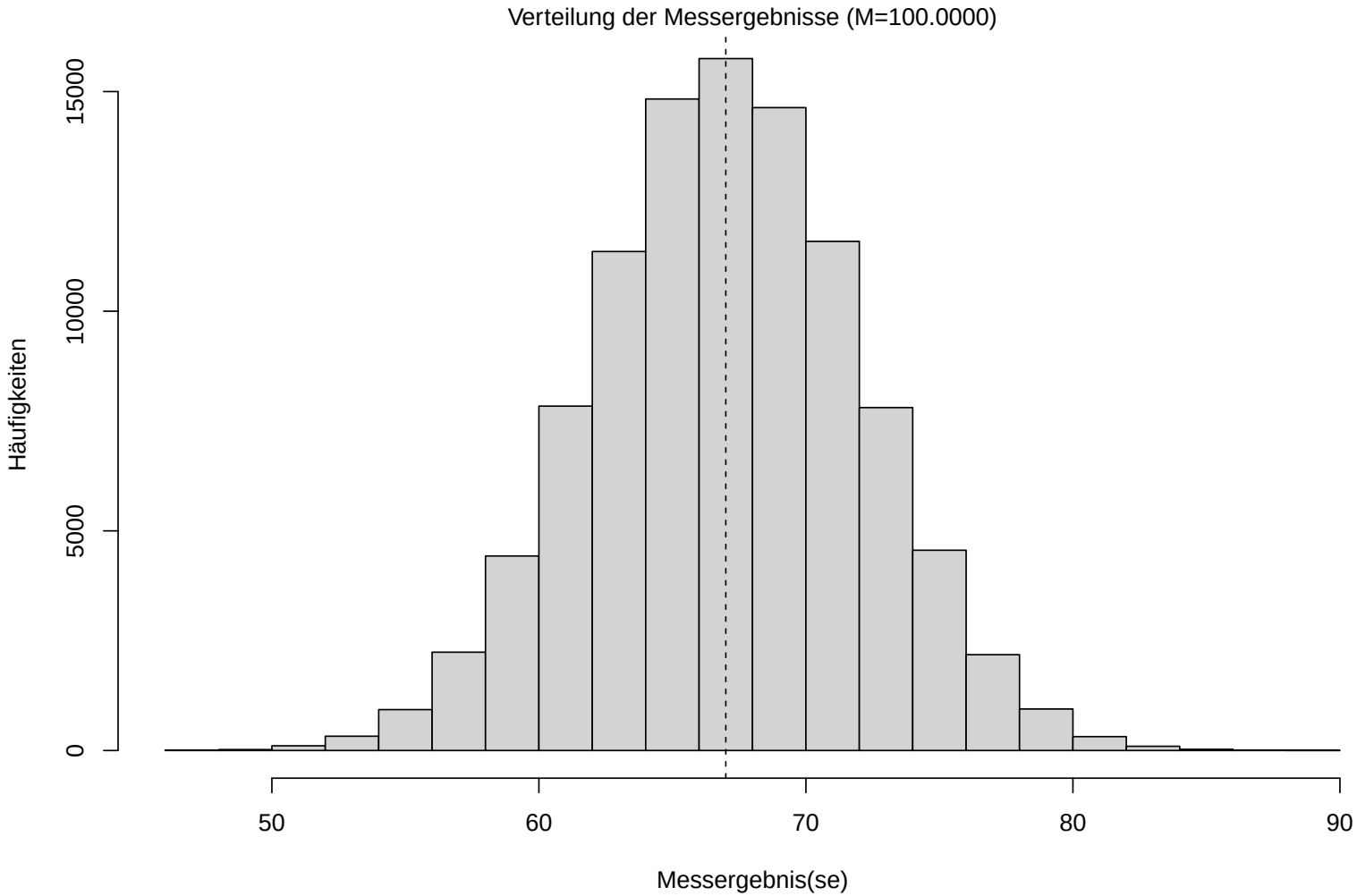
```
## [1] 27.8
```

Beispiel 2: Fluktuationsphänomen (verflixte Waage)

```r
true_weight <- 67
crazy_scale <- function(true_weight){
  # 5kg +/-
  rnorm(1, true_weight, 5)
}
# Anzahl der Messwiederholungen
M <- 100000
reps <- replicate(M, crazy_scale(true_weight))
mean(reps)
```

```
## [1] 67.00488
```

Grafik: Die verflixte Waage



Übungsaufgabe 1: Selbstexperiment

Example

Versuchen Sie es selbst! Nutzen Sie den Code zur Übungsaufgabe 1 in `04-KTT.R`. Geben Sie ihr eigenes Alter/Gewicht ein und verändern Sie den Wert für `M`.

1. Wie verändert sich der Wert mit steigender Wiederholungszahl?
2. Ab wann sind die Veränderungen konstant?

- Zeit: 5 Minuten
- Replikation: `set.seed(123)`
- Anmerkung: Konzepte verstehen $\gg$ Codes verstehen!

True Score Logik: ein Gedankenexperiment

..in (etwas) abstrakteren Worten entspricht die True-Score-Logik der KTT einem Gedankenexperiment (siehe VL-Folien)

Gedankenexperiment

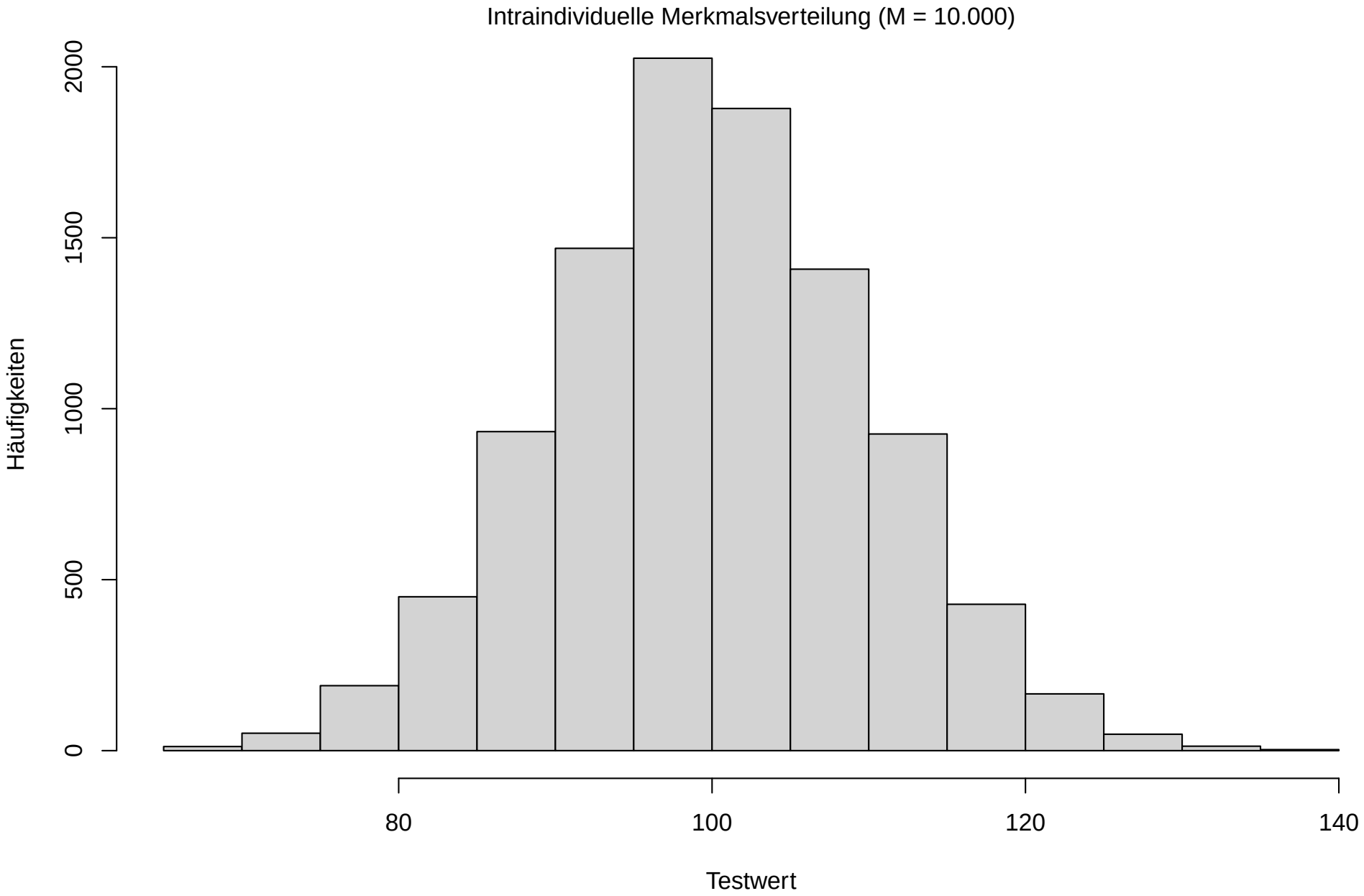
Hätten wir die Möglichkeit eine Person i unzählige Male zu unter den gleichen Bedinungen zu messen (m=1,...,M), dann erhalten wir eine (Normal-)verteilung seiner Messwerte (unter der Annahme $E(\epsilon) = 0$)), wobei im Mittel über alle *intraindividuellen* Messwerte ($X_j = 1, \ldots, M$) die Beobachtungswerte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

$$E(X_{j=1,\ldots,M}) = \tau \quad | \quad \epsilon \sim Normal(0, \sigma) \tag{5}$$

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

```r
set.seed(123)
# Anzahl der Messwiederholungen
M <- 10000
# True Score(s) (z.B. Intelligenztest)
T <- 100
# Zufallsfehler; Schwankungsbreite +/- 10
E <- rnorm(M, 0, 10)
# Beobachtungswerte
X <- T + E
# (Imaginäre() Wiederholung der Testung
reps <- replicate(M, sample(X, 1))
```

Ergebnisse: Intraindividuelle Merkmalsverteilung



Ergebnisse: Beobachtungs- & Erwartungswert

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

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

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

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

Übungsaufgabe 2: Selbstexperiment

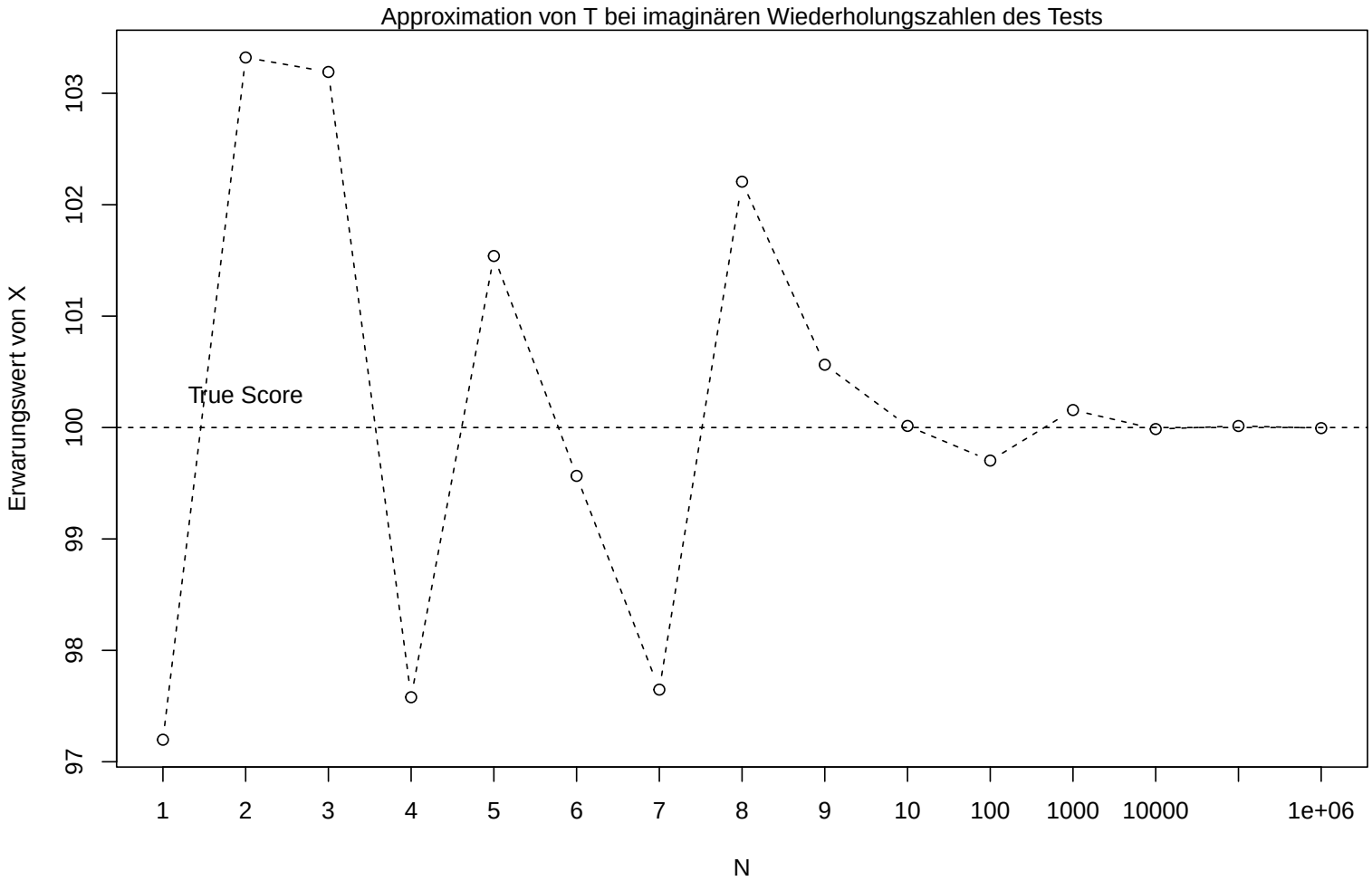
Example

Versuchen Sie es selbst! Nutzen Sie den Code zur Übungsaufgabe 2 in `04-KTT.R`. Geben Sie ihr eigenes Alter/Gewicht ein und verändern Sie den Wert für `M`.

1. Wie verändert sich der Wert mit steigender Wiederholungszahl?
2. Ab wann sind die Veränderungen konstant?

- Zeit: 10 Minuten
- Replikation: `set.seed(123)`
- Anmerkung: Konzepte verstehen $\gg$ Codes verstehen!

Auflösung: Veränderung in Abhängigkeit der Wiederholungszahl



Erweiterung: die True Score Logik der KTT

Um die True Score Logik zu verinnerlichen, wollen wir die einzelnen Konzepte in R umsetzen und mit den Kennwerten "herumspielen" um eine Intuition für das Konzept zu bekommen; v.a.:

1. Interindividuelle Merkmalsverteilung
2. Intrainduviduelle Merkmalsverteilung
3. Wahrer Wert, Beobachtungswert & Fehlerkomponente (τ, X, ϵ)

Zur Verortung: Kernkonzepte

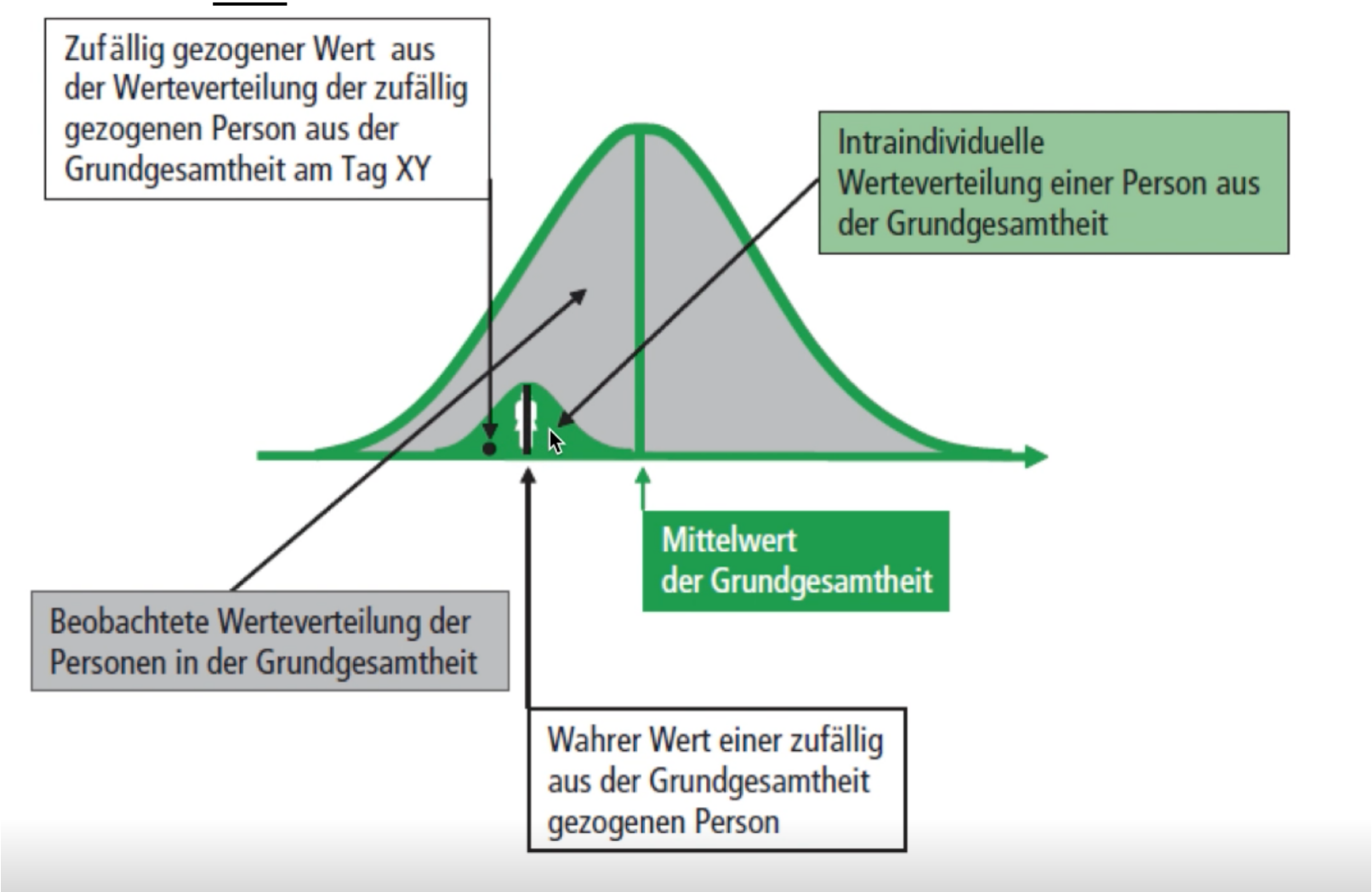


Figure 1: Merkmalsverteilungen

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.

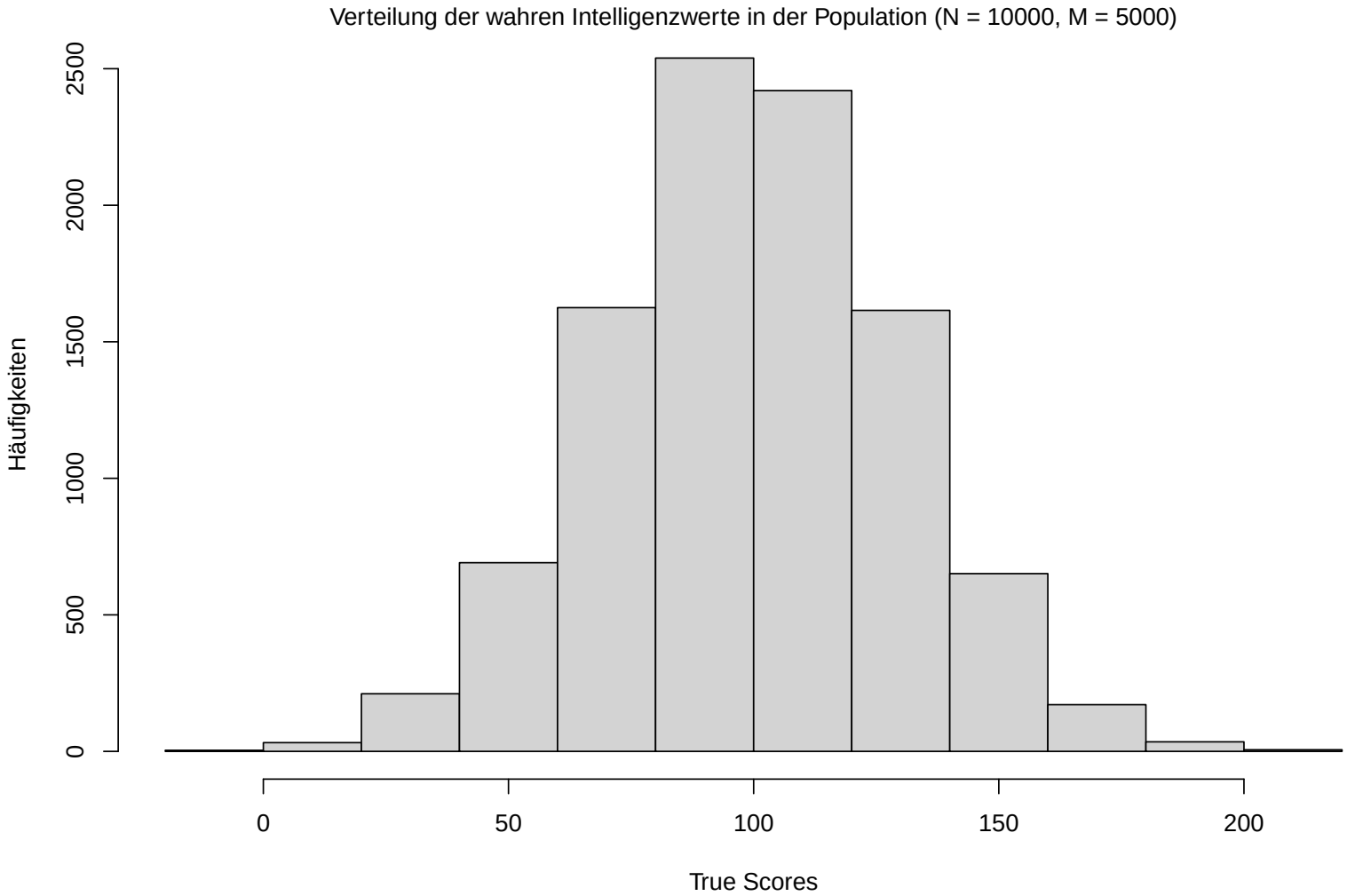
Populationen: Individuen & Messwerte

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

```
# Populationsgröße
N <- 10000
# Generierung der True Scores
# 100: Mittlere Intelligenz in der Population
# 30: Abweichungen vom Populationnsmittelwert
T <- round(rnorm(N, 100, 30), digits = 0)
# Anzahl der (imaginären) Testwiederholungen
M <- 5000
X <- lapply(T, function(T) rnorm(M, T , 5))
names(X) <- T
```

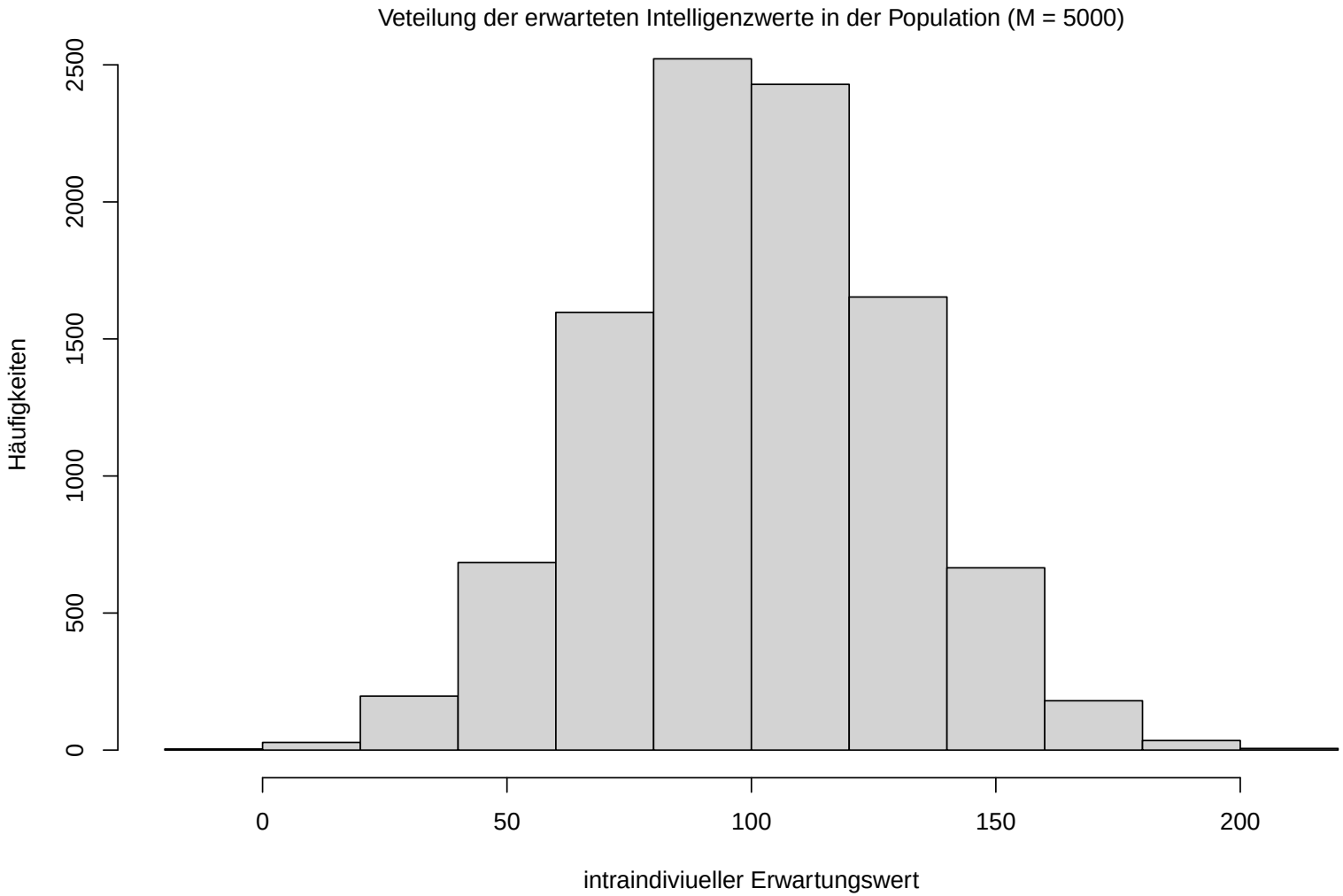
Verteilung der True Scores in der Population

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



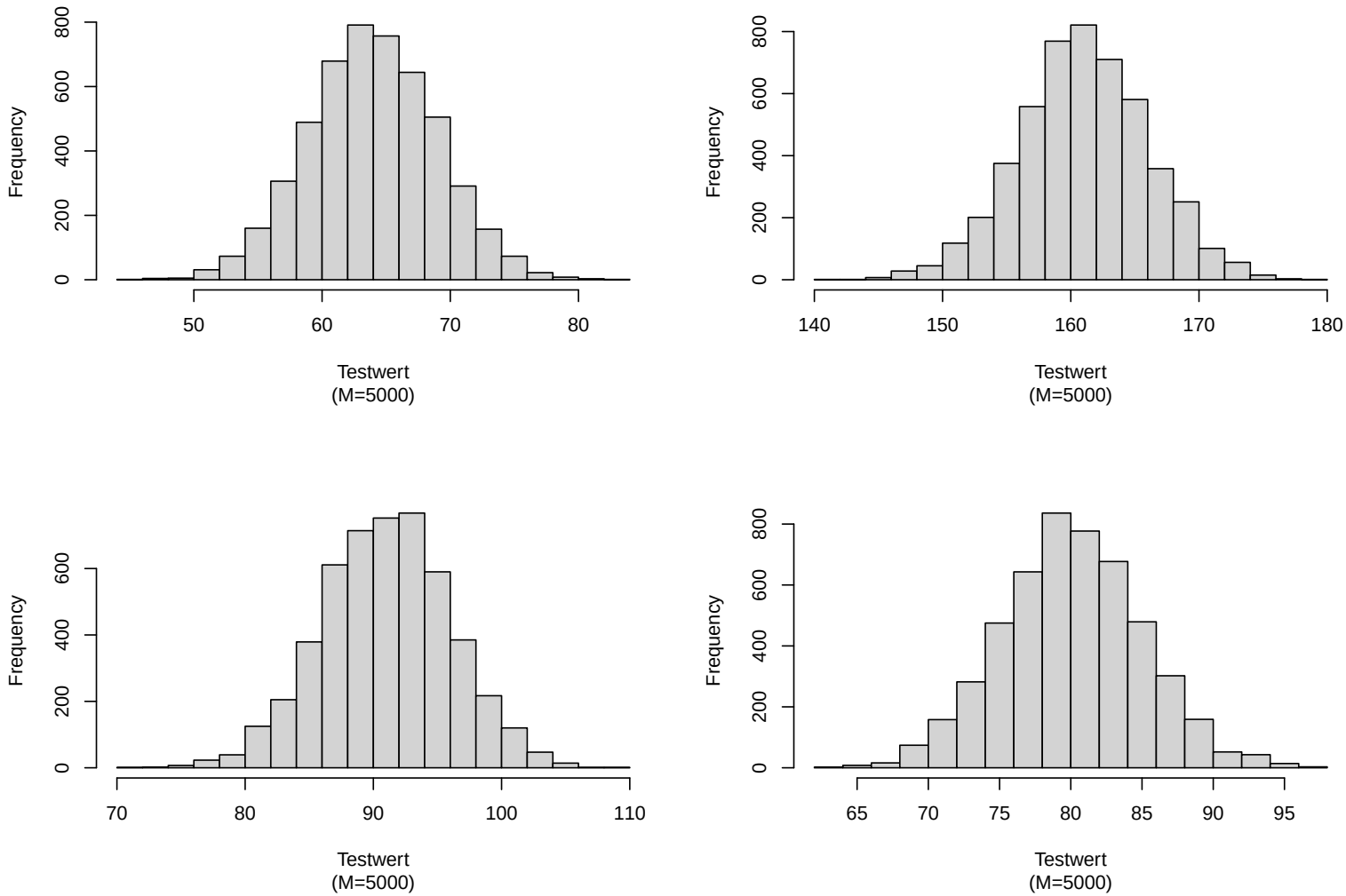
Interindividuelle Merkmalsverteilung

. . . mitteln wir die 5000 durchgeführten Testergebnisse ($X_{j=1,\ldots,5000}$) für jede Person ($i = 1, \ldots, 10000$) 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 einzes 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. Sehen Sie selbst...

```
# Zufallszug eines Individuums aus der Population
X_i <- sample(X, 1)
# Zufallszug eines Testwertes der Person
(X_ij <- sample(X_i[[1]], 1)) ; names(X_i)
```

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

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

Übungsaufgabe 3: Selbstexperiment

Example

Versuchen Sie es selbst! Nutzen Sie den Code zur Übungsaufgabe 3 in `04-KTT.R`. Ziehen Sie aus unserer Population immer wieder zufällig neue Probanden. Die nachfolgende Funktion `rsample()` erleichtert Ihnen den Prozess. Führen Sie diese deshalb immer wieder aus.

Wie verändern sich Wahrer Wert, Beobachtungswert und Messfehler?

- Zeit: 10 Minuten
- Replikation: `set.seed(123)`
- Anmerkung: Konzepte verstehen $\gg$ Codes verstehen!

Hilfsfunktion: rsample_i()

```
rsample_i <- function(){
  # Zufallszug eines Individuums
  X_i <- sample(X, 1)
  # Zufallszug eines Testwertes
  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))))}
# Automatisierung
rsample_i()
```

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

Übungsaufgabe 4: Selbstexperiment

Example

Versuchen Sie es selbst! Nutzen Sie den Code zur Übungsaufgabe 4 in `04-KTT.R`. Erstellen Sie ihre eigenen Populationen und variieren Sie systematisch die Werte M und N. (siehe: Übungsaufgabe 4)

1. Wie äußern sich Veränderungen in der Wiederholungszahl (M) ?& Populationsgröße (N)
2. Wie wirken sich die Veränderungen auf den Zusammenhang zwischen Wahrem Wert, Beobachtungswert und Messfehler aus (Tipp: $\tau = X + \epsilon$)

- Zeit: 10 Minuten
- Replikation: `set.seed(123)`
- Anmerkung: Konzepte verstehen $\gg$ Codes verstehen!

Zwischenfazit: Was wissen wir bisher?

Figure 2: Mentimeter-Umfrage 1

Eine (bisher) noch ungeklärte Frage

"Das zentrale Konzept der klassischen Testtheorie ist die Reliabilität, das ist die Zuverlässigkeit bzw. Genauigkeit[...], mit der ein Testwert den wahren Wert erfasst." (wikipedia.de)

Rückfrage: Reliabilität $\sim$ Messfehler

Jetzt haben wie so viel über Messfehler gesprochen. Wie hängen nun aber die zwei zentralen Konzepte der KTT (Messfehler & Reliabilität) zusammen?

Anmerkungen: Nehmen Sie dazu an der nachfolgenden Mentimenter-Umfrage teil!

Brainstorming: Reliabilität & Messfehler

Wie hängen Reliabilität und Messfehler zusammen?

Figure 3: Mentimenter-Umfrage 2

Section 4

Reliabilität

Was ist Reliabilität?

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

Definition: Reliabilität

$$Rel(Y_j) = \frac{Var(\tau_j)}{Var(Y_j)} = \frac{Var(\tau_j)}{Var(\tau_j) + Var(\epsilon_j)} \tag{6}$$

- $Rel \in [0, 1]$
- $Rel = 0$: ausschließlich Messfehler
- $Rel = 1$: gar kein Messfehler

Zusammenhang: Reliabilität & Messfehler

```r
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
```

Übungsaufgabe 5: Selbstexperiment

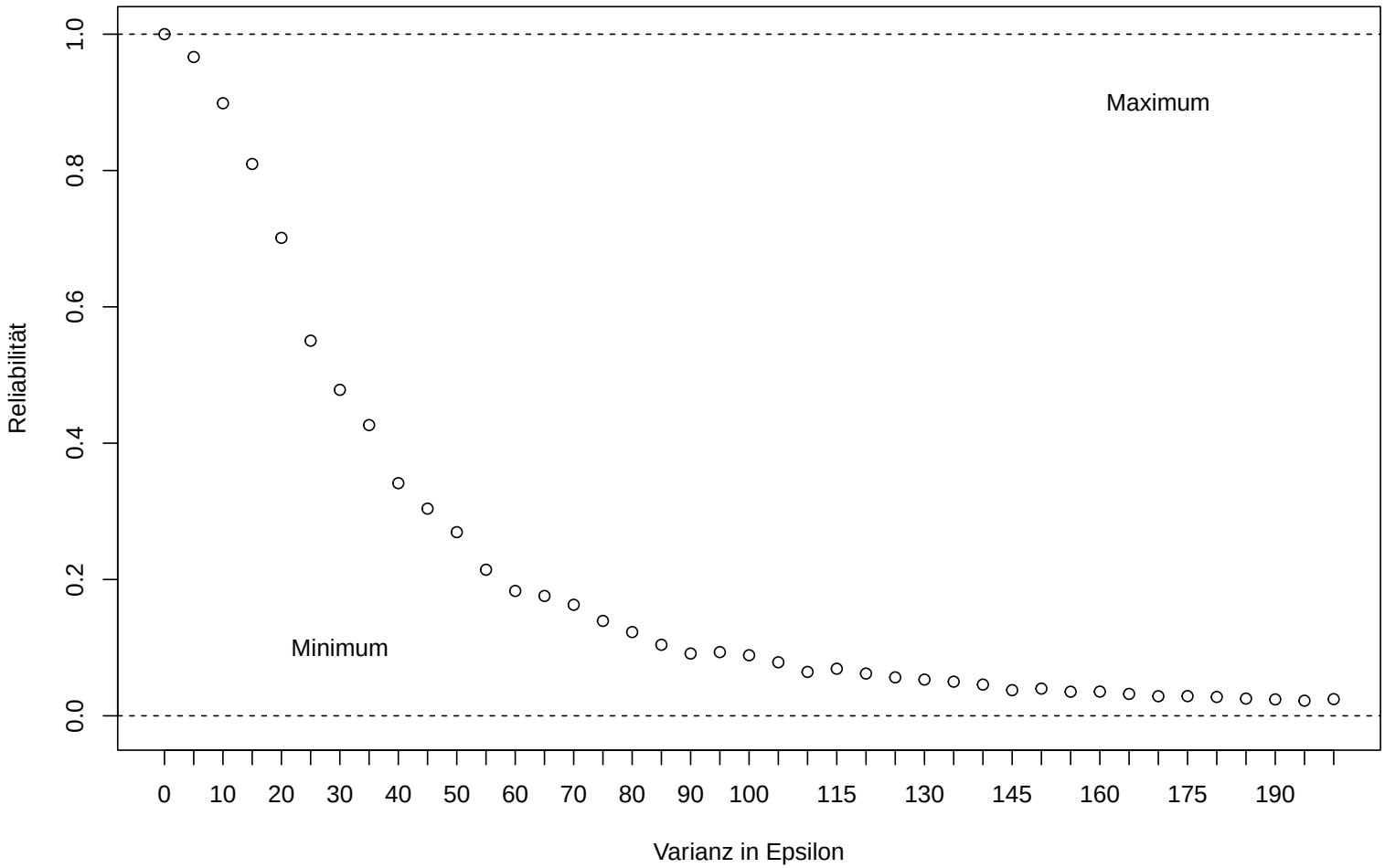
Example

Versuchen Sie es selbst! Nutzen Sie den Code zur Übungsaufgabe 5 in `04-KTT.R`. Verändern Sie die Varianz im Fehlerterm und schauen Sie auf die Veränderung in der Reliabilität der Testergebnisse.

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

- Zeit: 10 Minuten
- Replikation: `set.seed(123)`
- Anmerkung: Konzepte verstehen $\gg$ Codes verstehen!

Auflösung: Reliabilität & Messfehler



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

Zusammenfassung: KTT in a Nutshell

"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 True Score Logik

- 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.

$\Rightarrow$ Moderne Ansätze zur Verebsserung (..to be continued!)

Section 5

Bonusmaterial: Lokale Unabhängigkeit

Lokale Unabhängigkeit

Lokale Unabhängigkeit

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

- Die Logik lantenter Variablen (2.0)

$$\zeta : gen.process \rightarrow VAR(j = 1, \ldots, M) : cor(j, k) > 0) \tag{7}$$

- Die Logik lokaler Unabhängigkeit

$$cor(j, k)|\zeta = 0 \tag{8}$$

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

Grafik: Logik Lokaler Unabhängigkeit

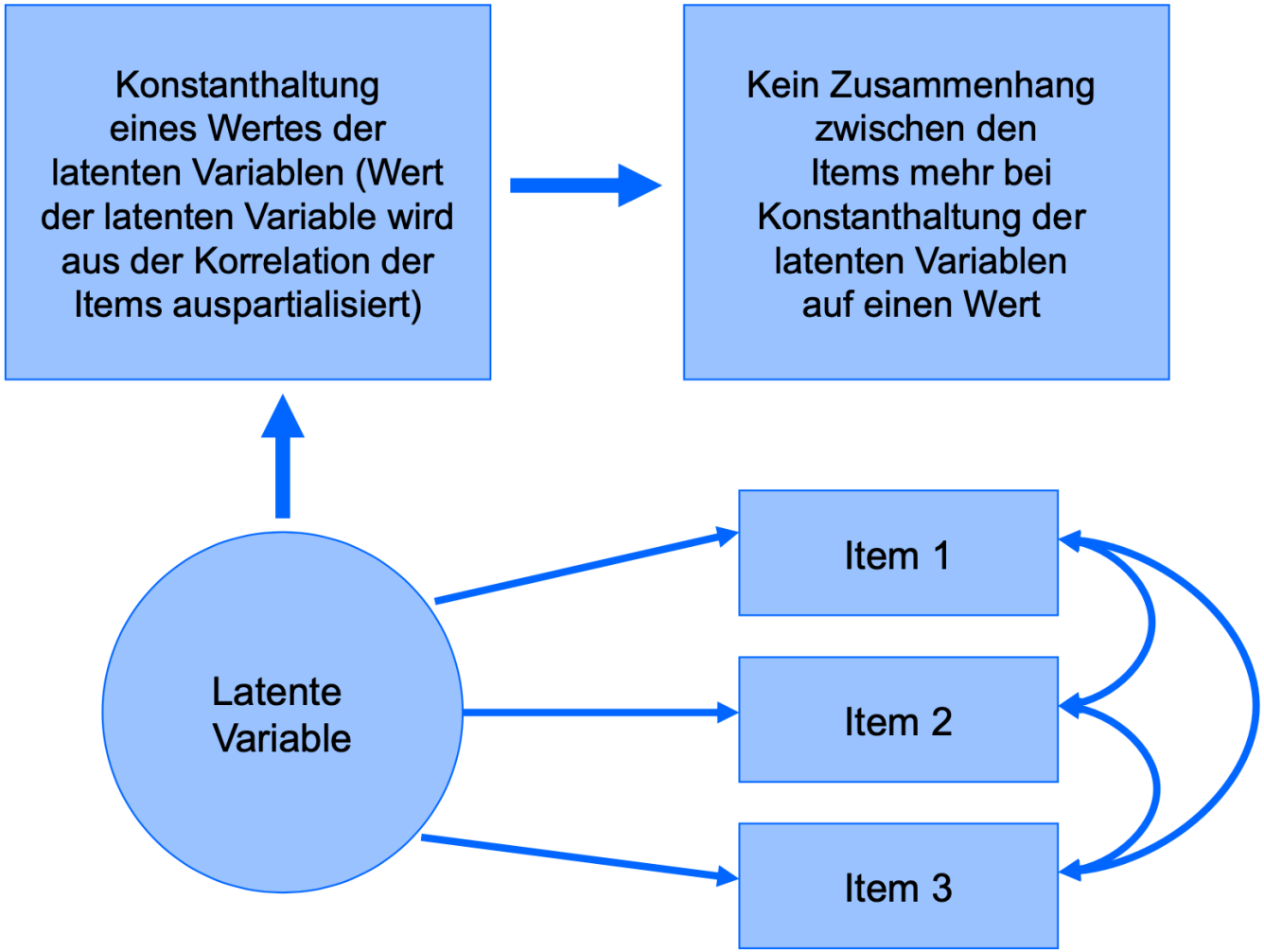
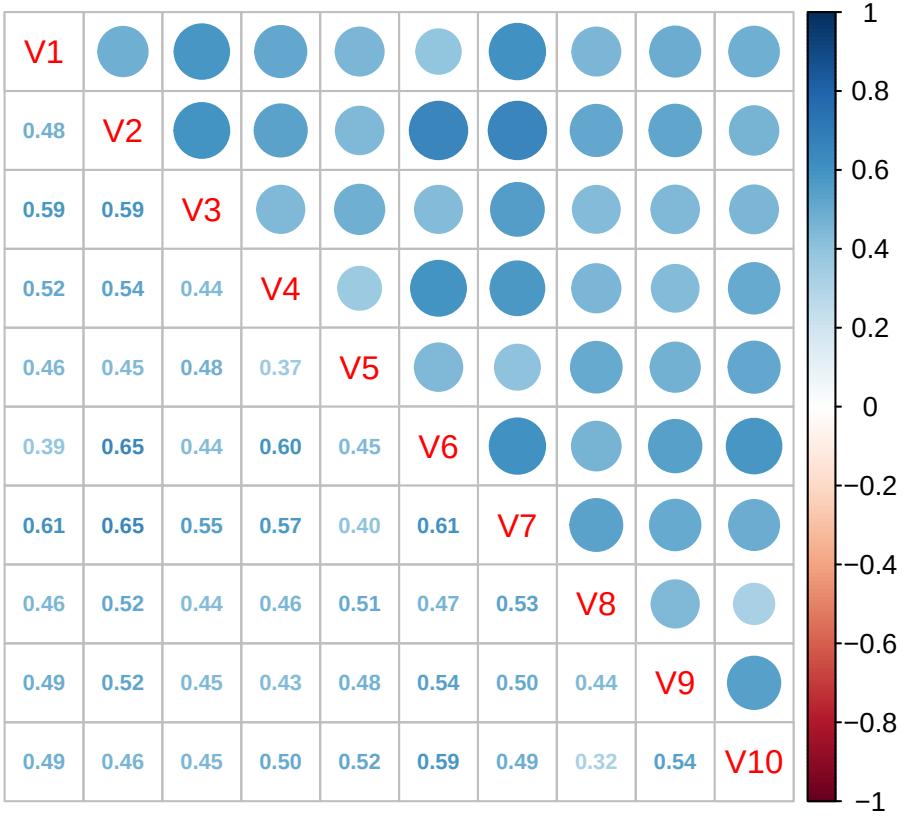


Figure 4: Logik Lokaler Unabhängigkeit

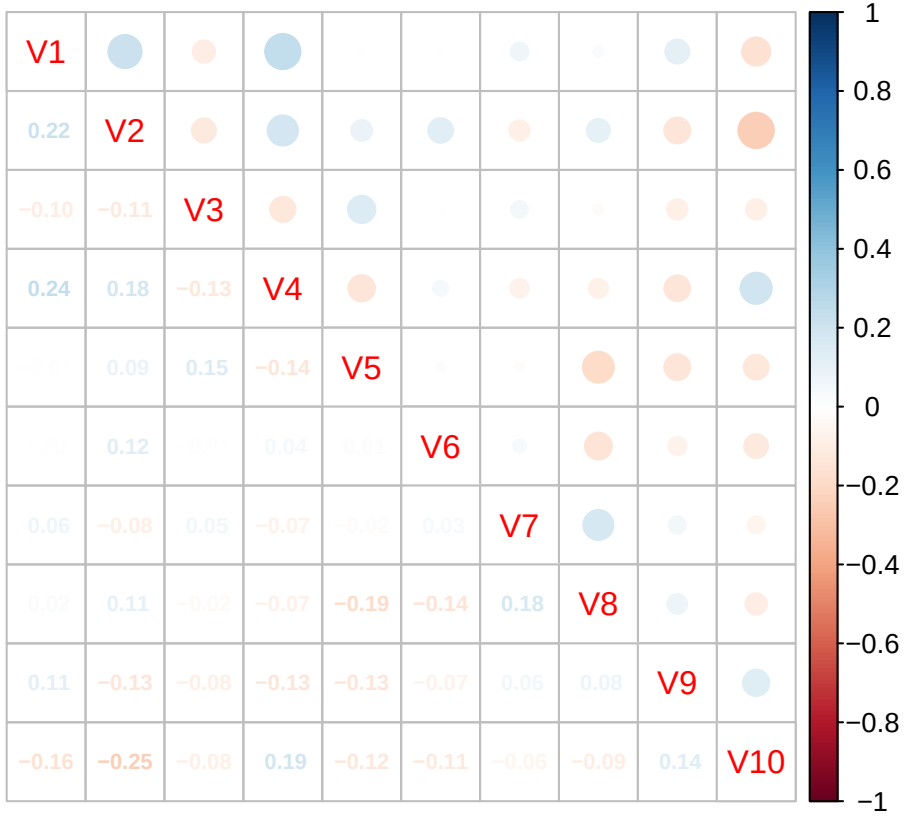
Grafik: Lokale Unabhängigkeit & Korrelationsmatrix

Wie überträgt sich dieses Prinzip auf die Itemantworten und damit auf die Korrelationsmatrix?

Vor Kontrolle der Traitausprägung



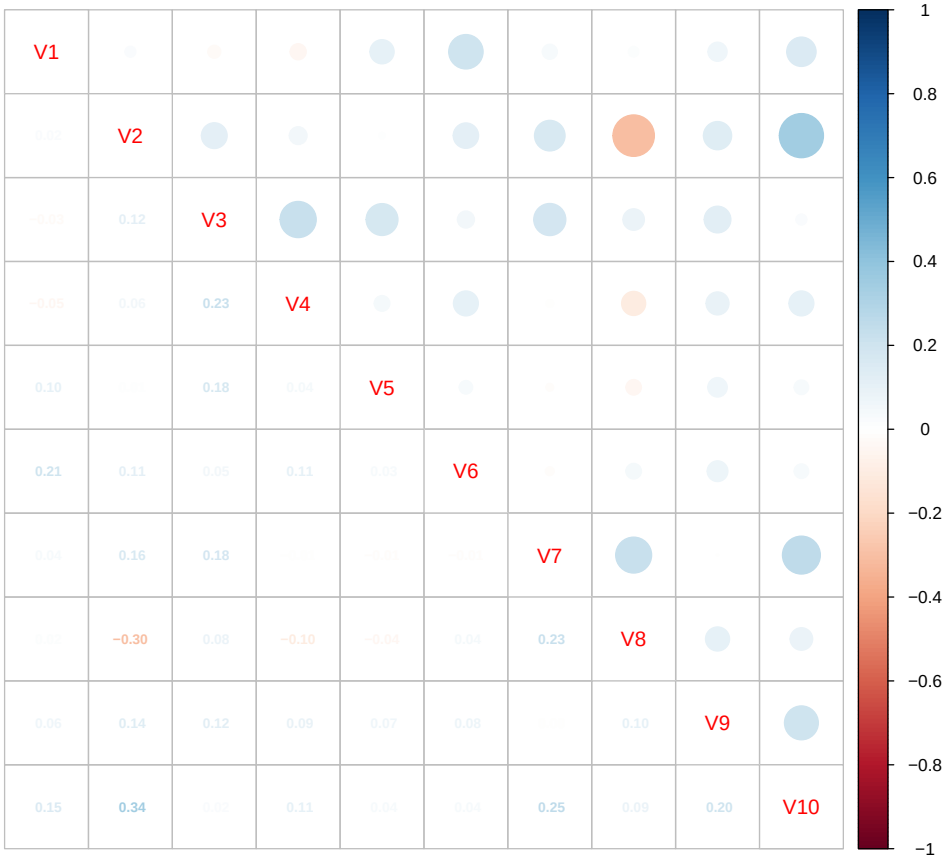
Nach Kontrolle der Traitausprägung



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 lokal statistischer Unabhängigkeit? Warum?



Section 6

Selbststudium

Zum Problem(?) systematischer Störeinflüsse

Ausganssituation

Da die KTT nur unsystamtische 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?

Auswkirungen 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
```

Auswikrungen 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
```

```
# Manipulatied predictor
beta_ast <- lm(y ~ I(x+3))$coef[[2]]
# Manipulatied outcome
lm(I(y+3) ~ I(x))$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)$coef[[2]]
# Manipulatied predictor
x[1:50] <- x[1:50] + 6
# Manipulatied coefficient
lm(y ~ x)$coef[[2]]
```

```
## [1] -0.03411985
```

Literaturverzeichnis I

Francois, Romain. 2020. *Bibtex: Bibtex Parser*. https://github.com/romainfrancois/bibtex.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Revelle, William. 2021. *Psych: Procedures for Psychological, Psychometric, and Personality Research*. https://personality-project.org/r/psych/%0Ahttps://personality-project.org/r/psych-manual.pdf.

Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook.