IV. Klassische Testtheorie

1. Allgemeines

- Älteste Messtheorie (Gulliksen, 1950)
- Messfehlertheorie / Reliabilitätstheorie
- Liefert die theoretische Begründung der Reliabilität
- Grundannahmen (Axiome) der KTT bilden die Basis zur mathematischen Ableitung der Reliabilität

Beispiel:
Intelligenzmessung von Person X.

	IQ
1. Messung	114
2. Messung	109
3. Messung	112
4. Messung	115
5. Messung	110
Mittelwert	**112**
Fehlerstreuung	2.28

Messfehler = Gesamtheit aller unsystematischen Einflussgrößen

- **Testkonstruktion** (z.B. mehrdeutige Items)
- **Testdurchführung** (z.B. variierende Testsituation, Motivation, Auftreten des Testleiters)
- **Testauswertung** (z.B. Fehler bei der Bestimmung des Testwertes)

Führt zu Schwankungen in den Ergebnissen

Die KTT setzt sich nur mit den unsystematischen Messfehlern auseinander. Reliabilitätstests sind so nur bezogen auf unsystematische Messfehler. Systematische Messfehler (wie Ja-Sage-Tendenz, Soziale Erwünschtheit) müssen separat über die Testkonstruktion bearbeitet werden.

$$X = T + E$$

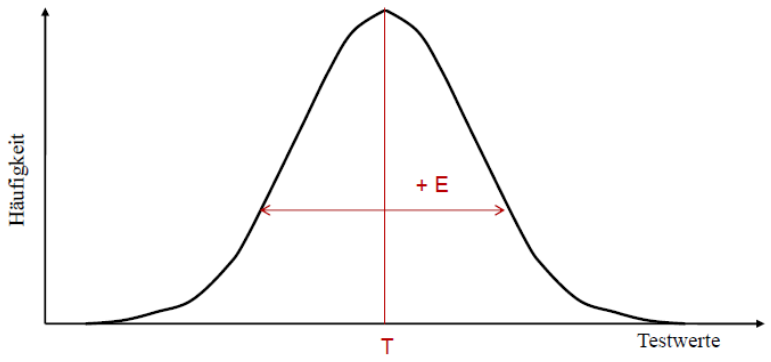
X = beobachteter Wert
T = wahrer Wert (true score)
E = Messfehler (error Score)

2. Axiome

1. Das Existenzaxiom

- Der Erwartungswert der Messung entspricht dem wahren Wert einer Person
- Erwartungswert = Mittelwert über Messwiederholungen

$$Erw(X) = T$$



Unsystematische Messfehler (E) gehen in beide Richtungen und mitteln sich aus, daher ist Erw(X)=T

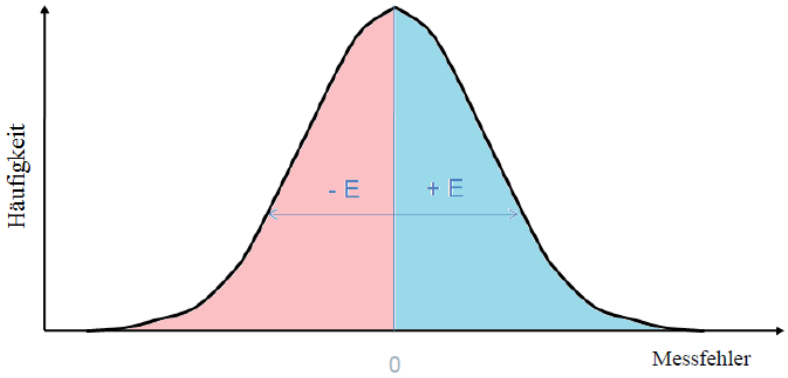
2. Das Verknüpfungsaxiom

- Der beobachteten Wert setzt sich aus dem wahren Wert und dem Messfehler zusammen

$$X = T + E$$

- Unter Berücksichtigung von Axiom 1 ergibt sich: Der Erwartungswert des Messfehlers ist Null

$$Erw(E) = 0$$



3. Das Unabhängigkeitsaxiom

- Die Korrelation zwischen den Messfehlern E und den wahren Werten T eines Tests sind unabhängig voneinander

$$Corr\ (T, E) = 0$$

- d.h. der Test misst im unteren Wertebereich genauso genau wie im mittleren oder oberen Wertebereich.

Weitere Annahmen:

- Die Messfehler zweier Tests (A und B) sind unabhängig voneinander

$$Corr\ (EA, EB) = 0$$

- Die Messfehler in einem Test A sind unabhängig von den wahren Werten in Test B

$$Corr\ (EA, TB) = 0$$

➔ *Messfehler sind unabhängig von allem, korrelieren also mit nichts.*

3. Ableitung der Reliabilität

(auf Basis der Axiome)

Was ist Reliabilität ?

= Die Reliabilität beschreibt die Genauigkeit, mit der ein Merkmal erfasst wird.
Die Reliabilität eines Tests ist der Anteil der Varianz der wahren Werte (T) an der Varianz der beobachteten Werte (X).
Reliabilität von 1 würde also bedeuten es gäbe keine Schwankungen.

$$\text{Rel} = \frac{\text{Var(T)}}{\text{Var(X)}}$$

Da Messfehler extra-Varianz produzieren, ist Var (X) immer größer als Var (T).
Je stärker sich Messfehler auswirken, desto größer ist Var(X), desto kleiner ist die Reliabilität
Var(X) ist gut erfassbar, Var(T) nur schwer erfassbar

Wie bekommt man die wahre Varianz (Var (T))?

- mehrfache Wiederholung der Messung pro Person
- Mittelwert = wahrer Wert (Erw(X) = T)
- Varianz der Mittelwerte = Varianz der wahren Werte

Nicht praxistauglich, da nicht immer mehrere Wiederholungsmessungen gemacht werden können.

Herleitung über zwei Messungen

- Test (t) wird unter identischen Bedingungen mit den gleichen Personen durchgeführt (t‘)
- In beiden Fällen gilt: **X = T + E** bzw. **X‘= T‘ + E‘**

$$Cov(X\,X') = Cov(T\,T') + Cov(T\,E') + Cov(T'\,E) + Cov(E\,E')$$

Die Kovarianz wird in additive Komponenten zerlegt, da sich X und X‘ aus additiven Komponenten bildet. Addition der Kovarianzen der additiven Komponenten = Kovarianz der beiden Testdurchführungen

Aus den Axiomen ergibt sich:

$$Cov(T\,E') = 0$$
$$Cov(T'\,E) = 0$$
$$Cov(E\,E') = 0$$

- *Die Korrelation zwischen den Messfehlern E und den wahren Werten T eines Tests sind unabhängig voneinander*
- *Die Messfehler zweier Tests (A und B) sind unabhängig voneinander*
- *Die Messfehler in einem Test A sind unabhängig von den wahren Werten in Test B*

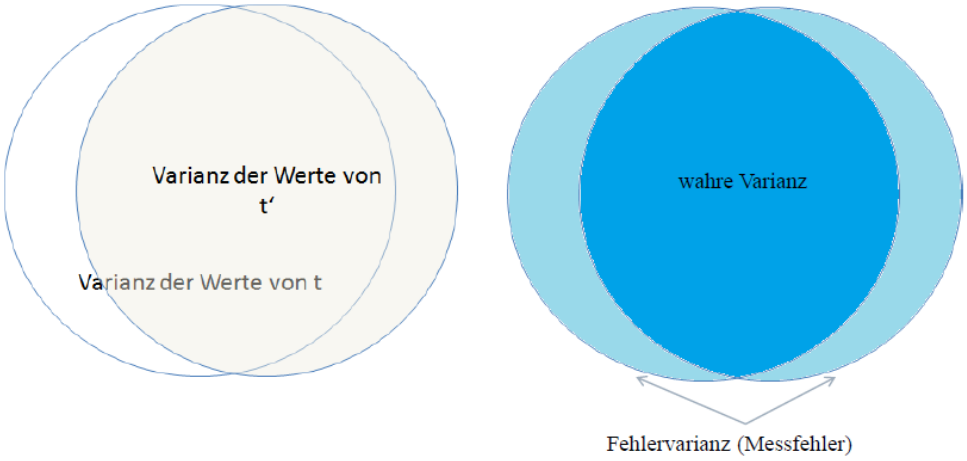
$$Cov(X\,X') = Cov(T\,T') + \cancel{Cov(T\,E') + Cov(T'\,E) + Cov(E\,E')}$$

$$\rightarrow \boldsymbol{Cov(X\,X') = Cov(TT') \Rightarrow Cov(X\,X') = Var(T)}$$

, da Cov und Var ähnliche Formeln sind.

Trennlinie: die Covarianz kann hier in die Varianzen von x und y aufgeteilt werden. Ist x = y wie bei der Messwiederholung, dann ist **Cov = Var**

$$\text{cov}_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$



Nur die TrueScores können kovariieren = Überlappung der beiden Tests

$$Rel = \frac{Var(T)}{Var(X)}$$

Cov(X X‘)

SD^2
→ SD * SD
→ SD(X) * SD(X‘)

Mehr Messungen integriert: Bessere Schätzung

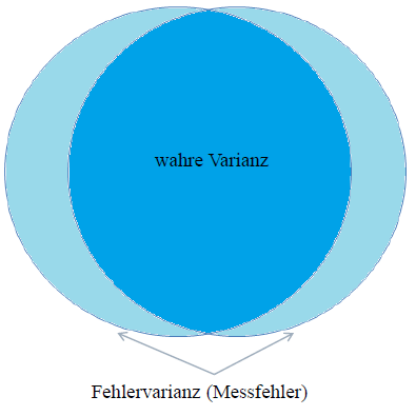
$$Rel = \frac{Cov\,(X\,X')}{SD(X)\cdot SD(X')} = Corr(X\,X')$$

= Retestreliabilität

$$r = \frac{\text{cov}_{xy}}{s_x \cdot s_y}$$

= **Retestreliabilität**
2 Messungen miteinander korrelieren, -> Ableitung aus KTT

4. Interpretation der Reliabilität



Rel = .80

→ 80 % der Varianz bezieht sich auf die Varianz wahrer Werte (zwischen Personen)

→ 20 % = Fehlervarianz (zwischen Personen)

5. Implikationen für die Praxis

(Reliabilitäten werden zu Interpretation von (absoluten) Testwerten herangezogen)

1. Standardmessfehler
2. Konfidenzintervall
3. Kritische Differenz
4. Minderungskorrektur

Standardmessfehler

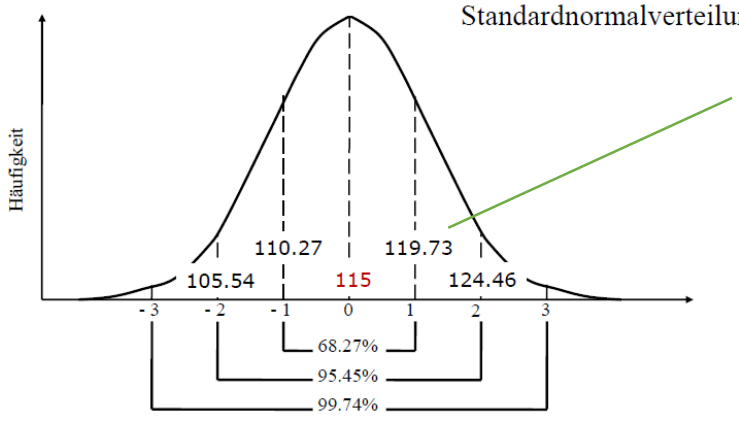
= gibt an, wie stark die Messfehler um die wahren Werte der Person(en) streuen.

$$s_E = s_X \cdot \sqrt{1 - Rel}$$

- s_E: Durchschnittliche Schwankung einer Person
- s_X: Beobachtete Streuung
- $1 - Rel$: Messfehleranteil

Multiplikation:
Sx* Wurzel(Messfehleranteil), damit Messfehleranteil auf das Maß der Streuung transformiert wird.

Beispiel:	Intelligenzmessung:	115
	Streuung der Intelligenzwerte (s_X):	15
	Reliabilität:	.90
	s_E:	4.73



Konfidenzintervalle: True score liegt mit einer Wahrscheinlichkeit von z.B.: 95% im Intervall xy

Der Standardmessfehler von S_E=4.73 besagt, dass (bei Normalverteilung) die beobachteten Werte in 68% (1 SD) der Fälle maximal 4.73 Punkte vom wahren Wert abweichen.

Im Beispiel: Der beobachtete Wert der Messung ist 115. Über die Formel gelangt man zu einem S_E=4.73. Das heißt, der wahre Wert liegt mit einer Wahrscheinlichkeit von 68% zwischen den Werten 110.27 (=115-4.73) und 119.73 (=115+4.73).

In der Praxis ist eine Sicherheits-wahrscheinlichkeit von 68% unüblich. Meist gibt man sich 90, 95 oder 99 % vor.

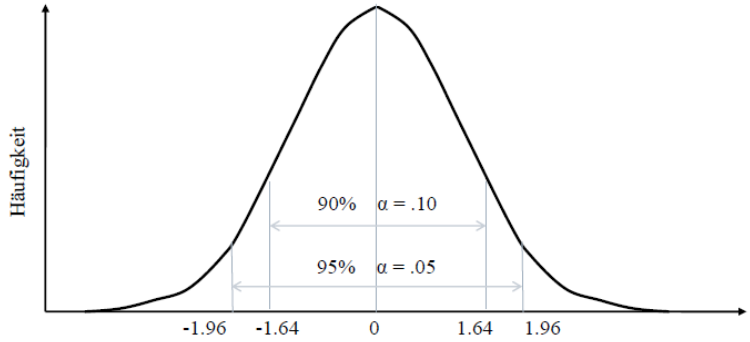
Konfidenzintervall

Das Konfidenzintervall (Vertrauensintervall) gibt den Bereich an, in dem der wahre Testwert einer Person bei einer zuvor festgelegten Irrtumswahrscheinlichkeit liegt.

$$CI = X \pm z_{\alpha/2} \cdot s_E$$

Der z-Wert $z_{a/2}$ bezieht sich auf die Standardnormalverteilung. Er gibt an, wie viele Standardabweichungen ein Wert vom Mittelwert der Verteilung entfernt liegen kann, damit noch x Prozent der Fläche unter der Verteilungskurve abgedeckt sind. Bei Standardnormalverteilung:

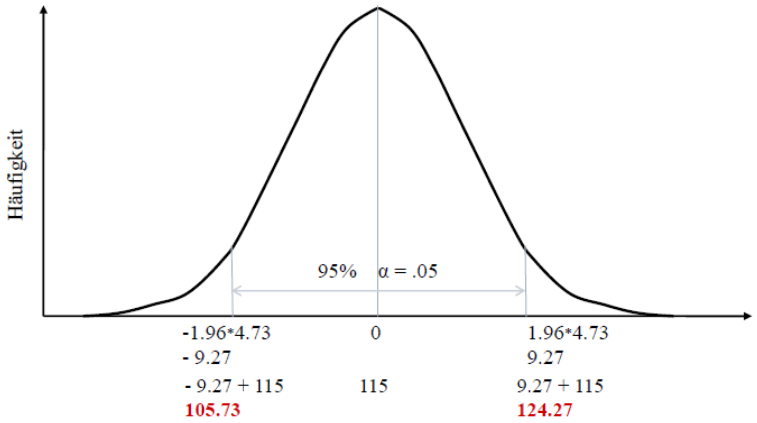
z= 1,96 -> die so begrenzte Fläche unter der Verteilungskurve umfasst 95% der Gesamtfläche
z=1,64 -> 90%
z=1 -> 68%



Durch Angeben der Sicherheitswahrscheinlichkeit, wird die Breite des Intervalls bestimmt und bestimmt, mit wie großer Wahrscheinlichkeit der True Score in dem Intervall liegt

CI = X ± 1.96 * 4.73
CI = X ± 9.27

X = 115



Bei Rel = .90,
Streuung=15,
X = 115,
wird SE = 4,73

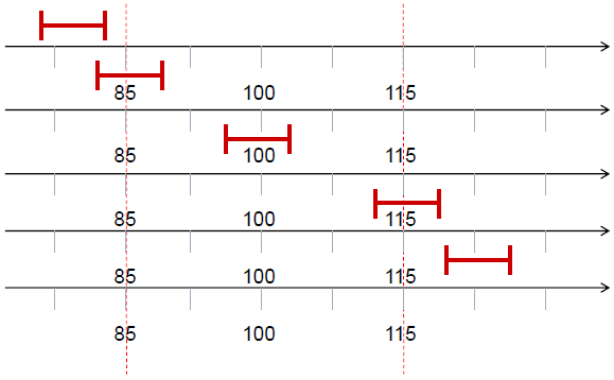
-> mit 95% Wahrscheinlichkeit liegt der IQ im Intervall von 105,73 - 124,27

Sicherheitswahrscheinlichkeit	z-Wert zweiseitig	z-Wert einseitig
99 Prozent	2,58	2,33
95 Prozent	1,96	1,64
90 Prozent	1,64	1,28

Zweiseitige vs. Einseitige Fragestellung:

Bei einer zweiseitigen Fragestellung interessieren wir uns gleichermaßen für Abweichungen vom beobachteten Wert nach oben und nach unten. Der beobachtete Wert liegt in der Mitte des Konfidenzintervalls. Bei bestimmten Fragestellungen interessiert man sich jedoch nur für Abweichungen nach oben oder nach unten; man spricht dann von einer einseitigen Fragestellung. Beispielsweise möchte man wissen, ob ein Kind mit einem IQ von 138 tatsächlich hochbegabt ist. Liegt sein wahrer Wert vielleicht unter der kritischen Grenze von 130? Ob sein IQ in Wahrheit noch höher sein kann als 138 ist in diesem Fall nicht von Interesse.

Praxisbeispiel Intelligenz (Klassifikation auf Basis des Konfidenzintervall)



Einbezug von CI bei Klassifikation:

Wenn Wert und Konfidenzintervall Bereiche überschneiden, dann kann man nicht klassifizieren.

Achtung: es werden nur unsystematische Messfehler berücksichtigt, systematische Messfehler sind nicht berücksichtigt.

Kritische Differenz

Wenn zwei Testwerte (z.B. von zwei Personen oder von einer Person in zwei Tests) verglichen werden sollen, stellt sich die Frage, ob der zweite Wert vom ersten signifikant abweicht. Beispielsweise will man wissen, beträgt der IQ einer Person im sprachlichen Bereich 105 und im rechnerischen 110. Ist die Person also eher rechnerisch begabt? Ein beobachteter Unterschied kann grundsätzlich auch auf den Messfehler zurückzuführen ist.

Die kritische Differenz beschreibt, wie große eine Differenz sein muss, um nicht mehr alleine mit Messfehlern erklärt werden zu können

$$D_{krit} = z_{\alpha/2} \cdot s_{E.Diff}$$

Standardmessfehler der Differenz

$$s_{E.Diff} = \sqrt{s_{E1}^2 + s_{E2}^2}$$

verbale Intelligenz (Reliabilität)

$$s_E = s_X \cdot \sqrt{1 - Rel}$$

numerische Intelligenz (Reliabilität)

$$s_E = s_X \cdot \sqrt{1 - Rel}$$

$$s_{E.Diff} = \sqrt{s_{E1}^2 + s_{E2}^2}$$

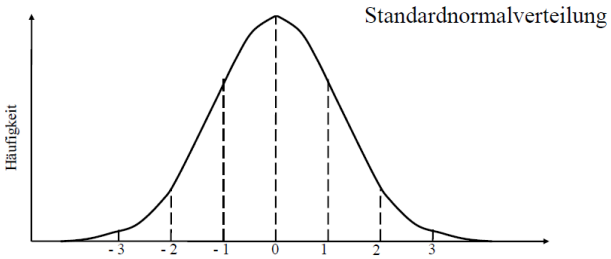
$$D_{krit} = z_{\alpha/2} \cdot s_{E.Diff}$$

Beide Messungen sind Messfehlerbehaftet, daher SE -> MW-Differenzen haben Standardmessfehler.

Schwankungen der MW-Differenzen sind von S_E1 und S_E2 abhängig. Je größer diese sind, desto größer ist S_E.Diff.

D_krit: Ab welcher Differenz kann man davon ausgehen, dass die Differenz signifikant ist / Wahrscheinlichkeit, dass Nullhypothese gilt ->

Je größer S_E.Diff, desto größer D_krit, (d.h. die Differenz, ab der man ausgehen kann, dass die Differenz signifikant ist, wird größer) -> schwieriger Signifikanz zu erreichen (erst bei größeren D_Krit's)



α – Niveau (α/2)	z-Wert (α/2)
5 % (2.5%)	1.96
1% (0.5%)	2,57

Praxisbeispiel Intelligenz (Kritische Differenzen)

- Bedeutsamkeit von Untertestdifferenzen

1. Bildung der kritischen Differenz
2. Vergleich der kritischen Differenz mit der beobachteten Differenz

Entscheidungsregeln

- $D_{kritisch} \geq D_{beobachtet}$ → Unterschied kann auf Messfehler der Messwerte zurückgeführt werden (diagnostisch nicht bedeutsamer Unterschied)
- $D_{kritisch} < D_{beobachtet}$ →Unterschied lässt sich nicht allein durch den Messfehler erklären (diagnostisch bedeutsamer Unterschied)

numerische Intelligenz:	130	SD = 15	Rel = .95	$s_{E1} = 3.35$
verbale Intelligenz:	115	SD = 15	Rel = .90	$s_{E2} = 4.73$

$$s_{E.Diff} = \sqrt{s_{E1}^2 + s_{E2}^2} = 5.80$$

$$D_{krit} = z_{\alpha/2} \cdot s_{E.Diff} = 1.96 * 5.80 = 11.36$$

Kritische Differenz = 11.36

Beobachtete Differenz = 130 – 115 = 15

Kritische Differenz ≮ Beobachtete Differenz → diagnostisch bedeutsamer Unterschied

Minderungskorrekturen

Messfehlerbehaftete Messwerte wirken sich mindernd auf die Korrelation mit einer anderen Variable aus. Angenommen zwei Tests, die das gleiche Merkmal erfassen sollen, sind so schlecht konstruiert, dass sie nur aus Messfehlern bestehen. Den Axiomen der KTT zufolge werden diese beiden Tests nicht miteinander korrelieren. Je niedriger die Reliabilität, also je größer der Messfehleranteil, desto geringer muss die Korrelation ausfallen. Diese Minderung der Korrelation durch die Messfehler wird durch die Minderungskorrektur behoben.

Doppelte Minderungskorrektur

= Reliabilität beider Variablen wird berücksichtigt.

Liefert eine Schätzung für die Korrelation der wahren Werte zweier Varibalen, wenn deren Reliabilitätskoeffizienten bekannt sind. Damit wird gleichsam die „Minderung“ korrigiert, welcher Korrelationskoeffizienten unterliegen, wenn die miteinander korrelierten Messwerte fehlerbehaftet sind.

$$r_{corr\ 1,2} = \frac{r_{1,2}}{\sqrt{Rel_1} * \sqrt{Rel_2}}$$

Korrelation zweier Variablen kann nicht größer ausfallen als die Wurzel aus dem Produkt der beiden Reliabilitätskoeffizienten dieser Variablen.

Einfache Minderungskorrektur

= Liefert eine Schätzung für die Korrelation eines Tests mit einem Kriterium unter der Annahme, dass der Test messfehlerfrei ist.

$$r_{corr\ c} = \frac{r_{tc}}{\sqrt{Rel_c}}$$

$$r_{tc} = Korrelation: Test - Kriterium$$

$$Rel_c = Reliabilität\ des\ Kriteriums$$

6. Grenzen der KTT

Vorzüge

- Sparsame Theorie, die mit wenigen Grundannahmen auskommt
- Praktische Ableitungen zur Testkonstruktion, Reliabilitäts-schätzung und darauf bezogene Minderungskorrektur, Konfidenzintervalle

Grenzen

- Die Axiome der KTT sind empirisch nicht überprüfbar und nicht durchweg plausibel (z.B. Unabhängigkeit des Messfehlers vom wahren Wert)
- Messfehler verteilen sich nicht immer zufällig um den wahren Wert
- Annahme eines invarianten wahren Werts
- Die Parameter der KTT sind populations-und stichprobenabhängig
- Substichproben können mit unterschiedlichen Reliabilitäten einhergehen
- Das Skalenniveau wird häufig missachtet