Grundlagen der Diagnostik

Lerneinheit 5

Beobachter- und Beurteilerübereinstimmung II

Wahr-oder-Falsch Aussagen zur letzten Sitzung

1. Eine symmetrische Kreuztabelle deutet auf gleiche Wahrnehmungsschwellen hin.

2. Unjustierte Maße bestrafen sowohl mangelnde Konsistenz als auch unterschiedliche Wahrnehmungsschwellen.

3. Yules Y ist ein Maß, das explizit Konsistenz bewertet und Unterschiede in Wahrnehmungsschwellen nicht berücksichtigt.

4. Die prozentuale Übereinstimmung ist gerade bei hoher zufälliger Übereinstimmung ein valides Maß.

5. Cohens κ ist von der Basisrate unabhängig.

Indizes für Übereinstimmung abhängig vom Skalenniveau

Bei nominalskalierten Daten:

- Prozentuale Übereinstimmung
- Cohens κ (kappa) und Scotts π (pi)
- Odds Ratio (auch: „Risikoverhältnis“) und Yules Y

Bei ordinal- und intervallskalierten Daten:

- Rangkorrelationen
- Intra-Klassen-Korrelation (Intra-Class-Correlation, ICC)

1. Überblick über Maße für ordinal- und intervallskalierte Daten:
   - „Normale“ Korrelationen
   - Intra-Klassen-Korrelationen (inkl. ihrer Varianzquellen)

2. Eigenschaften unterschiedlicher ICCs

3. Ausblick: ICC-Modelle (Shrout & Fleiss, 1979) & ICCs in R

Lernziele

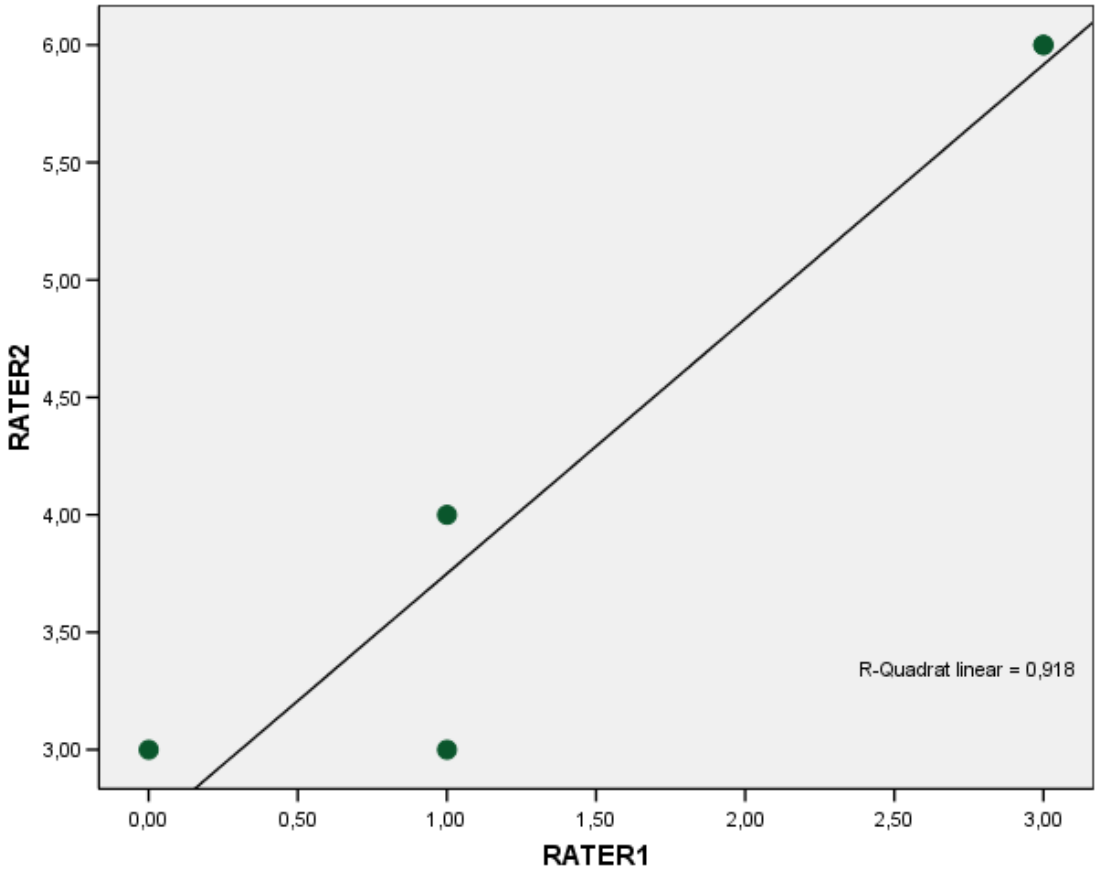
1. Überblick über Maße für ordinal- und intervallskalierte Daten

Möglichkeit 1: „Normale“ Korrelationskoeffizienten

- Die Korrelation zwischen beiden Ratern als Schätzung für die Übereinstimmung verwenden
- Die Pearson-Korrelation zwischen beiden Ratern beträgt r = .96

Aber:

- Nur 2 Rater
- Justiert: Nur Rangreihe



Rückblick: Unjustierte und Justierte Maße

Unjustierte Maße:

- Bestrafen unterschiedliche **Strenge** (Wahrnehmungsschwellen bei nominalskalierten Daten)
- Bestrafen mangelnde **Konsistenz**
- Konsequenz: **Absolute Übereinstimmung** wird bewertet

Justierte Maße:

- Bestrafen ***nicht*** unterschiedliche **Strenge**
- Bestrafen ***ausschließlich*** mangelnde **Konsistenz**
- Konsequenz: Nur Einhaltung der Rangreihe wird bewertet („**relative Übereinstimmung**“)

Konsistenz vs. Strenge bei Ordinal-/Intervallskalen

Beobachtung	Beurteiler A	Beurteiler B
#1	1	3
#2	1	3
#3	2	4
#4	2	4
#5	3	5
#6	3	5

→ konsistent, aber nicht gleich streng

Möglichkeit 2: Intra-Klassen-Korrelationen

Varianzanalytischer Ansatz:

- Erklärung der Unterschiede zwischen den realisierten Messwerten (z.B. Beurteilungen): $VAR(X_i)$
- Wiederholung:
  - Klassische Testtheorie (→ VL2 Testtheorie): $VAR(X_i) = VAR(\tau_i) + VAR(\varepsilon_i)$
  - Reliabilität (→ VL9 Testtheorie): $REL(X_r) = \frac{VAR(\tau_i)}{VAR(X_i)} = \frac{VAR(\tau_i)}{VAR(\tau_i)+VAR(\varepsilon_i)}$
- ICCs sind eine Schätzung der Reliabilität basierend auf (verschiedenen) Schätzungen für $VAR(\tau_i)$ und $VAR(\varepsilon_i)$
- Uneinigkeiten der Rater bei der Beurteilung werden in $VAR(\varepsilon_i)$ erfasst
- Die Varianz der „wahren Werte“ $VAR(\tau_i)$ wurde in Testtheorie auch als die „systematischen“ Unterschiede bezeichnet. Wir erinnern uns, diese ist nicht immer identisch ist mit den wahren Unterschieden der Personen im Merkmal: $VAR(\theta)$

Möglichkeit 2: Intra-Klassen-Korrelationen

Varianzanalytischer Ansatz zur Schätzung (i steht jetzt für Rater):

- Zusätzliche Möglichkeit der Berücksichtigung der Ratervarianz (→ zweifaktorielle Varianzanalyse):

$$VAR(X_i) = VAR(\theta) + VAR(rater_i) + VAR(\varepsilon_i)$$

- $VAR(\theta)$ ist die Varianz der wahren Unterschiede zwischen den Personen
- Werden die Raterunterschiede modelliert, sind sie ein Faktor in der Analyse, der Varianz in den Messwerten, also in $VAR(X_i)$, erklären kann (= eine weitere UV)
- Werden die Raterunterschiede nicht separat modelliert, ist $VAR(rater_i)$ in $VAR(\varepsilon_i)$ enthalten (systematische Raterunterschiede als Teil der „Fehler“).

Hinweis: Eine Interaktion zwischen Personen und Ratern wäre theoretisch denkbar, und damit auch eine Varianz $VAR(\tau*\text{rater})$ dieser Interaktion, aber diese wird typischerweise außen vor gelassen (d.h. als 0 betrachtet)

Möglichkeit 2: Intra-Klassen-Korrelationen

Varianzanalytischer Ansatz:

- Zusätzliche Möglichkeit der Berücksichtigung der Ratervarianz (→ zweifaktorielle Varianzanalyse):

$$VAR(X_i) = VAR(\theta) + VAR(rater_i) + VAR(\varepsilon_i)$$

- Wichtig: Es sind getrennte Entscheidungen, ob man …
  - … die Ratervarianz **modelliert** (d.h. im Modell aufnimmt, und dann durch Schätzung quantifiziert)
  - … die Ratervarianz **als Fehler berücksichtigt** (vgl. später ICC-Modell 2 und 3)

Höhe der ICC-Koeffizienten

Theoretischer Wertebereich: -1 bis 1; praktisch < 0 = 0

- 0 = Varianz ist ausschließlich auf Messfehler zurückzuführen (keine Reliabilität)
- 1 = Varianz ist ausschließlich auf wahre Werte zurückzuführen (perfekte Reliabilität)

Faustregeln zur Beurteilung (Fleiss, 1981; Cicchetti & Sparrow, 1981):

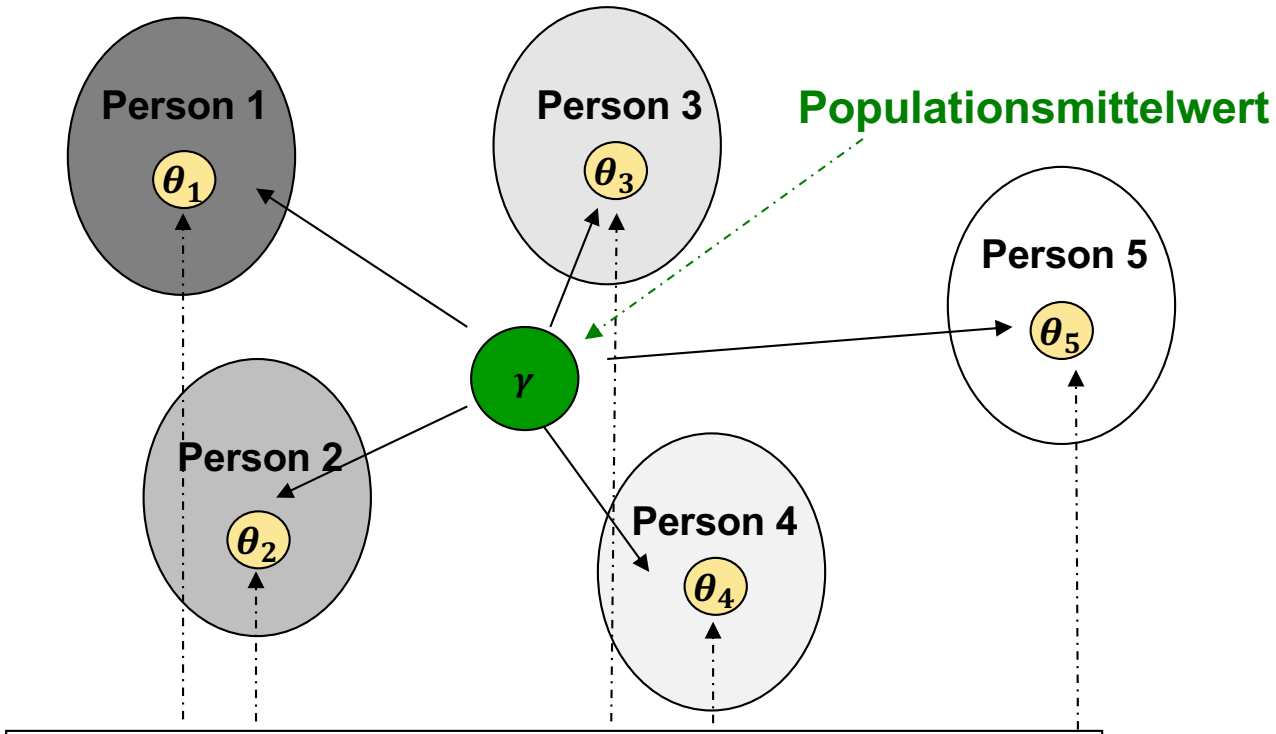
< 0.40	=	schlecht
0.40 – 0.59	=	befriedigend
0.60 – 0.74	=	gut
> 0.74	=	sehr gut

Bedeutung der verschiedenen Varianzquellen

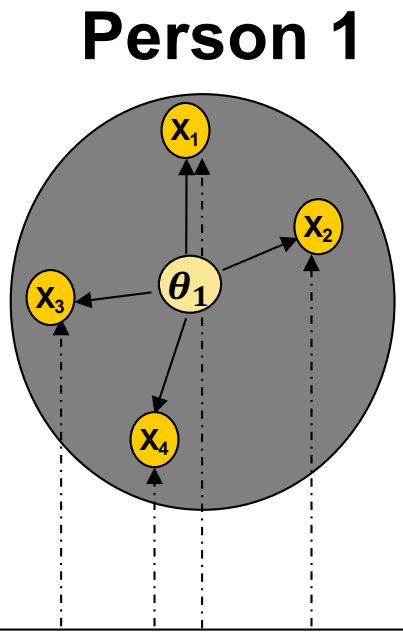
- **Faktor 1 = wahre Unterschiede**
  - Wahre Unterschiede zwischen den Personen = Schätzung der Varianz im Merkmal (bzw. der latenten Variable) → $VAR(\theta)$
- **Optionaler Faktor 2 = Raterunterschiede**
  - Unterschiede zwischen den mittleren Ratings der Rater über alle Personen hinweg = Schätzung unterschiedlicher Strenge (Wahrnehmungsschwellen) → $VAR(rater_i)$
- **Fehler**
  - Unterschiede der Rater in den Bewertungen für die Personen, die auf (sonstige) Fehler zurückgehen = Schätzung Fehlervarianz der Messwerte → $VAR(\varepsilon_i)$

Intra-Klassen-Korrelationen: Schätzung der verschiedenen Varianzquellen

große Kreise = „Raterraum“: Mögliche Raterurteile



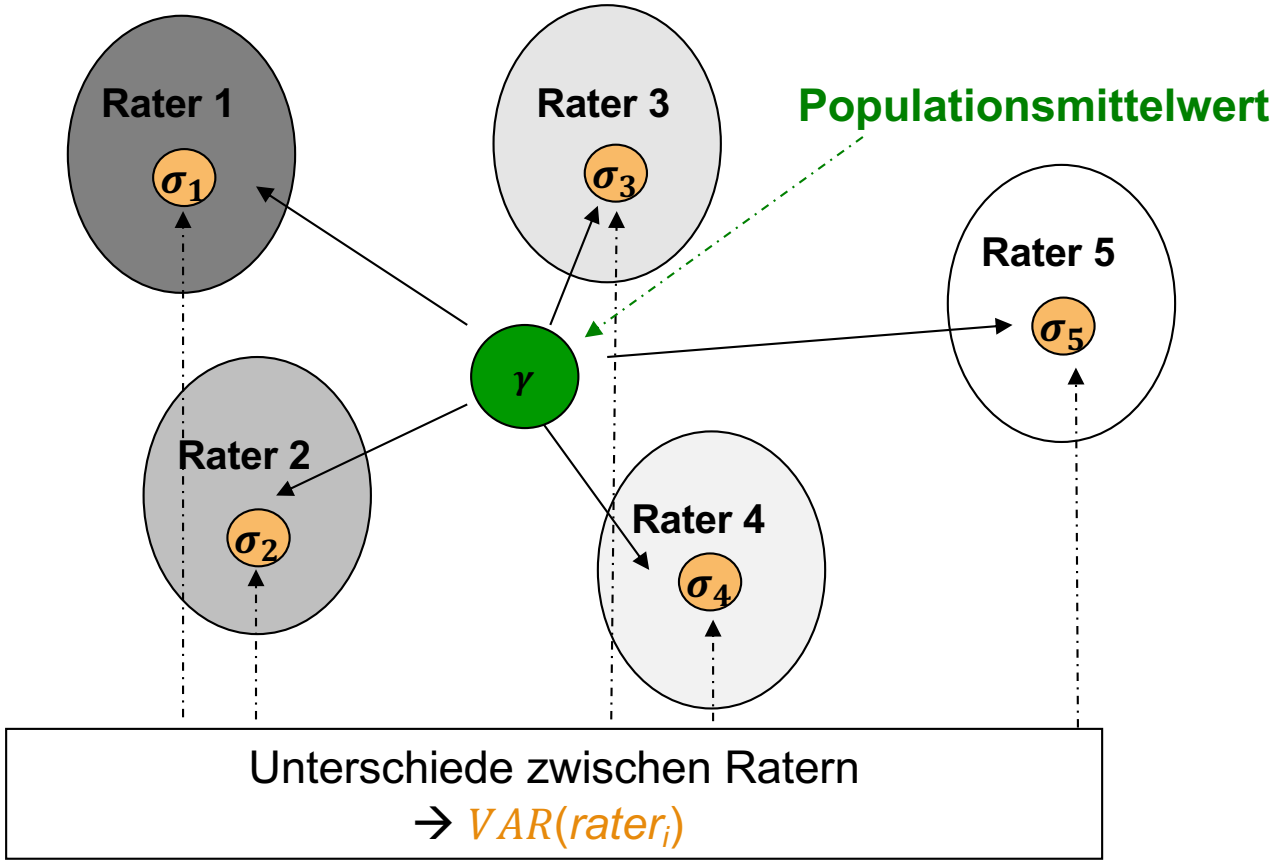
Tatsächliche Unterschiede zwischen Personen
→ $VAR(\theta)$



Abweichung der Ratings von den jeweiligen wahren Werten
→ $VAR(\varepsilon_i)$

- Ist $VAR(\theta)$ groß, dann unterscheiden sich die Personen bezüglich des Merkmals stark
- Ist $VAR(\varepsilon_i)$ groß, dann unterscheiden sich einzelne Ratings für die gleiche Person stark

Intra-Klassen-Korrelationen: Schätzung der verschiedenen Varianzquellen



- Ist $VAR(rater_i)$ groß, dann unterscheiden sich die Rater in ihrer mittleren Bewertung stark

- Die Varianz der wahren Unterschiede im Merkmal $VAR(\theta)$, Ratervarianz $VAR(rater_i)$ und Fehlervarianz $VAR(\varepsilon_i)$ sind nicht beobachtbar und müssen in einem statistischen Modell geschätzt werden.
- Die ursprüngliche Schätzmethode basiert auf varianzanalytischen Modellen (ein- und zweifaktorielle ANOVAs). Heute werden in der Regel sogenannte **Gemischte lineare Modelle** (LMMs) verwendet.
- Mithilfe dieser Modelle erhalten wir die Schätzwerte: $\widehat{VAR(\theta)}$, $\widehat{VAR(rater_i)}$ und $\widehat{VAR(\varepsilon_i)}$

2. Eigenschaften unterschiedlicher ICCs

Eigenschaften unterschiedlicher ICCs

Es gibt verschiedene ICC-Koeffizienten, mit unterschiedlicher Aussage und Interpretation (z.B. bzgl. Generalisierung, Art von Übereinstimmung, …).

Zwei Szenarien sind hier relevant:

- Wenn man die Situation der Datenerhebung **noch gestalten** kann:
  - Überlegungen zu welcher Aussage man kommen möchte und Situation entsprechend planen
- Wenn die Situation / die Daten **schon gegeben** sind:
  - Eingeschränkte Entscheidung, welche ICC möglich ist und Aussagekraft entsprechend berücksichtigen

Eigenschaften unterschiedlicher ICCs

→ Fragenkatalog, der zur richtigen ICC Variante führt:

- 1-faktorielle vs. 2-faktorielle ICC
- Random vs. Fixed ICC
- Unjustierte vs. Justierte ICC
- Single vs. Average ICC

1-faktorielle vs. 2-faktorielle ICC

Wird die Varianz der Rater explizit modelliert oder nicht?

- Werden die Rater und deren Varianz $VAR(rater_i)$ als weiterer Faktor in das Modell aufgenommen?
  - Ja: ICC 2-faktoriell; Modell: $VAR(X_i) = VAR(\theta) + VAR(rater_i) + VAR(\varepsilon_i)$
  - Nein: ICC 1-faktoriell; Modell: $VAR(X_i) = VAR(\tau_i) + VAR(\varepsilon_i)$
- Eine 2-faktorielle ICC kann nur berechnet werden, wenn es eine Überschneidung zwischen Personen und Ratern gibt. Im Idealfall wird jede Person von jedem Rater beurteilt.

Hinweis: $VAR(\varepsilon_i)$ ist in den beiden Modellen numerisch nicht gleich, man könnte hier auch einen zusätzlichen „Varianznamen“ vergeben. $VAR(\varepsilon_i)$ im einfaktoriellen Modell entspricht $VAR(rater_i)$ + $VAR(\varepsilon_i)$ im zweifaktoriellen Modell.

Random vs. Fixed ICC

Möchte man die Ergebnisse hinsichtlich der Ratervarianz auf eine Population von Ratern generalisieren oder sind die untersuchten Rater die Einzigen von Interesse?

- ICC random: Rater sind eine „repräsentative“ Stichprobe aus einer Population von Ratern → Generalisierung möglich
- ICC fixed: Es gibt nur diese Rater → Aussage nur für diese Rater möglich
- Da sich diese Unterscheidung auf die Ratervarianz bezieht, unterscheiden sich nur 2-faktorielle ICCs in der Eigenschaft random / fixed

Hinweis: Bei 1-faktoriellen ICCs fällt manchmal der Begriff „random“ - dies bezieht sich dann allerdings nicht auf die Ratervarianz (da diese nicht geschätzt wird; auch wenn man hier trotzdem von zufällig gezogenen Ratern ausgeht), sondern auf die bewerteten Personen, die in allen Modellen als „random“ (d.h. als Zufallsstichprobe) behandelt werden

Unjustierte vs. Justierte ICC

Interessiert man sich für die absolute Übereinstimmung oder nur die Rangreihe der Urteile?

- ICC unjustiert: absolute Übereinstimmung wird betrachtet
  → Beobachterinnen müssen konsistent (gleiche Rangreihe) *und* gleich streng urteilen (Übereinstimmung der absoluten Bewertung)
- ICC justiert: nur relative Übereinstimmung wird betrachtet
  → Es reicht, wenn Beobachterinnen die Personen in die gleiche Rangreihe bringen (d.h. konsistent sind)
- Dieser Unterschied kann als *Entscheidung* nur bei 2-faktoriellen ICCs einfließen: Eine 1-faktorielle ICC ist immer unjustiert

Single vs. Average ICC

Interessiert man sich für die Reliabilität eines Urteils von nur einem Rater oder für die Reliabilität eines gemittelten Urteils von mehreren Ratern?

- ICC single: Wie gut ist die Übereinstimmung eines Urteils mit dem von anderen Ratern?
- ICC average: Wie messgenau ist das gemittelte Urteil mehrerer Rater?
- Dies hängt davon ab, ob Entscheidungen letztendlich basierend auf Einzelurteilen oder gemittelten Urteilen stattfinden
- Für alle ICCs sind die Betrachtungen von „single“ und „average“ möglich

Analogie zu Testtheorie: *Rater entsprechen Items*

- Single: Das Merkmal der Person wird nur mithilfe eines Raters (*eines Items*) gemessen.
- Average: Der Mittelwert von mehreren Ratern (*Itemmittelwert*) dient als Messwert für die Merkmalsausprägung der Person.

3. Ausblick: ICC Modelle (Shrout & Fleiss, 1979) & ICCs in R

Ausblick: ICC Modelle (Shrout & Fleiss, 1979)

- Es gibt 3 verschiedene ICC-Modelle, die als Reliabilitätsschätzung herangezogen werden können
- Diese können jeweils als *single* und *average* ICC geschätzt werden, so dass sich nach Shrout & Fleiss (1979) insgesamt 6 ICC-Varianten mit unterschiedlichen Eigenschaften ergeben:

ICC Modelle (Shrout & Fleiss, 1979)	**ICC single**	**ICC average**
ICC Modell 1 (1-faktoriell, unjustiert)	ICC(1,1)	ICC(1,k)
ICC Modell 2 (2-faktoriell, unjustiert, random)	ICC(2,1)	ICC(2,k)
ICC Modell 3 (2-faktoriell, justiert, fixed)	ICC(3,1)	ICC(3,k)

k = Anzahl der Ratings aus denen der Mittelwert gebildet wird,
z.B. bei 3 Ratern und Modell 1: ICC(1,3)

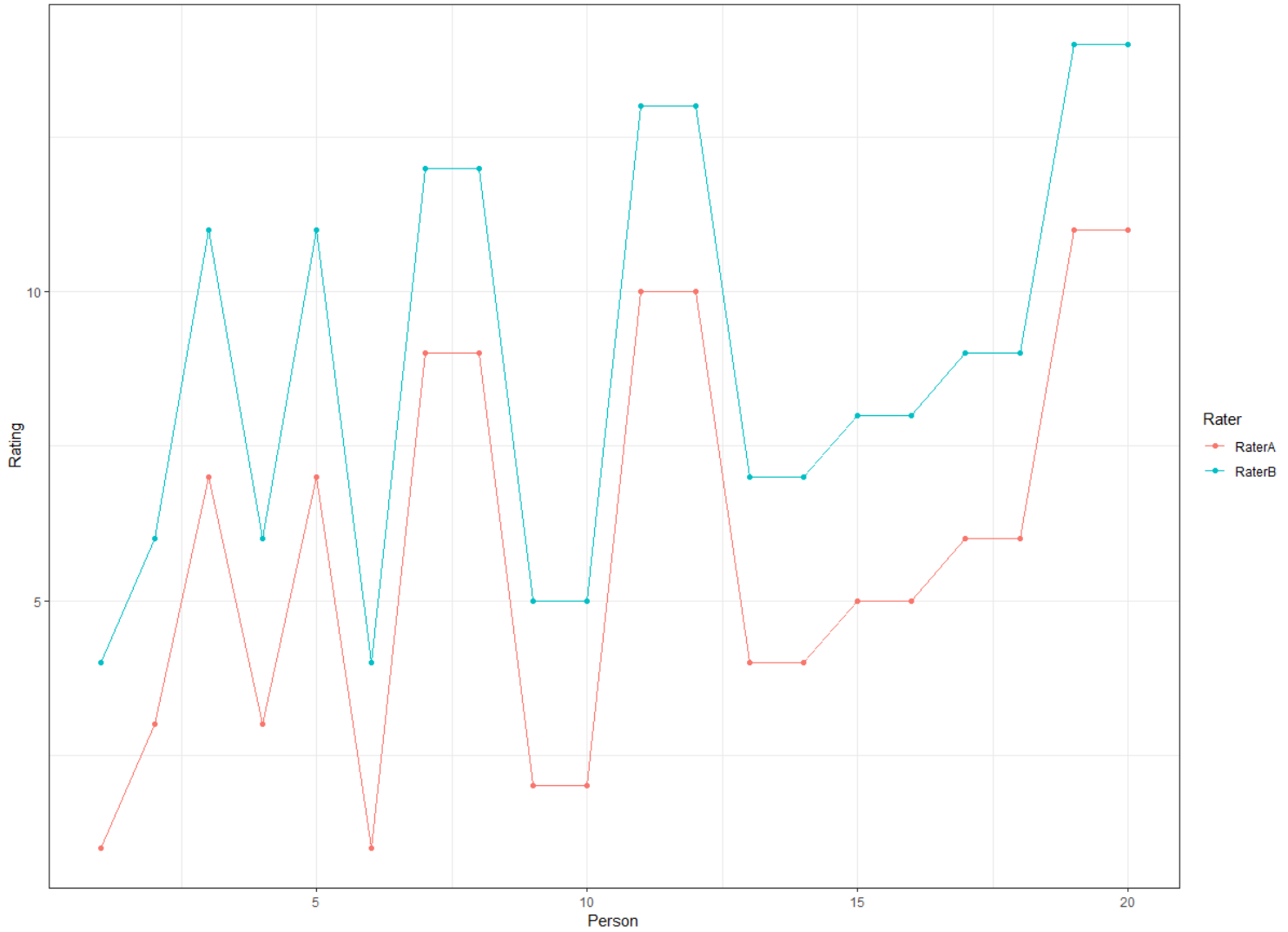
Exkurs: Erweiterungen der ICC-Modelle

- McGraw & Wong (1996) schlagen Varianten von den 2-faktoriellen Modellen 2 und 3 vor, so dass unjustiert/justiert beliebig mit random/fixed kombiniert werden kann
- Diese sind allerdings nur auf Ebene der Interpretation als Varianten zu betrachten, nicht auf Ebene der mathematischen Modelle
- Nur die von Shrout & Fleiss aufgestellten Modelle können mit den beschriebenen Eigenschaften (d.h. random für Modell 2 und fixed für Modell 3) als Korrelationen im engeren Sinne interpretiert werden. Diese Interpretation als Korrelation werden wir uns nächste Woche kurz anschauen.
- Wir werden uns im Folgenden nur mit den Modellen von Shrout & Fleiss beschäftigen.

ICCs in R → Bezug ICC zur Korrelation

data-Objekt

Person	RaterA	RaterB
1	1	4
2	3	6
3	7	11
4	3	6
5	7	11
6	1	4
7	9	12
8	9	12
9	2	5
10	2	5
11	10	13
12	10	13
13	4	7
14	4	7
15	5	8
16	5	8
17	6	9
18	6	9
19	11	14
20	11	14



ICCs in R

data-Objekt (anders sortiert):

Person	RaterA	RaterB
1	1	4
6	1	4
9	2	5
10	2	5
2	3	6
4	3	6
13	4	7
14	4	7
15	5	8
16	5	8
17	6	9
18	6	9
3	7	11
5	7	11
7	9	12
8	9	12
11	10	13
12	10	13
19	11	14
20	11	14

Code:

```
library(psych)
ICC(data[, c("RaterA", "RaterB")])
```

Output:

```
Call: ICC(x = data)

Intraclass correlation coefficients
                         type  ICC     F df1 df2       p lower bound upper bound
Single_raters_absolute   ICC1 0.65   4.7  19  20 6.2e-04      0.3040        0.84
Single_random_raters     ICC2 0.70 476.0  19  19 1.7e-21     -0.0022        0.93
Single_fixed_raters      ICC3 1.00 476.0  19  19 1.7e-21      0.9894        1.00
Average_raters_absolute ICC1k 0.78   4.7  19  20 6.2e-04      0.4663        0.91
Average_random_raters   ICC2k 0.82 476.0  19  19 1.7e-21     -0.0043        0.96
Average_fixed_raters    ICC3k 1.00 476.0  19  19 1.7e-21      0.9947        1.00

 Number of subjects = 20     Number of Judges =  2
```

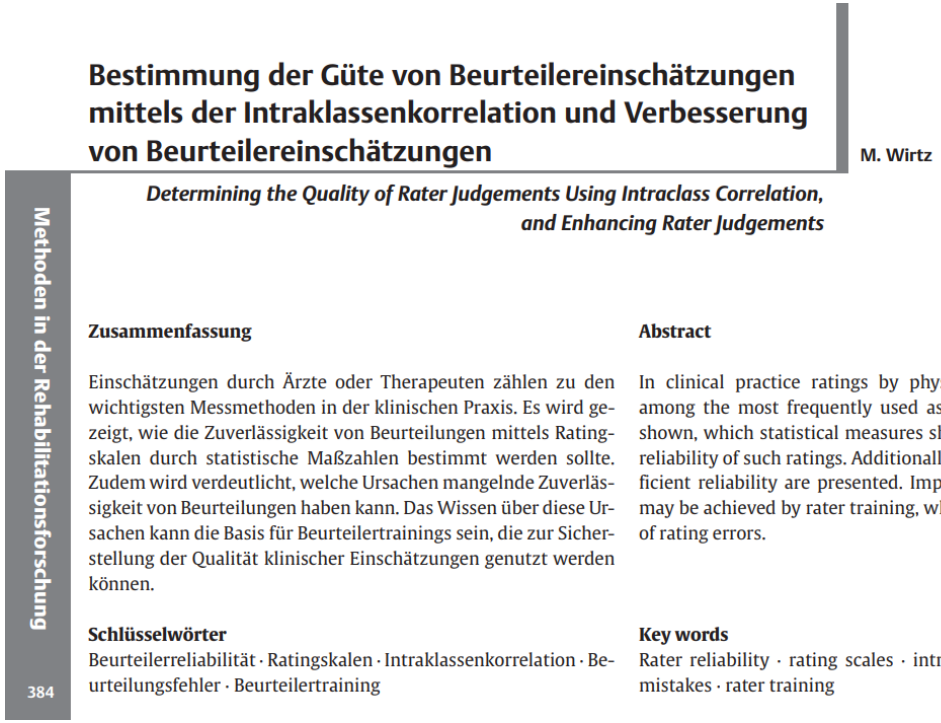
Konfidenzintervalle:

Zwischenfazit

- ICCs sind ein Maß für die Beobachter-/Beurteilerübereinstimmung bei **ordinal- oder intervallskalierten** Daten
- Für die Berechnung der ICC wird die beobachtete Varianz in den Bewertungen auf **verschiedene Varianzquellen** aufgeteilt
- Dadurch, dass die Varianz der wahren Werte auch geschätzt wird, können ICCs als **Schätzung für die Reliabilität** der Beurteilung herangezogen werden
- Es gibt **verschiedene ICCs**, die verschiedene Eigenschaften aufweisen (1-faktoriell vs. 2-faktoriell, unjustiert vs. justiert, random vs. fixed, single vs. average) und daher für verschiedene Situationen sinnvoll sind (dazu mehr im nächsten Foliensatz)

Beobachter- und Beurteilerübereinstimmung II

Wirtz, M. & Caspar, F. (2004). Beobachterübereinstimmung (ab S. 47) Kapitel 4.1.3, 4.1.4 und Kapitel 6. Weinheim: Juventa.



Bestimmung der Güte von Beurteilereinschätzungen mittels der Intraklassenkorrelation und Verbesserung von Beurteilereinschätzungen

M. Wirtz

Determining the Quality of Rater Judgements Using Intraclass Correlation, and Enhancing Rater Judgements

Methoden in der Rehabilitationsforschung

384

Zusammenfassung

Einschätzungen durch Ärzte oder Therapeuten zählen zu den wichtigsten Messmethoden in der klinischen Praxis. Es wird gezeigt, wie die Zuverlässigkeit von Beurteilungen mittels Ratingskalen durch statistische Maßzahlen bestimmt werden sollte. Zudem wird verdeutlicht, welche Ursachen mangelnde Zuverlässigkeit von Beurteilungen haben kann. Das Wissen über diese Ursachen kann die Basis für Beurteilertrainings sein, die zur Sicherstellung der Qualität klinischer Einschätzungen genutzt werden können.

Schlüsselwörter

Beurteilerreliabilität · Ratingskalen · Intraklassenkorrelation · Beurteilungsfehler · Beurteilertraining

Abstract

In clinical practice ratings by physicians and therapists are among the most frequently used assessment procedures. It is shown, which statistical measures should be used to assess the reliability of such ratings. Additionally, potential causes of insufficient reliability are presented. Improvement of rating quality may be achieved by rater training, which is based on an analysis of rating errors.

Key words

Rater reliability · rating scales · intraclass correlations · rating mistakes · rater training

Wirtz, M. (2004). Bestimmung der Güte von Beurteilereinschätzungen mittels der Intraklassenkorrelation und Verbesserung von Beurteilereinschätzungen. *Die Rehabilitation, 4*, 384-389.

Über die UB erreichbar (E-Medien-Login / PSYNDEX Datenbank)