Grundlagen der Diagnostik

Sitzung 4

Beobachter- und Beurteilerübereinstimmung I

Organisatorisches

Update!

Schriftliche Klausur, in Präsenz, auf Papier, Dauer 90 Minuten

- **Wann?** Mittwoch, 05. August 2026, 12:00-13:30 Uhr (s.t.)
- **Wo?**
  - Nachnamen A-H → Raum B001, Oettingerstr. 67
  - Nachnamen I-Q → Raum 2U01, Leopoldstr. 13
  - Nachnamen R-Z → Raum B138, Theresienstraße 39
- **Was?** Wahr-Falsch Aussagen (in in Diagnostik I bzw. Testtheorie)
  - Auch Aussagen zu R Outputs oder der Berechnung von Werten!
  - Mehr Details im Syllabus!

Sitzung	Datum	Thema	Themenblock
1	16.04.26	Einführung I	Definitionen; diagnostischer Prozess; gesetzlicher Rahmen; diagnostische Entscheidungen; Gütekriterien
2	23.04.26	Einführung II	
3	30.04.26	Einführung II fertig	
4	07.05.26	Verhaltensbeobachtung	Verhaltensbeobachtung als diagnostisches Verfahren
/	14.05.26	*entfällt wegen Feiertag*	
5	21.05.26	Beobachterübereinstimmung I	

➛ In der heutigen Vorlesung befassen wir uns mit Maßen der Beobachterübereinstimmung für dichotome Ratings, die uns helfen die Objektivität von aus Verhaltensbeobachtung und Interviews gewonnenen Daten zu beurteilen

Rückblick & Ausblick

Objektivität und Reliabilität bei der Verhaltensbeobachtung und anderen Beurteilungen (z.B. in Interviews)

Statistische Analysen zur Objektivität:

- Berechnung von Maßen der Beobachter- bzw. Beurteilerübereinstimmung

Statistische Analysen zur Reliabilität:

- Wie viel Fehler passiert beim Rückschluss von den Ratings eines oder mehrerer Beobachterinnen auf das wahre Merkmal (latente Variable)?
- Beobachterinnen können sich objektiv einig sein und trotzdem ist die Messung unreliabel (z.B. Beurteilungsfehler aus Sitzung 3).
- Bestimmte Maße der Beurteilerübereinstimmung (z.B. ICCs in Sitzung 5) können auch als Reliabilitätsschätzung verwendet werden, aber die Trennung von Übereinstimmung und Reliabilität ist nicht immer einfach.

Verschiedene Indizes für Beobachterübereinstimmung, abhängig vom Skalenniveau…

Bei nominalskalierten Daten:

- Prozentuale Übereinstimmung
- Cohens κ (kappa) und Scotts π (pi)
- Odds Ratio und Yules Y

Bei ordinal- und intervallskalierten Daten:

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

Hinweise:

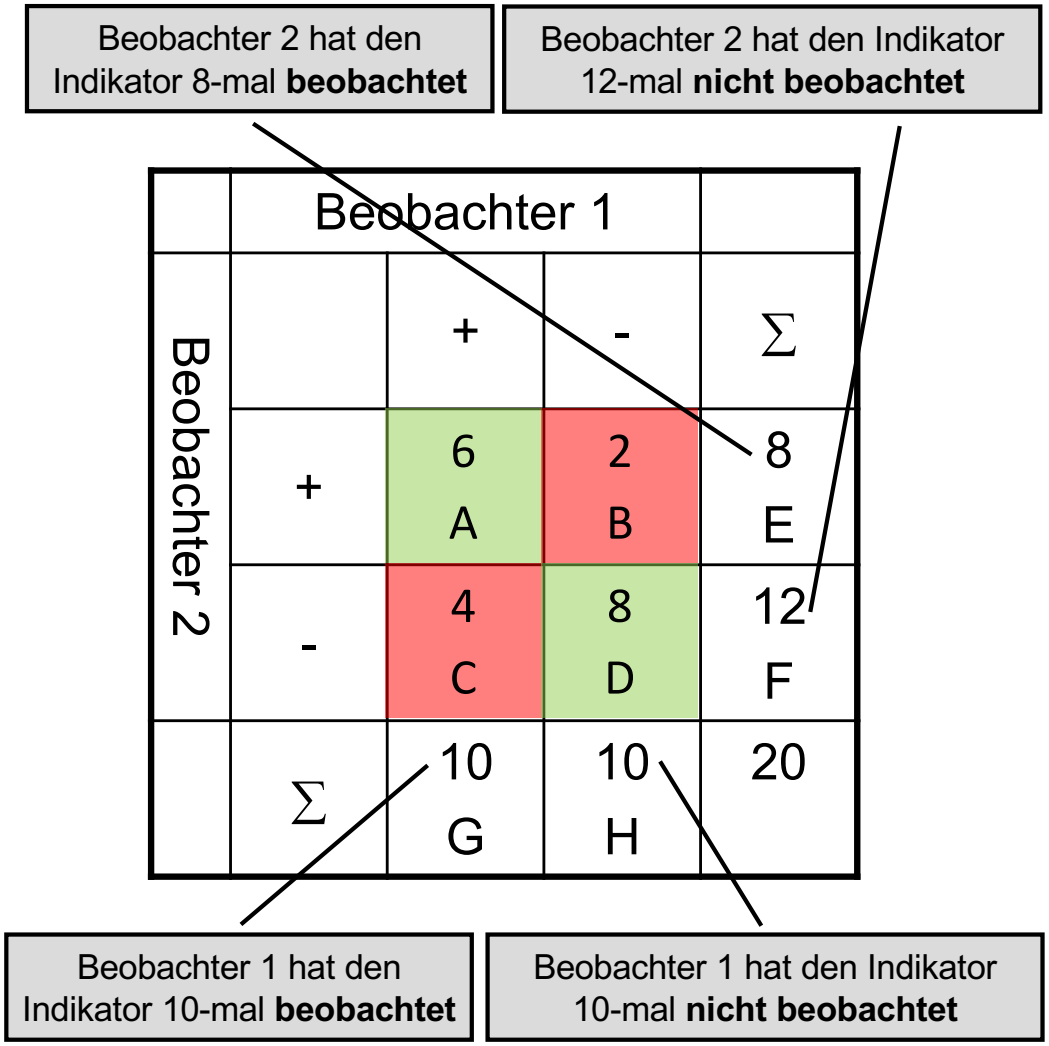
- Wir besprechen den Fall für zwei Beobachter, aber es gibt auch Varianten für mehr als zwei.
- Für ordinalskalierte Daten gibt es eine Abwandlung von Cohens κ, das „gewichtete Cohens κ“ und andere Koeffizienten (z.B. „Krippendorffs Alpha“), aber auf diese wird in der Vorlesung nicht eingegangen, da hierfür die Rangkorrelationen & ICCs besprochen werden

1. Aufbereitung der Daten für dichotome Ratings

Maße für **nominalskalierte** Daten

Beispiel:

- Assessment-Center mit 20 Bewerberinnen
- Zwei Beobachterinnen (sog. Rater)
  → Beobachter 1 & Beobachter 2
- Mögliche Urteile:
  - Indikator **beobachtet**, d.h., Bewerberin geeignet: **+**
  - Indikator **nicht beobachtet**, d.h., Bewerberin nicht geeignet: **-**



Wie kommen wir von einem Time-Sampling-Protokoll zu einer Kreuztabelle?

Beobachter 1	**Zeitintervall (1 bis 20)**																				Σ
Indikator:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
„ähm“ gesagt	x	x	x	x	x	x			x	x	x	x									10

In diesen Intervallen haben **beide** Beobachter den Indikator **beobachtet**

In diesen Intervallen hat **ein Beobachter den Indikator beobachtet, der andere nicht**

In diesen Intervallen haben **beide** Beobachter den Indikator **nicht beobachtet**

G

Beobachter 2	**Zeitintervall (1 bis 20)**																				Σ
Indikator:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
„ähm“ gesagt	x	x	x	x	x	x	x	x													8

A **B** **C** **D** **E**

		Beobachter 1		
		+	-	Σ
Beobachter 2	+	6 A	2 B	8 E
	-	4 C	8 D	12 F
	Σ	10 G	10 H	20

- Wenn möglich, direkt die Kreuztabelle interpretieren, um die Qualität der Übereinstimmung zu beurteilen!

  - Wie sehr stimmen die Beobachter überein?
  - Wenn die Beobachter nicht übereinstimmen, welches Muster ergibt sich für die „Fehler“?
  - Was lässt sich über die „wahre Verteilung“ der zugrundeliegenden Merkmalsausprägungen vermuten?
  - Beobachten die Beobachter die Merkmalsausprägungen unterschiedlich häufig, bzw. Unterscheiden sich die Beobachter in ihrer **Strenge**? → **Unterschiedliche Wahrnehmungs-schwellen?**
  - Sind sich die Beobachter hinsichtlich der **Rangreihe** einig (trotz unterschiedlicher Strenge), also welche Beobachtungen tendenziell „positiv“ beurteilt werden? → **Konsistente Beurteilung?**

	Beobachter 1		
Beobachter 2	+	-	Σ
+	6 A	2 B	8 E
-	4 C	8 D	12 F
Σ	10 G	10 H	20 N

- In der Praxis wünscht man sich häufig eine Quantifizierung der Qualität der Beobachterübereinstimmung über die deskriptive Betrachtung der Kreuztabelle hinaus
- Für eine solche Quantifizierung existieren zwei Strategien:

Modellbasierte Übereinstimmung:

- Interpretation von Parametern in speziellen Messmodellen zur Beschreibung des Beobachtungs- bzw. Beurteilungsprozesses
- Methodisch anspruchsvoll (Schätzung statistischer Modelle)
- *Können wir leider nur kurz ansprechen (siehe Anhang)*

Klassische Übereinstimmungsmaße und Indizes:

- Maßzahlen für Übereinstimmung basierend auf der Kreuztabelle
- Zusammenfassung der Übereinstimmung in einer einzigen Zahl
- Unterschiedliche Interpretation verschiedener Maßzahlen
- *Standard in der Praxis und Fokus dieser Vorlesung*

2. Berechnung der klassischen Übereinstimmungsmaße für dichotome Ratings

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**“)

Maße für nominalskalierte Daten:

1. Prozentuale Übereinstimmung

2. Cohens κ und Scotts π

3. Odd‘s Ratio und Yules Y

Lernziele

2.1. Prozentuale Übereinstimmung

1. Prozentuale Übereinstimmung

Berechnung:

$$P_O = \frac{\text{Häufigkeit der Übereinstimmung}}{\text{Anzahl aller Urteile}} \cdot 100 = \frac{\text{A+D}}{\text{A+B+C+D}} \cdot 100$$

Beispiel:

$$P_O = \frac{6+8}{6+2+4+8} = 70\%$$

		Beobachter 1		
Beobachter 2		+	-	$\sum$
	+	6 A	2 B	8 E
	-	4 C	8 D	12 F
	$\sum$	10 G	10 H	20

1. Prozentuale Übereinstimmung

Aber Achtung!

Prozentuale Übereinstimmung kann irreführend sein, wenn **Merkmale sehr häufig bzw. sehr selten beobachtet werden**,

z.B. sehr häufige / seltene Krankheiten (hohe / geringe *Basisrate*, siehe Sitzung 8)

	Beobachter 1		
Beobachter 2	+	-	Σ
+	17 A	1 B	18 E
-	1 C	1 D	2 F
Σ	18 G	2 H	20

Beispiel:

$$P_o = \frac{17 + 1}{17 + 1 + 1 + 1} = 90\%$$

2.2. Cohens κ und Scotts π

2. Cohens κ und Scotts π

Cohens κ und Scotts π beachten zusätzlich die **„durch Zufall“ erwartete Übereinstimmung**:

$$\kappa, \pi = \frac{Beobachtete\ Übereinstimunng(P_O) - Erwartete\ Übereinstimmung\ (P_E)}{1 - Erwartete\ Übereinstimmung\ (P_E)}$$

- Wertebereich: -1 bis +1 (in der Praxis meist > 0)
- *Zähler:* „Wie viel mehr Übereinstimmung liegt vor als zufällig erwartbar?“
- *Nenner:* „Wie viel mehr Übereinstimmung wäre maximal möglich?“
  (d.h. der größtmögliche Zähler)

2. Cohens κ und Scotts π

Die **„durch Zufall“ erwartete Übereinstimmung (P_E)** wird für Cohens κ und Scotts π und unterschiedlich berechnet

- Cohens κ: Schätzung der erwarteten Übereinstimmung aus den **kombinierten** Randsummen
- Scotts π: Schätzung der erwarteten Übereinstimmung aus den **mittleren** Randsummen

2. Cohens κ und Scotts π

Schritt 1:

Umwandlung der absoluten Häufigkeiten in relative Häufigkeiten

		Beobachter 1		
		+	-	Σ
Beobachter 2	+	6 A	2 B	8 E
	-	4 C	8 D	12 F
	Σ	10 G	10 H	20

→

		Beobachter 1		
		+	-	Σ
Beobachter 2	+	.30 A	.10 B	.40 E
	-	.20 C	.40 D	.60 F
	Σ	.50 G	.50 H	1

Im Beispiel:

A: 6 / 20 = .30
B: 2 / 20 = .10
C: 4 / 20 = .20
D: 8 / 20 = .40

E: 8 / 20 = .40
F: 12 / 20 = .60
G: 10 / 20 = .50
H: 10 / 20 = .50

2. Cohens κ und Scotts π

Schritt 2:

Die „durch Zufall“ erwartete Übereinstimmung berechnen (P_E)

Wie hoch ist die Wahrscheinlichkeit, dass die Beobachter zufällig übereinstimmen, also zufällig beide „positiv“ oder beide „negativ“ beurteilen?

		Beobachter 1		
		+	-	Σ
Beobachter 2	+	.30 A	.10 B	.40 E
	-	.20 C	.40 D	.60 F
	Σ	.50 G	.50 H	20 N

≙ Wie hoch ist die Wahrscheinlichkeit, dass eines von zwei Ereignissen eintritt, die sich gegenseitig ausschließen?

→ **Addition der Wahrscheinlichkeiten**

2. Cohens κ und Scotts π

Schritt 2:

Die „durch Zufall“ erwartete Übereinstimmung berechnen (P_E)

Wie hoch ist die Wahrscheinlichkeit, dass die Beobachter zufällig beide „positiv“ beurteilen (analog für „negativ“)?

		Beobachter 1		
		+	-	Σ
Beobachter 2	+	.30 A	.10 B	.40 E
	-	.20 C	.40 D	.60 F
	Σ	.50 G	.50 H	20 N

≙ Wie hoch ist die Wahrscheinlichkeit, dass zwei unabhängige Ereignisse gleichzeitig eintreten, die jeweils eine bestimmte Wahrscheinlichkeit haben?

→ **Multiplikation der Wahrscheinlichkeiten**

2. Cohens κ und Scotts π

Schritt 2:

Die „durch Zufall“ erwartete Übereinstimmung berechnen (P_E)

- P_E für Cohens κ: $(E \cdot G)+(F \cdot H)$
- *Annahme:* Jeder Beobachter hat eine eigene Wahrscheinlichkeit, eine Ausprägung zu beobachten.
- „**kombinierte Randhäufigkeiten**“
- Im Beispiel:
  P_E für Cohens $\kappa = (.40 \cdot .50) + (.60 \cdot .50) = .50$

		Beobachter 1		
		+	-	Σ
Beobachter 2	+	.30 A	.10 B	.40 E
	-	.20 C	.40 D	.60 F
	Σ	.50 G	.50 H	20 N

2. Cohens κ und Scotts π

Schritt 2:

Die „durch Zufall“ erwartete Übereinstimmung berechnen (P_E)

- P_E für Scotts π: $\left(\frac{E+G}{2}\right)^2 + \left(\frac{F+H}{2}\right)^2$
- *Annahme:* Beide Beobachter haben die gleiche Wahrscheinlichkeit, eine Ausprägung zu beobachten. Damit ist der Mittelwert der beiden relativen Häufigkeiten ein sinnvoller Schätzwert für diese Wahrscheinlichkeit.
- „**mittlere Randhäufigkeiten**“
- Im Beispiel: P_E für Scotts π $= \left(\frac{.40+.50}{2}\right)^2 + \left(\frac{.60+.50}{2}\right)^2 = .505$

		Beobachter 1		
		+	-	Σ
Beobachter 2	+	.30 A	.10 B	.40 E
	-	.20 C	.40 D	.60 F
	Σ	.50 G	.50 H	20 N

2. Cohens κ und Scotts π

Schritt 2:

Die „durch Zufall“ erwartete Übereinstimmung berechnen (P_E)

Zusammenfassung:

- P_E für Cohens κ: $(E \cdot G) + (F \cdot H)$
  - D.h. kombinierte Randhäufigkeiten
- P_E für Scotts π: $\left(\frac{E+G}{2}\right)^2 + \left(\frac{F+H}{2}\right)^2$
  - D.h. mittlere Randhäufigkeiten

		Beobachter 1		
		+	-	Σ
Beobachter 2	+	.30 A	.10 B	.40 E
	-	.20 C	.40 D	.60 F
	Σ	.50 G	.50 H	20 N

2. Cohens κ und Scotts π

Man kann **Schritt 1 und 2** auch zusammenfassen und die Umwandlung in relative Häufigkeiten in die Formeln einbauen:

- P_E für Cohens κ: $\left(\frac{E}{N} \cdot \frac{G}{N}\right) + \left(\frac{F}{N} \cdot \frac{H}{N}\right)$
- P_E für Scotts π: $\left(\frac{\frac{E}{N}+\frac{G}{N}}{2}\right)^2 + \left(\frac{\frac{F}{N}+\frac{H}{N}}{2}\right)^2$

		Beobachter 1		
		+	-	Σ
Beobachter 2	+	6 A	2 B	8 E
	-	4 C	8 D	12 F
	Σ	10 G	10 H	20 N

2. Cohens κ und Scotts π

Schritt 3:

κ bzw. π berechnen

$$\kappa, \pi = \frac{P_O - P_E}{1 - P_E}$$

		Beobachter 1		
		+	-	Σ
Beobachter 2	+	.30 A	.10 B	.40 E
	-	.20 C	.40 D	.60 F
	Σ	.50 G	.50 H	20 N

Im Beispiel:

- P_O = .70
- P_E = .50 (κ) bzw. .505 (π)
- Cohens $\kappa = \frac{.70-.50}{1-.50} = 0.40$
- Scotts $\pi = \frac{.70-.505}{1-.505} = 0.39$

2. Cohens κ und Scotts π

Interpretation

- κ, π = 1: „perfekte“ Übereinstimmung
- κ, π = 0: keine Übereinstimmung
- κ, π = -1: „perfekte“ Nicht-Übereinstimmung

- Positive Werte: Besser als „durch Zufall“ zu erwarten wäre
- Negative Werte: Schlechter als „durch Zufall“ zu erwarten wäre
- -1: keine beobachteten Übereinstimmungen bei gleichzeitig maximaler Wahrscheinlichkeit zufälliger Übereinstimmungen

2. Cohens κ und Scotts π

Cohens κ ist das am häufigsten angewandte Maß der Übereinstimmung für nominalskalierte Daten (vgl. Wirtz, 2007).

Beurteilungsgüte

Ungefähre Richtlinie für Kappa nach McHugh (2012):

- κ > .90 = fast perfekte Übereinstimmung
- κ = .80 - .90 = starke Übereinstimmung
- κ = .60 - .79 = moderate Übereinstimmung
- κ = .40 - .59 = schwache Übereinstimmung
- κ = .21 - .39 = minimale Übereinstimmung
- κ = .00 - .20 = keine Übereinstimmung

Cohens κ und Scotts π in R

data-Objekt:

	RaterA	RaterB
1	1	1
2	1	1
3	1	1
4	1	1
5	1	1
6	1	1
7	1	0
8	1	0
9	1	0
10	1	0
11	0	1
12	0	1
13	0	0
14	0	0
15	0	0
16	0	0
17	0	0
18	0	0
19	0	0
20	0	0

Code:

```
library(irrCAC)
kappa2.table(table(data)) # cohen‘s kappa
scott2.table(table(data)) # scott‘s pi
```

Outputs:

```
> kappa2.table(table(data))
     coeff.name coeff.val  coeff.se     coeff.ci coeff.pval
1 Cohen's Kappa       0.4 0.2007984 (-0.02,0.82)  6.094e-02

> scott2.table(table(data))
  coeff.name coeff.val  coeff.se       coeff.ci coeff.pval
1 Scott's Pi 0.3939394 0.2064653 (-0.038,0.826)  7.162e-02
```

2. Cohens κ und Scotts π

Spezielle Eigenschaften

- Ein Wert von 1 kann nur erreicht werden, wenn beide Rand**verteilungen** gleich sind
- Ein Wert von -1 kann nur erreicht werden, wenn alle vier Rand**summen** gleich sind („symmetrische Kreuztabelle“)

Exkurs: Cohens κ kann korrigiert werden, so dass es auch bei unterschiedlichen Randverteilungen/-summen maximal werden kann (-1 und +1)

Hinweis: Für den Spezialfall einer Vierfeldertafel, bei der nur eine der 4 Zellen belegt ist, sind Cohens κ und Scotts π nicht definiert

Gleiche & ungleiche Randverteilungen

		Beobachter 1		
Beobachter 2		+	-	Σ
	+	3 A	8 B	11 E
	-	8 C	1 D	9 F
	Σ	11 G	9 H	20 N

gleiche Randverteilung

		Beobachter 1		
Beobachter 2		+	-	Σ
	+	6 A	2 B	8 E
	-	4 C	8 D	12 F
	Σ	10 G	10 H	20 N

ungleiche Randverteilung

Gleiche & ungleiche Randsummen

		Beobachter 1		
Beobachter 2		+	-	Σ
	+	3 A	8 B	11 E
	-	8 C	1 D	9 F
	Σ	11 G	9 H	20 N

ungleiche Randsummen

		Beobachter 1		
Beobachter 2		+	-	Σ
	+	2 A	8 B	10 E
	-	8 C	2 D	10 F
	Σ	10 G	10 H	20 N

gleiche Randsummen

Man kann zwei Fehlerquellen betrachten:

1. Unterschiedliche Wahrnehmungsschwellen (auch: Strenge)

- Ab wann wird ein Verhalten als Beobachtung registriert
- Beispiel: „ähms“ zählen → die eine Person registriert bereits bei „äh“, die andere will ein voll ausgeprägtes „ähm“ hören
- Sichtbar an **ungleichen Randverteilungen** (mit „McNemar“-Test könnte die Hypothese gleicher Randverteilungen in der Population geprüft werden)

2. Mangelnde Konsistenz:

- Beobachterinnen nutzen grundlegend andere Definitionen oder Kriterien
- Sichtbar, wenn Übereinstimmung nicht gut ist, obwohl die Beobachterinnen insgesamt gleiche Wahrnehmungsschwellen haben / gleich streng sind (z.B. haben beide das Merkmal 5-mal beobachtet, aber jeweils bei 5 verschiedenen Personen)

(vgl. Bühner & Ziegler, 2017, S 359 ff.)

Maße für **nominalskalierte** Daten

	Beobachter 1		
Beobachter 2	+	-	Σ
+	40 A	61 B	101 E
-	62 C	37 D	99 F
Σ	102 G	98 H	200 N

Cohens κ und Scotts π sind unjustierte Maße

- Das heißt, dass sie **unterschiedliche Wahrnehmungsschwellen** (= Randverteilungen) der Rater **„bestrafen“**
- Bei unterschiedlichen Randverteilungen fallen unjustierte Koeffizienten (wie Cohens κ & Scotts π) aufgrund dieser „Bestrafung“ niedriger als justierte Koeffizienten (wie Odds Ration & Yules Y)

- **Sind die Randverteilungen ungleich**…
  - …liegen systematische Unterschiede der Rater in der Einschätzung der Grundwahrscheinlichkeiten eines Merkmals vor
  - …sind unterschiedliche Wahrnehmungsschwellen (d.h., Strenge) der Rater Ursache für unterschiedliche Randverteilungen
  - …fällt Cohens κ höher als Scotts π aus
    → Unterschiedliche Wahrnehmungsschwellen bei Scotts π stärker bestraft
    → Diese Eigenschaft kann je nach Fragestellung genutzt werden

Weitere Gründe für mangelnde Übereinstimmung:

- Sind die **Randverteilungen gleich** (d.h., gleiche Strenge) ist eine **mangelnde Übereinstimmung** eindeutig als ein Effekt **mangelnder Konsistenz** interpretierbar
  - D.h. Beobachter nutzen andere Definitionen oder Kriterien
- Vorsicht: Für einen geringen *Koeffizienten* gibt es aber auch andere Ursachen, wie z.B. eine hohe zufällig erwartbare Übereinstimmung durch sehr geringe/hohe **Basisrate**

Weitere Einflussgröße: Basisrate

Tabelle 2: Drei Beispiele für Vierfeldertafeln bei dichotomen Einschätzungen

		Beispiel A			Beispiel B			Beispiel C		
		Beurteiler B			Beurteiler B			Beurteiler B		
		0	1	Σ	0	1	Σ	0	1	Σ
Beurteiler A	0	74	25	**99**	145	18	**163**	94	73	**167**
	1	24	77	**101**	17	20	**37**	4	29	**33**
	Σ	**98**	**102**	**200**	**162**	**38**	**200**	**98**	**102**	**200**
	Cohens κ	,51			,43			,24		

Aus: Wirtz (2006). In Petermann, F. & Eid, M. (Eds.). Handbuch der Psychologischen Diagnostik (S.371).

Gleiche Randverteilungen Mittlere Basisrate von 0/1-Urteilen	Gleiche Randverteilungen Geringe Basisrate von „1“-Urteilen	Ungleiche Randverteilungen Basisrate unklar

Wichtige Eigenschaften von Cohens κ & Scotts π

- ✓ **„Bereinigt“**
  - Von der durch Zufall erwarteten Übereinstimmung
- ✓ **Unjustiert**
  - Unterschiedliche Wahrnehmungsschwellen werden bestraft
- ✓ **Abhängig von der Basisrate**
  - Aber anders als die prozentuale Übereinstimmung: Wenn Merkmale sehr häufig oder sehr selten vorkommen, dann ist die bei Zufall erwartete Übereinstimmung sehr hoch und Kappa / Pi werden sehr klein

- Beide Koeffizienten gibt es auch für mehr als zwei Rater oder Ratings mit mehr als zwei Kategorien („Krippendorffs Alpha“ und „Fleiss‘ Kappa“)

2.3. Odds Ratio / Yules Y

Odds Ratio/ Yules Y sind justierte Maße

- Das heißt sie betrachten nur die **Konsistenz** von Beobachtern:
  - Wie stark stimmen Beobachter überein, ohne dass unterschiedliche Wahrnehmungsschwellen der Beobachter eine Rolle spielen?
  - Im dichotomen Fall bedeutet Konsistenz also: Im Rahmen der aufgetretenen Beobachtungshäufigkeiten überschneiden sich die Beobachtungen
- Es ist also nicht wichtig, ob Beobachter 1 das Merkmal öfter beobachtet (z.B. mehr Personen für geeignet hält) als Beobachter 2
- Justierte Maße sollten dementsprechend dann angewandt werden, wenn unterschiedliche Wahrnehmungsschwellen keine Rolle (!) spielen sollen

Odds Ratio

Schritt 1: Odds (Chance)

Odds (Übereinstimmung) $= \frac{\frac{A}{G}}{1-\frac{A}{G}} = \frac{\frac{A}{G}}{\frac{C}{G}} = \frac{A}{C}$

Lies: B2 vergibt (+), gegeben B1 hat bereits (+) vergeben

Odds (Nicht-Übereinstimmung) $= \frac{\frac{B}{H}}{1-\frac{B}{H}} = \frac{\frac{B}{H}}{\frac{D}{H}} = \frac{B}{D}$

Lies: B2 vergibt (+), gegeben B1 hat bereits (-) vergeben

Exkurs

	Beobachter 1		
Beobachter 2	+	-	Σ
+	6 A	2 B	8 E
-	4 C	8 D	12 F
Σ	10 G	10 H	20

Schritt 2: Odds Ratio (Chancenverhältnis)

$$q = \frac{Odds\ (\text{Übereinstimmung})}{Odds\ (Nicht-\text{Übereinstimmung})} = \cdots = \frac{A \cdot D}{B \cdot C}$$

Interpretation:

- **Gegeben ein Urteil**, um welchen Faktor wächst die *Chance* für nochmal das gleiche Urteil?
- Anders ausgedrückt: Faktor, um den die *Chance* für ein positives (bzw. negatives) Urteil höher ist, wenn die andere Beobachterin bereits ein positives (bzw. negatives) Urteil vergeben hat
- Vorsicht: *Chance* ist nicht das Gleiche wie Wahrscheinlichkeit (WK), sondern *WK / (1- WK)*

Odds Ratio

- Wertebereich liegt zwischen 0 und ∞
- **Interpretation**
  - q = 1: Übereinstimmung und Nicht-Übereinstimmung sind gleich wahrscheinlich = Zufall
  - q < 1: Systematische Nicht-Übereinstimmung
  - q > 1: Systematische Übereinstimmung

Yules Y

Koeffizient dient ausschließlich der **Normierung der Odds Ratio**:

$$\text{Yules Y} = \frac{\sqrt{q}-1}{\sqrt{q}+1}$$

- Liegt zwischen -1 und +1: Ermöglicht den Vergleich mit anderen Zusammenhangsmaßen, die zwischen -1 und +1 liegen
- Scotts π, Cohens κ und Yules Y sind nur bei symmetrischen Kreuztabellen gleich (= alle vier Randsummen sind gleich)

3. Odds Ratio/ Yules Y

Schritt 1:

Odds Ratio berechnen

Im Beispiel:

$$q = \frac{A \cdot D}{B \cdot C} = \frac{6 \cdot 8}{2 \cdot 4} = 6$$

	Beobachter 1			
Beobachter 2		+	-	Σ
	+	6 A	2 B	8 E
	-	4 C	8 D	12 F
	Σ	10 G	10 H	20

→ Gegeben ein Urteil, wächst die Chance für nochmal das gleiche Urteil um den Faktor 6

→ Chance für ein positives Urteil steigt um das 6-fache, wenn bekannt ist, dass die andere Beobachterin auch ein positives Urteil vergeben hat

3. Odds Ratio/ Yules Y

Schritt 2:

Yules Y berechnen

$$Y = \frac{\sqrt{q} - 1}{\sqrt{q} + 1}$$

Im Beispiel:

$$Y = \frac{\sqrt{6} - 1}{\sqrt{6} + 1} = .42$$

		Beobachter 1		
		+	-	Σ
Beobachter 2	+	6 A	2 B	8 E
	-	4 C	8 D	12 F
	Σ	10 G	10 H	20

Einfluss der Basisrate im Vergleich

Tabelle 2: Drei Beispiele für Vierfeldertafeln bei dichotomen Einschätzungen

		Beispiel A			Beispiel B			Beispiel C		
		Beurteiler B			Beurteiler B			Beurteiler B		
		0	1	Σ	0	1	Σ	0	1	Σ
Beurteiler A	0	74	25	**99**	145	18	**163**	94	73	**167**
	1	24	77	**101**	17	20	**37**	4	29	**33**
	Σ	**98**	**102**	**200**	**162**	**38**	**200**	**98**	**102**	**200**
	Cohens κ		,51			,43			,24	
	Odds Ratio		9,50			9,48			9,34	
	Yules *Y*		,51			,51			,51	

Aus: Wirtz (2006). In Petermann, F. & Eid, M. (Eds.). Handbuch der Psychologischen Diagnostik (S.371).

Wichtige Eigenschaften von Odds Ratio & Yules Y

- ✓ **Justiert**:
  - Unterschiedliche Wahrnehmungsschwellen werden nicht besonders bestraft, nur Konsistenz wird berücksichtigt
- ✓ **Robust** gegenüber Veränderungen der Basisrate

3. Beispiele und Fazit zu dichotomen Ratings

Beispiel 1

Eine Beratungsstelle für Hochbegabung untersucht „Hochbegabte“, die aufgrund sehr guter Leistungen von Lehrerinnen in der Beratungsstelle gemeldet werden.

		Beobachter 1		
Beobachter 2		+	-	Σ
	+	17 A	1 B	18 E
	-	1 C	1 D	2 F
	Σ	18 G	2 H	20

Beispiel 1

- Prozentuale Übereinstimmung = .90 (90%)
- Cohens κ (kappa) = .44
- Scotts π (pi) = .44
- Yules Y = .61

		Beobachter 1		
Beobachter 2		+	-	Σ
	+	17 A	1 B	18 E
	-	1 C	1 D	2 F
	Σ	18 G	2 H	20

Beispiel 2

Mit einer Verhaltensbeobachtung soll die Anzahl an Füllwörtern beobachtet werden, aber der Indikator ist nicht konkret genug formuliert, und Beobachter 2 ist etwas unkonzentrierter als Beobachter 1.

		Beobachter 1		
Beobachter 2		+	-	Σ
	+	6 A	2 B	8 E
	-	4 C	8 D	12 F
	Σ	10 G	10 H	20 N

Beispiel 2

- Prozentuale Übereinstimmung = .70
- Cohens κ (kappa) = .40
- Scotts π (pi) = .39
- Yules Y = .42

		Beobachter 1		
		+	-	Σ
Beobachter 2	+	6 A	2 B	8 E
	-	4 C	8 D	12 F
	Σ	10 G	10 H	20 N

Fazit

- Die heute besprochenen Maße sind auf **nominalskalierte Daten** anwendbar. Die Berechnungen sind mit angepassten Formeln auch für nicht-dichotome Beurteilungen möglich.
- Sollen **unterschiedliche Wahrnehmungsschwellen** der Rater besonders bestraft werden, dann sollte Scotts π verwendet werden, sonst Cohens κ.
- Spielt **nur die Konsistenz** des Ratings eine Rolle, dann sollte Yules Y verwendet werden.
- Wenn Yules Y hoch ist, aber Cohens κ / Scotts π nicht, dann weiß man dass die Konsistenz ok ist, aber die Wahrnehmungsschwellen oder die Basisrate ein Problem sind.
- **Wenn möglich, sollte man immer auch direkt die Kreuztabelle betrachten und interpretieren, bevor man sich auf den komplizierten Vergleich der Übereinstimmungsmaße einlässt!**

Flowchart für nominalskalierte Daten

Wann verwende ich welches Maß?

Muss ich die Beobachterübereinstimmung unbedingt quantitativ zusammen fassen?

nein → Kreuztabelle direkt interpretieren

ja ↓

Soll für die bei Zufall erwartete Wahrscheinlichkeit korrigiert werden? Sollen unterschiedliche Wahrnehmungsschwellen berücksichtigt werden (unjustiert vs. justiert)?

nein → **Justiert: Konsistenz:** Prozentuale Übereinstimmung oder Odd‘s Ratio / Yules Y

ja ↓

Sollen die erwarteten Übereinstimmungen aus den ***mittleren*** **Randhäufigkeiten** geschätzt werden?
= Besonders streng

ja → **Unjustiert: Konsistenz + gleiche Wahrnehmungsschwellen:** Scotts π

nein ↓

Sollen die erwarteten Übereinstimmungen aus den ***kombinierten*** **Randhäufigkeiten** geschätzt werden?

ja → **Unjustiert: Konsistenz + gleiche Wahrnehmungsschwellen:** Cohens κ

Das war‘s schon fast für heute …

- *Ausblick:* In der nächsten Vorlesung beschäftigen wir uns mit den Maßen der Beobachterübereinstimmung für ordinal- bzw. intervallskalierte Daten mithilfe von ICCs.

- Aber zuerst: Gibt es offene Fragen zur heutigen Vorlesung?

- Weiterführende Literatur:

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


REHA/794/13.4.2007/Macmillan Originalarbeit 1

Analyse der Beurteilerübereinstimmung für kategoriale Daten mittels Cohens Kappa und alternativer Maße

Analyzing Interrater Agreement for Categorical Data Using Cohen's Kappa and Alternative Coefficients

Autoren M. Wirtz[1], M. Kutschmann[2]

Institute [1] Institut für Psychologie, Pädagogische Hochschule Freiburg
[2] Fakultät für Gesundheitswissenschaften, Universität Bielefeld

Wirtz, M. & Kutschmann, M. (2006). Methoden zur Bestimmung der Beurteilerübereinstimmung. Handbuch der psychologischen Diagnostik (S. 369-380). Göttingen: Hogrefe. https://www.researchgate.net/publication/321255287_Methoden_zur_Bestimmung_der_Beurteilerubereinstimmung

Exkurs: Messmodelle für dichotome Ratings

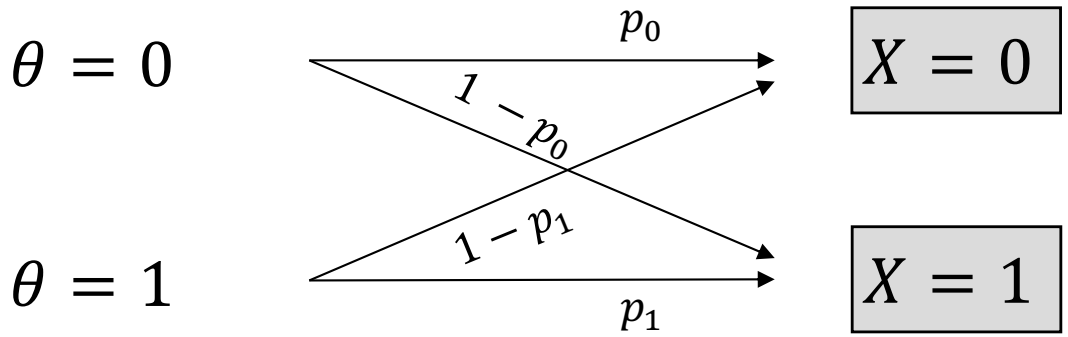
WH: Essentiell paralleles bzw. essentiell τ-äquivalentes Modell für kontinuierliche Ratings

- $X_{iPerson} = \sigma_i + \theta_{Person} + \varepsilon_{iPerson}$ für alle Rater i
- Im Gegensatz zur Vorlesung Testtheorie werden Items durch Rater ersetzt
- Messgenauigkeit quantifiziert durch Reliabilität oder Standardmessfehler
- Grundlage für Maße der Beobachterübereinstimmung bei intervallskalierten Daten (siehe Sitzung 5)

Exkurs: Raschmodell für dichotome Ratings

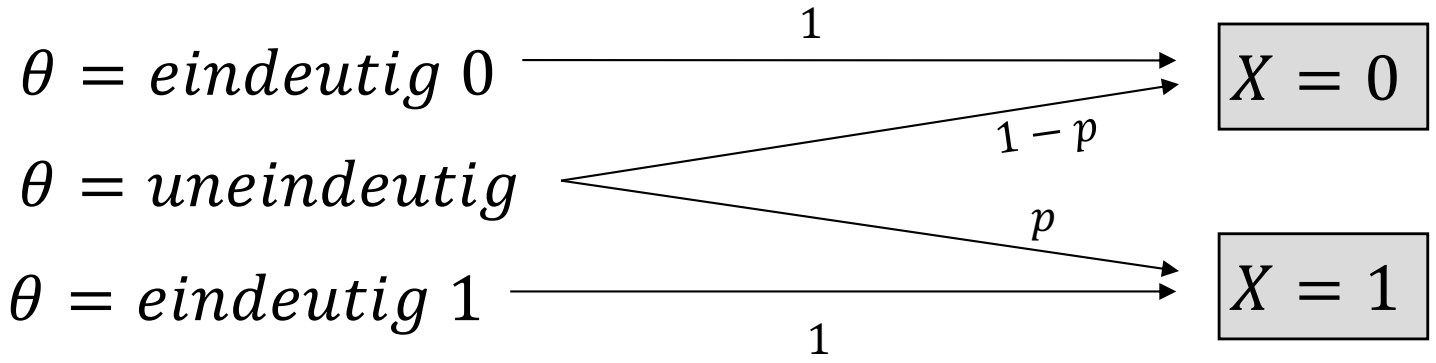
- $P(X_{iPerson} = 1|\theta_{Person}) = \frac{e^{\sigma_i + \theta_{Person}}}{1 + e^{\sigma_i + \theta_{Person}}}$ für alle Rater i
- Annahme: θ_{Person} kontinuierlich (aber $X_{iPerson}$ diskret bzw. dichotom)
- Analogie zu Statistik II: Logistische Regression mit latenter Merkmalsausprägung der Person als Prädiktor
- Testmodelle der *Item Response Theorie* → Vorlesung im Master

Latente Klassenanalyse: Messfehlerperspektive



- Latente Variable und Itemantwort jeweils diskret mit nur zwei Ausprägungen 0 und 1
- Wahrscheinlichkeiten $p_0, p_1, 1-p_0, 1-p_1$ beschreiben den Antwortprozess, z.B. $p_1 = P(X=1|\theta=1)$ und $1-p_1 = P(X=0|\theta=1)$
- Ideale Messung:
  - p_0, p_1 nahe an 1 und $1-p_0, 1-p_1$ nahe an 0
    → Hohe Messgenauigkeit (wenig Messfehler)
  - $p_0, p_1, 1-p_0, 1-p_1$ sind jeweils für alle Rater gleich
    → Keine unterschiedlichen Wahrnehmungsschwellen

Latente Klassenanalyse: Übereinstimmungsperspektive



- Drei latente Zustände: „eindeutig 0“, „eindeutig 1“ und „uneindeutig“
- Eindeutige Beobachtungen werden **immer** richtig beurteilt. Die Wahrscheinlichkeiten p und $1 - p$ beschreiben den Antwortprozess bei uneindeutigen Beobachtungen, z.B. $p = P(X = 1|\theta = uneindeutig)$
- Ideale Messung:
  - $P(\theta = uneindeutig)$ möglichst klein
  - p entspricht dem Anteil von „eindeutig 1“ unter allen „eindeutigen“
    → Kein Bias bei uneindeutigen Beobachtungen
  - p und $1 - p$ sind jeweils für alle Rater gleich
    → Keine unterschiedlichen Wahrnehmungsschwellen

Zusammenhang von Cohens κ & Scotts π mit Messmodellen für dichotome Ratings

- **Messfehlerperspektive:**
  - Keine direkte Entsprechung, aber…
  - *Scotts π entspricht (für großes N) der Intraklassenkorrelation ICC(1,1) bei kontinuierlichen Ratings → Sitzung 5*
  - *Cohens κ entspricht (für großes N) der Intraklassenkorrelation ICC(2,1) bei kontinuierlichen Ratings → Sitzung 5*
- **Übereinstimmungsperspektive:**
  - Annahme einer idealen Messung:
    → Kein Bias bei uneindeutigen Beobachtungen
    → Keine unterschiedlichen Wahrnehmungsschwellen
  - *Cohens κ & Scotts π entsprechen dem Anteil der eindeutig klassifizierbaren Beobachtungen an allen Beobachtungen*