IRT 3: Parameterschätzung, Modelltests und Modellvergleiche

Wiederholung: Skalierung

- In den letzten beiden Kapiteln wurden IRT Modelle für dichotome und ordinale Items eingeführt. Dabei wurde bei der Interpretation der Modelle implizit immer davon ausgegangen, dass sowohl die Werte aller Itemparameter, als auch die Werte aller Personen auf der latenten Variable bekannt sind.

- In der Praxis sind diese Werte natürlich nicht bekannt. Schließlich sollen die IRT Modelle zur Skalierung eines psychologischen Tests verwendet werden.

- Wiederholung Skalierung:
  Ein psychologischer Test gilt als skalierbar, wenn die Zuordnung der Messwerte zu den Personen auf der Basis eines empirisch nachgewiesenen testtheoretischen Modells geschieht.

Ablauf Skalierung

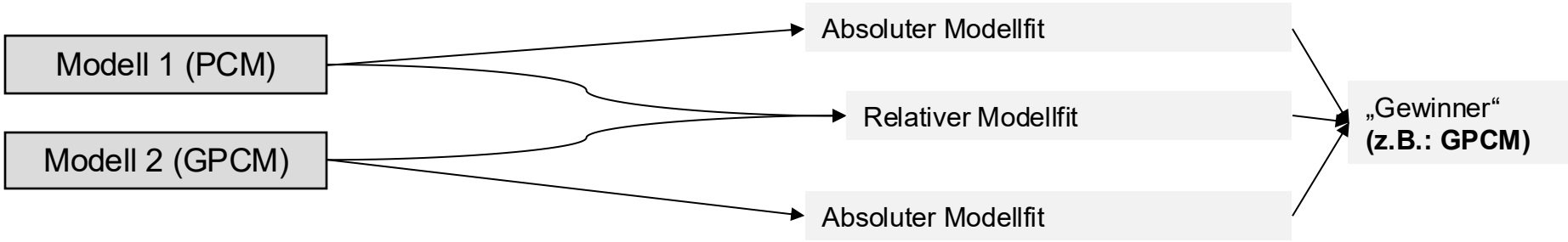
- Wähle für den vorliegenden Datentyp (z.B. dichotome Items) eine Reihe geeigneter Testmodelle aus (z.B. das 1PL und das 2PL Modell).
- Schätze für jedes der ausgewählten Modelle die Itemparameter anhand des vorliegenden Normdatensatzes.

- Finde ein Testmodell, dass empirisch auf die vorliegenden Daten „passt“.
  - Absoluter Modellfit: Kann für ein Modell empirisch nachgewiesen werden, dass alle Modellannahmen erfüllt sind?
  - Vergleichender / Relativer Modellfit: Welches der geschätzten Modelle passt empirisch im Vergleich zu den anderen Modellen am besten?

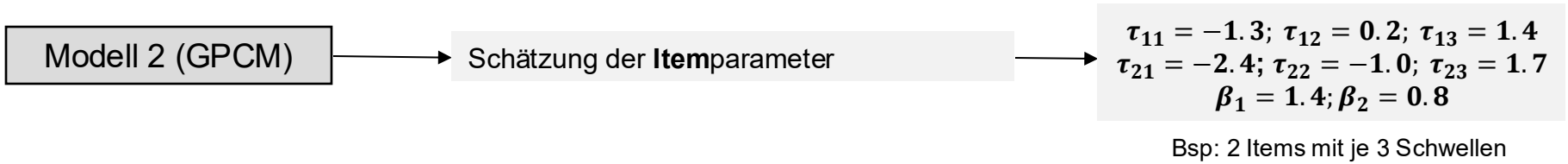
- Wenn ein „passendes“ Testmodell gefunden werden konnte:
  Schätze im Rahmen der Einzelfalldiagnostik für die interessierenden Personen den Wert auf der latenten Variable, mithilfe der geschätzten Itemparameter aus dem anhand der Normstichprobe nachgewiesenen Testmodell.

Überblick

1. Modell aus mehreren Kandidaten auswählen (Modellvergleich)

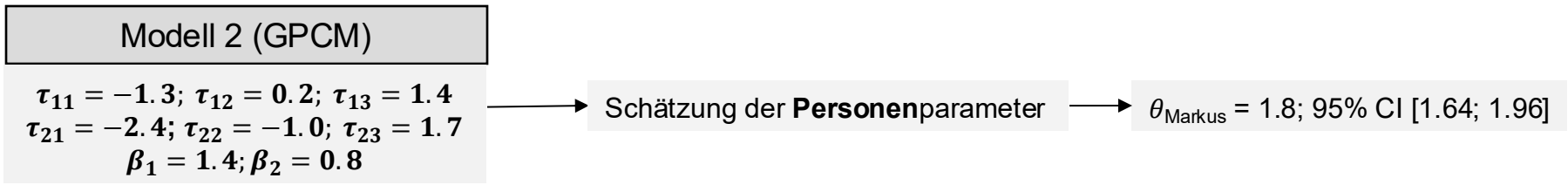


2. Finales Modell: Itemparameter schätzen (Normstichprobe)



Bsp: 2 Items mit je 3 Schwellen

3. Neue Personen (*nach* der Normierung): Personenparameter schätzen



Datenmatrix und Antwortmuster

- Beispiel mit dichotomen Items:

Beobachtete Datenmatrix

	Items 1 … I				
Personen 1 … P	x_{pi}	1	2	3	…	I
	1	1	1	0		0
	2		0			
	…					
	P		1			

$\underline{x_p} = (1; 1; 0; \ldots; 0)\ mit\ p = 1$

Bei I Items mit jeweils K verschiedenen Antwortkategorien sind K^I verschiedene Antwortmuster möglich, die jedoch oft nicht alle beobachtet werden.

- Der Vektor $\underline{x_p}$ enthält die beobachteten Antworten einer Person p auf alle I Items -> Man bezeichnet $\underline{x_p}$ auch als das beobachtete „Antwortmuster“ von Person p
- Wir definieren außerdem die zufällige Matrix X, die alle zufälligen Antworten der P Personen auf die I Items zusammenfasst. Die Zufallsvariable X realisiert sich nach Beobachtung der Itemantworten in der beobachteten Datenmatrix x.

Lokale stochastische Unabhängigkeit

- Für alle besprochenen Item IRT Modelle wird, zusätzlich zu den in der Modellgleichung enthaltenen Annahmen, immer auch die Annahme der lokalen stochastischen Unabhängigkeit getroffen:
  - Die Antwort einer Person p auf ein Item ist (gegeben dem Wert der Person auf der latenten Variable) unabhängig von der Antwort der gleichen Person auf ein anderes Item.
  - Mathematische Formulierung: $P\left(\underline{X}_p = \underline{x}_p \middle| \theta_p\right) = \prod_{i=1}^{I} P(X_{pi} = x_{pi} | \theta_p)$
- Die lokale stochastische Unabhängigkeit entspricht der Annahme unkorrelierter Fehlervariablen in der klassischen Testtheorie.
- Die Bezeichnung „lokale“ stochastische Unabhängigkeit soll betonen, dass die Unabhängigkeitsannahme für jeden möglichen Wert auf der latenten Variable gilt.
- Verletzung der Annahme der lokalen stochastischen Unabhängigkeit zum Beispiel bei logischen Abhängigkeiten zwischen Items, Übungseffekten, oder Mehrdimensionalität des gemessenen Konstrukts.

Schätzung der Itemparameter

Schätzung der Itemparameter mit Maximum Likelihood (ML) Methoden

- Idee der ML-Schätzung:
  Finde für die Itemparameter die Werte, bei denen die beobachteten Daten „am wahrscheinlichsten“ sind.
- Im Gegensatz zu einfacheren statistischen Modellen wie der logistischen Regression gibt es für probabilistische Testmodelle verschiedene ML-Methoden die sich darin unterscheiden, was „am wahrscheinlichsten“ genau bedeutet.
- Gemeinsames Prinzip aller ML-Methoden:
  - Herleiten einer Formel, mit der bei bekannten Parameterwerten eine relevante Wahrscheinlichkeitsaussage hinsichtlich der beobachteten Daten berechnet werden kann
  - Fasse nun diese Formel nicht mehr als Funktion der unbeobachteten Daten bei gegebenen Parametern auf, sondern als Funktion der unbekannten Parameter bei gegebenen beobachteten Daten („Likelihoodfunktion“)
  - Finde die Parameterwerte mit dem höchsten Wert der Likelihoodfunktion
  - Verwende die Werte mit der maximalen Likelihood als Punktschätzer für die Werte der Itemparameter in der Population („ML-Schätzer“)

Verschiedene ML-Methoden zur Schätzung der Itemparameter

- Gemeinsame („unconditional“) ML-Schätzung (UML):
  - Gleichzeitige Schätzung der Itemparameter und der Werte der Personen auf der latenten Variable durch Maximierung der gemeinsamen Likelihood
  - Problem: Instabile Schätzung der Itemparameter (je größer die Stichprobe, desto mehr „Personenparameter“ θ_p müssen mitgeschätzt werden)
- Bedingte („conditional“) ML-Schätzung (CML):
  - Getrennte Schätzung der Itemparameter durch Maximierung einer Likelihood, bei der auf den beobachteten Summenwert der Personen bedingt wird
  - Problem: Nur möglich für Testmodelle mit spezifischer Objektivität (also 1PL-Modell und PCM)
- Marginale ML-Schätzung (MML):
  - Schätzung der Itemparameter durch Maximierung einer Likelihood, aus der die Werte der Personen auf der latenten Variable „herausintegriert“ werden.
  - **Vorteil: Schätzmethode ist für alle besprochenen Testmodelle anwendbar und wird daher in allen unseren praktischen Anwendungen verwendet**

UML-Schätzung

- Obwohl die UML-Schätzung in der Praxis nicht verwendet wird, da sie instabile Schätzungen der Itemparameter liefert[1], ist deren Verständnis für die später folgende Schätzung der Werte der Personen auf der latenten Variable notwendig.
- Idee der UML-Schätzung: Finde die Parameterwerte (sowohl Itemparameter als auch Werte der Personen auf der latenten Variable), für die die beobachteten Daten am wahrscheinlichsten sind.
- Verwende als Likelihood *L* die Wahrscheinlichkeit, die konkret vorliegende Datenmatrix x zu beobachten, sofern die Werte der Personen auf der latenten Variable bekannt sind:

$$L = P(X = x|\theta_1, \dots, \theta_P) = \prod_{p=1}^{P} \prod_{i=1}^{I} P(X_{pi} = x_{pi}|\theta_p)$$

Das doppelte Produkt folgt aus Annahme der Unabhängigkeit der Personen sowie der der Annahme der lokalen stochastischen Unabhängigkeit

- Diese Formel gilt für alle besprochenen IRT Modelle. Dabei wird für $P(X_{pi} = x_{pi}|\theta_p)$ einfach die entsprechende Modellgleichung eingesetzt.

[1] nicht der Fall bei bayesianischer Schätzung, hier ist UML der Standard

Beispiel: UML-Schätzung

- Dichotomes Raschmodell $(P = 3, I = 3)$:

Beobachtete Datenmatrix

x_{pi}	1	2	3
1	1	0	0
2	1	1	0
3	0	0	1

Wahrscheinlichkeit der beobachteten Datenmatrix

$P(X_{pi} = x_{pi}\|\theta_p)$	1	2	3
1	$\frac{e^{(\theta_1-\sigma_1)}}{1+e^{(\theta_1-\sigma_1)}}$	$\frac{1}{1+e^{(\theta_1-\sigma_2)}}$	$\frac{1}{1+e^{(\theta_1-\sigma_3)}}$
2	$\frac{e^{(\theta_2-\sigma_1)}}{1+e^{(\theta_2-\sigma_1)}}$	$\frac{e^{(\theta_2-\sigma_2)}}{1+e^{(\theta_2-\sigma_2)}}$	$\frac{1}{1+e^{(\theta_2-\sigma_3)}}$
3	$\frac{1}{1+e^{(\theta_3-\sigma_1)}}$	$\frac{1}{1+e^{(\theta_3-\sigma_2)}}$	$\frac{e^{(\theta_3-\sigma_3)}}{1+e^{(\theta_3-\sigma_3)}}$

- $L(\sigma_1, \sigma_2, \sigma_3, \theta_1, \theta_2, \theta_3) = P(X = x|\theta_1, \theta_2, \theta_3) = \prod_{p=1}^{3}\prod_{i=1}^{3} P(X_{pi} = x_{pi}|\theta_p) =$

$$= \frac{e^{(\theta_1-\sigma_1)}}{1+e^{(\theta_1-\sigma_1)}} \cdot \frac{1}{1+e^{(\theta_1-\sigma_2)}} \cdot \frac{1}{1+e^{(\theta_1-\sigma_3)}} \cdot \frac{e^{(\theta_2-\sigma_1)}}{1+e^{(\theta_2-\sigma_1)}} \cdot \frac{e^{(\theta_2-\sigma_2)}}{1+e^{(\theta_2-\sigma_2)}} \cdot \frac{1}{1+e^{(\theta_2-\sigma_3)}} \cdot \frac{1}{1+e^{(\theta_3-\sigma_1)}} \cdot \frac{1}{1+e^{(\theta_3-\sigma_2)}} \cdot \frac{e^{(\theta_3-\sigma_3)}}{1+e^{(\theta_3-\sigma_3)}}$$

- Suche die Schätzwerte $\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ für die die beobachteten Daten am wahrscheinlichsten sind, d.h. die gemeinsame Likelihood aller Parameter $L(\sigma_1, \sigma_2, \sigma_3, \theta_1, \theta_2, \theta_3)$ am größten ist.
- Die Berechnung der Schätzwerte erfolgt durch einen numerischen Algorithmus.

Exkurs: CML-Schätzung

- Die CML Schätzung ist eine elegante Schätzmethode für diejenigen IRT Modelle, bei denen die Eigenschaft der spezifischen Objektivität gilt (1PL-Modell und PCM).

- Für den Teil der psychometrischen Community, welcher der spezifischen Objektivität sehr hohe Bedeutung zumisst und Modelle ohne diese Eigenschaft eher ablehnt, stellt die CML-Schätzung die am häufigsten verwendete Schätzmethode dar.

- Wie bereits erwähnt vertreten wir die Meinung, dass auch flexiblere Testmodelle ohne spezifische Objektivität (2PL-Modell und GPCM) für die Praxis sinnvoll sind. Für einen vorliegenden Datensatz wollen wir immer unterschiedlich flexible Modelle hinsichtlich ihrer Passung miteinander vergleichen, um ein möglichst angemessenes Testmodell zu finden. Da dies mit der CML-Schätzung nicht möglich ist, **werden wir uns mit dieser Schätzmethode nicht weiter beschäftigen**.

Hinleitung zur MML-Schätzung: Itemantworten als doppeltes Zufallsexperiment

- Bisher wurde in dieser Vorlesung die Beantwortung eines Items i durch eine feste Person p als einfaches Zufallsexperiment beschrieben.
  -> Markus sitzt vor uns und wir fragen uns, „was wird er wohl ankreuzen?“
- Analog zur klassischen Testtheorie, wird für die MML-Schätzung der Itemparameter erstmals das folgende „doppelte Zufallsexperiment“ betrachtet:
  - Erstes Zufallsexperiment: Ziehen einer zufälligen Person aus der Population. Wir definieren für den Wert der zufällig gezogenen Person auf der latenten Variable die Zufallsvariable Θ_p, die sich nach Durchführung des Zufallsexperiments in dem latenten Wert θ_p realisiert. Der Index p steht dabei nicht mehr für eine bestimmte Person, sondern kennzeichnet im Rahmen der Stichprobenziehung die Person mit der „Nummer“ p.
  - Zweites Zufallsexperiment: Wie bisher beschreibt die Zufallsvariable X_{pi} die Itemantwort der p-ten Person auf Item i, deren Wert θ_p auf der latenten Variable zu diesem Zeitpunkt bereits feststeht. X_{pi} realisiert sich nach Durchführung des Zufallsexperiments in der manifesten Itemantwort x_{pi}.

MML-Schätzung

- „Wie groß ist die Wahrscheinlichkeit $P(\underline{X}_p = \underline{x}_p)$, dass die p-te zufällig gezogene Person ein bestimmtes Antwortmuster $\underline{x}_p$ erzielt?“
- Nehmen wir als Vereinfachung zunächst an, die latente Variable wäre nicht kontinuierlich, sondern könnte nur endlich viele verschiedene Werte annehmen (z.B. $\theta_p \in \{1,2,3\}$)

$$P\left(\underline{X}_p = \underline{x}_p\right) = \sum_{\theta_p=1}^{3} P\left(\underline{X}_p = \underline{x}_p \middle| \Theta_p = \theta_p\right) \cdot P\left(\Theta_p = \theta_p\right) =$$

$$= \sum_{\theta_p=1}^{3} \left\{ \prod_{i=1}^{I} P(X_{pi} = x_{pi} | \Theta_p = \theta_p) \right\} \cdot P\left(\Theta_p = \theta_p\right)$$

Das Produkt folgt aus Annahme der lokalen stochastischen Unabhängigkeit

- Um die „marginale Wahrscheinlichkeit“ $P\left(\underline{X}_p = \underline{x}_p\right)$ berechnen zu können muss man also wissen, wie wahrscheinlich jeder mögliche Wert der latenten Variable ist. Dies erfolgt in der Praxis dadurch, dass man eine bestimmte Wahrscheinlichkeitsverteilung für die latente Variable in der Population annimmt.

Exkurs: MML-Schätzung

- Tatsächlich ist Θ_p jedoch eine kontinuierliche Variable. Damit ersetzt man die Summe durch das Integral und die Wahrscheinlichkeits- durch die Dichtefunktion.

$$P(\underline{X}_p = \underline{x}_p) = \int_{-\infty}^{+\infty} \left\{ \prod_{i=1}^{I} P(X_{pi} = x_{pi} | \Theta_p = \theta_p) \right\} \cdot f(\theta_p) d\theta_p$$

- In der Praxis nimmt man für die Verteilung der Θ_p in der Population in der Regel eine Standardnormalverteilung an. Damit ist die Dichte $f(\theta_p)$ für beliebige Ausprägungen der latenten Variable bekannt.
- (Die angenommene Verteilung der Θ_p entspricht also einer Prior-Verteilung in der Bayes-Statistik; vgl. Diagnostikseminar)
- Betrachtet man die gesamte Stichprobe von zufällig gezogenen Personen, ergibt sich die marginale Wahrscheinlichkeit für die komplette beobachtete Datenmatrix:

$$P(X = x) = \prod_{p=1}^{P} P(\underline{X}_p = \underline{x}_p) =$$

Das Produkt folgt aus Annahme der Unabhängigkeit der Personen

$$= \prod_{p=1}^{P} \int_{-\infty}^{+\infty} \left\{ \prod_{i=1}^{I} P(X_{pi} = x_{pi} | \Theta_p = \theta_p) \right\} \cdot f(\theta_p) d\theta_p$$

MML-Schätzung

- Idee der MML-Schätzung:
  - Verwende $P(X = x)$ als „marginale Likelihood“.
  - Suche die Werte der Itemparameter für die die marginale Likelihood am größten wird und verwende diese Werte als Schätzwerte für die Itemparameter in der Population („MML-Schätzer“).
- Durch die Betrachtung des doppelten Zufallsexperiments und der damit verbundenen Verwendung der marginalen Likelihoodfunktion ist es möglich, die Personenparameter bei der Schätzung der Itemparameter zu ignorieren.
- Dies gelingt deshalb, weil die marginale Likelihood durch die Annahme einer Verteilung für die Personenparameter nur noch von den Itemparametern und nicht mehr von den konkreten Werten der Personen auf der latenten Variable abhängt.

  -> Die Personenparameter werden aus der Likelihood „herausintegriert“

Exkurs: Beispiel: MML-Schätzung (vgl. Folie 14)

- $L(\sigma_1, \sigma_2, \sigma_3) = P(X = x) = \prod_{p=1}^{3} P\left(\underline{X}_p = \underline{x}_p\right) =$

$$= \int_{-\infty}^{+\infty} \left\{ \frac{e^{(\theta_1 - \sigma_1)}}{1 + e^{(\theta_1 - \sigma_1)}} \cdot \frac{1}{1 + e^{(\theta_1 - \sigma_2)}} \cdot \frac{1}{1 + e^{(\theta_1 - \sigma_3)}} \right\} \cdot f(\theta_1) d\theta_1 \cdot$$

$$\cdot \int_{-\infty}^{+\infty} \left\{ \frac{e^{(\theta_2 - \sigma_1)}}{1 + e^{(\theta_2 - \sigma_1)}} \cdot \frac{e^{(\theta_2 - \sigma_2)}}{1 + e^{(\theta_2 - \sigma_2)}} \cdot \frac{1}{1 + e^{(\theta_2 - \sigma_3)}} \right\} \cdot f(\theta_2) d\theta_2 \cdot$$

$$\cdot \int_{-\infty}^{+\infty} \left\{ \frac{1}{1 + e^{(\theta_3 - \sigma_1)}} \cdot \frac{1}{1 + e^{(\theta_3 - \sigma_2)}} \cdot \frac{e^{(\theta_3 - \sigma_3)}}{1 + e^{(\theta_3 - \sigma_3)}} \right\} \cdot f(\theta_3) d\theta_3 \cdot$$

- Suche die Schätzwerte $\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3$ der Itemparameter für die die beobachteten Daten am wahrscheinlichsten sind, d.h. die „marginale Likelihood“ $L(\sigma_1, \sigma_2, \sigma_3)$ am größten ist.
- Die Integrale werden numerisch approximiert, da sie nicht analytisch gelöst werden können. Die Maximierung der Likelihood erfolgt dann ebenfalls mit einem numerischen Algorithmus.

MML-Schätzung und Normierung

- Im Rahmen der MML-Schätzung wird die Verteilungsannahme $\Theta_p \sim N(0;1)$ getroffen. Diese Annahme wirkt auf den ersten Blick ziemlich streng.
- Dabei ist die Fixierung des Erwartungswerts und der Varianz der latenten Variable lediglich eine Form der „Normierung“.
- Wiederholung Normierung:
  Testtheoretische Modelle sind häufig nicht „eindeutig“. Es existieren unendlich viele Kombinationen von Itemparameterwerten und Werten der Personen auf der latenten Variable, die die gleiche Verteilung der Itemantworten implizieren. Um dennoch Modellparameter schätzen zu können, müssen bestimmte Festlegungen getroffen werden, die dafür sorgen, dass nur noch eine einzige Kombination existiert, für die die beobachteten Daten am wahrscheinlichsten sind.
- Solche Normierungen sind auch im Rahmen der klassischen Testtheorie notwendig (vgl. $E(\xi) = 0$ und $Var(\xi) = 1$ im τ-kongenerischen Modell).
- Hinweis: Die Normierung $\Theta_p \sim N(0;1)$ wird in der Regel nur im 2PL-Modell und GPCM getroffen. Im 1PL-Modell und PCM kann man die Varianz schätzen.

Standardfehler für die Schätzung der Itemparameter

- Alle drei ML-Methoden erlauben es, für die Schätzung der Itemparameter auch die Standardfehler zu bestimmen. Die genaue Schätzung der Standardfehler ist kompliziert und wird in dieser Vorlesung nicht behandelt.
- Analog zur logistischen Regression sind die ML-Schätzer der Itemparameter in allen IRT Modellen für große Stichproben approximativ normalverteilt. Damit lassen sich mithilfe der geschätzten Standardfehler approximative Konfidenzintervalle und Hypothesentests berechnen.
- Beispiel für den Schwierigkeitsparameter σ eines Items im 1PL-Modell:
  - $95\%\ KI: \left[\hat{\sigma} \pm z_{1-\frac{0{,}05}{2}} \cdot \widehat{SE}(\hat{\sigma})\right] = \left[\hat{\sigma} \pm 1{,}96 \cdot \widehat{SE}(\hat{\sigma})\right]$
  - $H_0: \sigma \leq \sigma_0$, oder $H_0: \sigma \geq \sigma_0$, oder $H_0: \sigma = \sigma_0$
  - Approximativ gilt: $\frac{\hat{\sigma}-\sigma_0}{\widehat{SE}(\hat{\sigma})} \sim N(0;1)$ (d.h., unter der H_0 ist diese Teststatistik annähernd normalverteilt)

Modelltests

Modelltests

- Es existieren eine Reihe verschiedener Modelltests und Informationskriterien, um die Passung eines IRT Modells empirisch zu überprüfen:
  - Absolute Modellpassung (Globale Modelltests):
    - Pearson-χ^2-Test
  - Modellvergleiche (Relative Modelltests):
    - Likelihood Ratio Test (LRT) für „geschachtelte“ Modelle
    - Informationskriterien: AIC und BIC
- Wie bei der Schätzung der Itemparameter werden wir im Folgenden nur Modelltests und Kriterien besprechen, die für alle besprochenen IRT Modelle gleichermaßen verwendet werden können.
- Vor allem für die Modelle mit spezifischer Objektivität existieren viele weitere Modelltests mit teilweise besseren statistischen Eigenschaften. Die meisten dieser Tests eignen sich in der Praxis dementsprechend nur, solange man sich auf die weniger flexiblen Raschmodelle (1PL und PCM) beschränken möchte.

Globaler Modelltest: Pearson-χ^2-Test

- Idee des Pearson-χ^2-Tests: „Stimmen die beobachteten Häufigkeiten der Antwortmuster im Datensatz mit den durch das Modell implizierten erwarteten Häufigkeiten aller möglichen Antwortmuster überein?“

$$\chi^2 = \sum_{s=1}^{K^I} \frac{\left\{N_{\underline{x}_s} - E(N_{\underline{x}_s})\right\}^2}{E(N_{\underline{x}_s})} \qquad mit \qquad E\left(N_{\underline{x}_s}\right) = N \cdot P\left(\underline{X}_s = \underline{x}_s\right)$$

- Die Teststatistik ist unter der Nullhypothese, dass das betrachtete Testmodell in der Population gilt approximativ χ^2-verteilt mit $df = K^I - N_{par} - 1$.

K: Anzahl der Antwortkategorien
I: Anzahl der Items
N: Anzahl der Personen
$\underline{x}_s$: Eines von K^I möglichen Antwortmustern
$N_{\underline{x}_s}$: Beobachtete Häufigkeit des Antwortmusters $\underline{x}_s$
$E(N_{\underline{x}_s})$: Erwartete Häufigkeit des Antwortmusters $\underline{x}_s$

Exkurs: Pearson-χ^2-Test mit Bootstrap

- Problem des Pearson-χ^2-Tests:
  Damit unter der Nullhypothese die χ^2-Verteilung der Teststatistik approximativ gilt, sollte im vorliegenden Datensatz jedes theoretisch mögliche Antwortmuster mindestens einmal beobachtet worden sein.
  Dies ist bei K^I möglichen Antwortmustern in der Praxis jedoch nur bei extrem großen Stichproben in Kombination mit relativ wenigen Items möglich.
- Zur Verbesserung dieses Problems ist es möglich, die tatsächliche Verteilung der Teststatistik mithilfe eines parametrischen Boostrapverfahrens zu approximieren:
  - Verwende die Schätzungen der Itemparameter aus der Stichprobe als bekannte Parameterwerte in der Population und simuliere wiederholt Datensätze im gleichen Format wie der beobachtete Datensatz.
  - Berechne die Teststatistik für jeden der simulierten Datensätze
- Vergleiche den Wert der Teststatistik für den tatsächlich beobachteten Datensatz mit der Verteilung der Teststatistik basierend auf den simulierten Datensätzen.

Vergleichender Modelltest: Likelihood-Ratio-Test (LRT)

- Idee des Likelihood-Ratio-Test : „Passt ein restriktives Modell (mit Likelihood L_0) auf die vorliegenden Daten genauso gut, wie ein flexibleres Obermodell (mit Likelihood L_1)?“

$$\chi^2 = -2 \cdot \ln\left(\frac{L_0}{L_1}\right)$$

- Die Teststatistik ist unter der Nullhypothese, dass in der Population beide Modelle gleich gut passen approximativ χ^2–verteilt mit $df = N_{par}^{(1)} - N_{par}^{(0)}$.

L_0: Wert der zur Schätzung des restriktiveren Modells verwendeten Likelihoodfunktion, ausgewertet am ML-Schätzer

L_1: Wert der zur Schätzung des flexibleren Modells verwendeten Likelihoodfunktion, ausgewertet am ML-Schätzer

$N_{par}^{(0)}$: Anzahl der zu schätzenden Modellparameter im restriktiveren Modell

$N_{par}^{(1)}$: Anzahl der zu schätzenden Modellparameter im flexibleren Modell

Genestete Modelle

- Der LRT darf nur verwendet werden, wenn das restriktivere Modell (mit Likelihood L_0) in dem flexibleren Modell (mit Likelihood L_1) „genestet“ (bzw. geschachtelt) ist.
- Ein Modell ist in einem anderen Modell genestet, wenn es sich durch eine echte Restriktion des flexibleren Obermodells ergibt (z.B., indem man freie Parameter des flexibleren Modells auf einen bestimmten Wert fixiert).

- Beispiele:
  - Das 1PL-Modell ist im 2PL-Modell genestet, da die Modellgleichung des 1PL-Modells der Modellgleichung des 2PL-Modells mit $\beta_i = 1$ für alle Items i entspricht.
  - Das PCM ist im GPCM genestet, da die Modellgleichung des PCM der Modellgleichung des GPCM mit $\beta_i = 1$ für alle Items i entspricht.

Modellvergleich mit Informationskriterien

- Zum Vergleich der Passung mehrerer statistischer Modelle können neben dem LRT auch „Informationskriterien“ verwendet werden.
- Praktisches Vorgehen:
  - Berechne das Informationskriterium für jedes der zu vergleichenden Modelle
  - Wähle das Modell mit dem **niedrigsten** Kriteriumswert aus. Je niedriger der Kriteriumswert, desto besser passen die Daten zum jeweiligen Modell.
- Die zwei am häufigsten verwendete Informationskriterien sind:
  - Akaike Informationcriterion:

$$AIC = -2 \cdot \ln(L) + 2 \cdot N_{par}$$

  - Bayesian Informationcriterion:

$$BIC = -2 \cdot \ln(L) + \ln(N) \cdot N_{par}$$

N: Größe der Stichprobe

N_{par}: Anzahl der zu schätzenden Modellparameter.

L: Wert der zur Schätzung des Modells verwendeten Likelihoodfunktion, ausgewertet am ML-Schätzer.

Modellvergleich mit Informationskriterien

- Informationskriterium als Gewichtung von Modellfit und Modellkomplexität:
  - Je besser ein Modell auf die vorliegende Daten passt, desto höher ist der Wert der Likelihood am ML-Schätzer („hohe Wahrscheinlichkeit der Daten“).
  - Die Modellpassung ist für ein flexibleres Modell tendenziell besser, als für ein restriktiveres Modell ($L_1 > L_0$). Jedoch hat ein flexibleres Modell tendenziell auch mehr zu schätzende Modellparameter $N_{par}^{(1)} > N_{par}^{(0)}$.
  - Im BIC werden Modelle mit vielen Parametern stärker „bestraft“ als im AIC ($\ln(N) \cdot N_{par} > 2 \cdot N_{par}$ für $N > 7$), d.h. BIC wählt eher das sparsame Modell
- Vorteile der Informationskriterien:
  - Modellvergleich möglich auch im Falle nicht genesteter Modelle
  - Bei kleinen Stichproben oft bessere Ergebnisse selbst bei genesteten Modellen
- Nachteile der Informationskriterien:
  - Der Wert des Informationskriteriums kann nicht absolut interpretiert werden
  - BIC und AIC kommen nicht immer zu gleichen Entscheidungen

Was tun, wenn keines der Testmodelle in der Praxis gilt?

- Inhaltlich steht hinter den Testmodellen eine zusammengesetzte Hypothese:
  - (A) Die latente Variable existiert und
  - (B) kann durch die Items des vorliegenden Tests erfasst werden und
  - (C) der Zusammenhang zwischen Items und latenter Variable ist wie im Modell spezifiziert.
- Wenn alle Testmodelle durch die absoluten Hypothesentests abgelehnt werden, kann entweder (A) (B) oder (C) falsch sein.
- Vorgehen, um herauszufinden, woran es liegen könnte:
  - überprüfen, ob es an (C) liegt: Komplexere (z.B. nonlineare) Modelle ausprobieren (aber irgendwann braucht man extrem große Stichproben)
  - überprüfen, ob es an (B) liegt: eventuell geben die Modelltests Aufschluss über Verbesserungsmöglichkeiten in Bezug auf den Test: z.B. könnte sich zeigen, dass einzelne Items Probleme verursachen
  - Falls weder (C) noch (B) der Grund für die Ablehnung des Modells sind:
    -> Revision der Theorie bezüglich (A)

Schätzung der Personenparameter

Schätzung der Werte der Personen auf der latenten Variable

- Sobald ein IRT Modell gefunden wurde, das auf die vorliegenden Daten ausreichend gut passt, liegen **Schätzwerte für die Itemparameter** vor.
- In der Einzelfalldiagnostik ist man nun daran interessiert, für Personen deren Wert auf der latenten Variable zu schätzen.
- Für diese **Schätzung der Werte der Personen** auf der latenten Variable stehen wiederum mehrere verschiedene Methoden zur Verfügung:
  - UML-Schätzung
  - WML-Schätzung
  - MAP/EAP-Schätzung
- Schätzt man die Itemparameter mit der CML-Methode, verwendet man für die Werte der Personen auf der latenten Variable in der Regel die WML-Schätzung.
- Wir werden in dieser Vorlesung für alle praktischen Anwendung entweder den MAP oder den EAP Schätzer verwenden, da diese Methoden mit der von uns verwendeten MML-Schätzung der Itemparameter theoretisch besser vereinbar sind.

UML-Schätzung der Personenparameter

- Idee der UML-Schätzung:
  Setze die ML-Schätzungen für die Itemparameter in die gemeinsame Likelihood ein und suche für jede Person den Wert auf der latenten Variable, bei dem der Wert der Likelihood am größten ist.
- Beispiel von Folie 11: Dichotomes Raschmodell $(I = 3)$
  - Schätzwerte der Itemparameter: $\hat{\sigma}_1 = -0{,}8;\ \ \hat{\sigma}_2 = 0{,}8;\ \ \hat{\sigma}_3 = 0{,}8$
  - „Person 1 hat das 1. aber nicht das 2. und 3. Item gelöst“ -> $\underline{x}_1 = (1; 0; 0)$
  - $P\left(\underline{X}_1 = \underline{x}_1 \middle| \theta_1\right) = \prod_{i=1}^{I} P(X_{1i} = x_{1i}|\theta_1) =$

$$= \frac{e^{(\theta_1 - (-0{,}8))}}{1 + e^{(\theta_1 - (-0{,}8))}} \cdot \frac{1}{1 + e^{(\theta_1 - 0{,}8)}} \cdot \frac{1}{1 + e^{(\theta_1 - 0{,}8)}}$$

  - Suche den Wert $\hat{\theta}_1$ für den die Likelihood $L(\theta_1) = P\left(\underline{X}_1 = \underline{x}_1 \middle| \theta_1\right)$, bei der bereits die Schätzwerte $\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3$ eingesetzt wurden, am größten ist.
  - Der Schätzwert $\hat{\theta}_1$ kann numerisch bestimmt werden.
    Für die restlichen Personen wird genauso verfahren.
- ! Potentielles Problem: Wenn Personen alle oder kein Item gelöst haben, kann deren Personenparameter nicht berechnet werden (die log-Likelihood wird unendlich)

WML-Schätzung der Personenparameter

- Die WML-Schätzung ist eine gewichtete Schätzmethode („Weighted ML“), die zwei Nachteile des UML-Schätzers ausgleichen soll:
  - Möglichkeit einen Wert auf der latenten Variable für Personen schätzen zu können, die alle Items gelöst oder kein Item gelöst haben
  - Reduzierung des Bias des UML-Schätzers (nicht behandelt)

- Idee der WML-Schätzung:
  Suche den Wert $\hat{\theta}_p$ für den die *gewichtete* Likelihood $L(\theta_p) \cdot w(\theta_p) = P(\underline{X}_p = \underline{x}_p | \theta_p) \cdot w(\theta_p)$, bei der bereits die Schätzwerte der Itemparameter eingesetzt wurden, den größten Wert annimmt.

- Die genaue Form der Gewichtungsfunktion $w(\theta_p)$ ist kompliziert und wird in dieser Vorlesung nicht behandelt.

MAP/EAP-Schätzung der Personenparameter

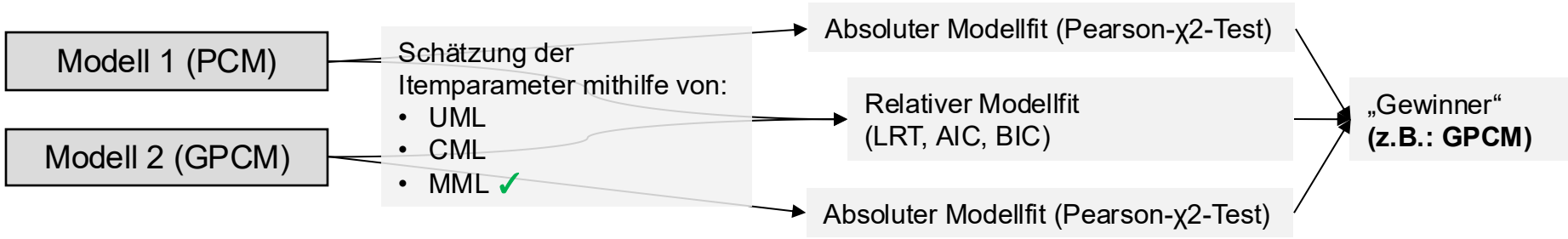
- Die MAP-Schätzung („Maximum-a-posteriori“) arbeitet mit einer alternativen Gewichtungsfunktion, bei der anders als bei der WML-Schätzung nicht der Bias der UML-Schätzung, sondern deren Varianz reduziert werden soll.
- Idee der MAP-Schätzung:
  Suche den Wert $\hat{\theta}_p$ für den die mit der Wahrscheinlichkeitsdichte $f(\theta_p)$ gewichtete Likelihood $L(\theta_p) \cdot f(\theta_p) = P(\underline{X}_p = \underline{x}_p | \theta_p) \cdot f(\theta_p)$, bei der bereits die Schätzwerte der Itemparameter eingesetzt wurden, am größten ist.
- Verwendet man zur Schätzung der Itemparameter die MML-Methode, so wurde dabei für die Werte der Personen auf der latenten Variable aus Gründen der Normierung bereits eine Verteilungsannahme getroffen. Es ist daher naheliegend, bei der MAP-Schätzung für $f(\theta_p)$ ebenfalls die Dichte der Standardnormalverteilung zu verwenden.
- Der mit dieser Methode gefundene Wert $\hat{\theta}_p$ entspricht dem Maximum der sogenannten „a-posteriori Verteilung“ des Parameters θ_p (MAP). Anstatt des Maximums kann auch der Erwartungswert (EAP) als Schätzer verwendet werden.

Standardmessfehler

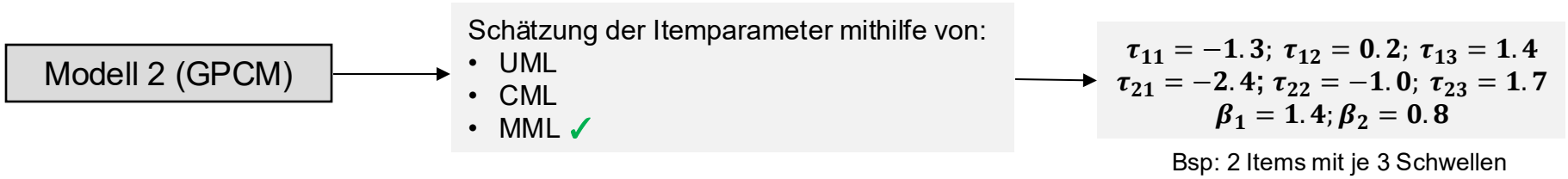
- Für alle vorgestellten Methoden zur Schätzung der Personenparameter ist es möglich, den Standardfehler der Schätzfunktion für θ_p zu bestimmen.
- Anders als in der klassischen Theorie ist dieser „Standardmessfehler“ nicht für alle Personen gleich, sondern hängt wiederum vom Schätzwert $\hat{\theta}_p$ ab (mehr dazu im Kapitel zum adaptiven Testen). Die genaue Schätzung des Standardmessfehlers ist kompliziert und wird in dieser Vorlesung nicht behandelt.
- Analog zu den Itemparametern ist der ML-Schätzer für θ_p in allen IRT Modellen approximativ normalverteilt. Damit lassen sich mithilfe des geschätzten Standardmessfehlers approximative Konfidenzintervalle und Hypothesentests berechnen:
  - $95\%\ KI: \left[\hat{\theta} \pm z_{1-\frac{0{,}05}{2}} \cdot \widehat{SE}(\hat{\theta})\right] = \left[\hat{\theta} \pm 1{,}96 \cdot \widehat{SE}(\hat{\theta})\right]$
  - $H_0: \hat{\theta} \leq \hat{\theta}_0$, oder $H_0: \hat{\theta} \geq \hat{\theta}_0$, oder $H_0: \hat{\theta} \neq \hat{\theta}_0$
  - Approximativ gilt: $\frac{\hat{\theta}-\hat{\theta}_0}{\widehat{SE}(\hat{\theta})} \sim N(0;1)$

Überblick

1. Modell aus mehreren Kandidaten auswählen (Modellvergleich)



2. Finales Modell: Itemparameter schätzen (Normstichprobe)



Bsp: 2 Items mit je 3 Schwellen

3. Neue Personen (*nach* der Normierung): Personenparameter schätzen