Modelle zur Leistungsbewertung SS 2002

D. Lind BUGH Wuppertal

I Deskriptive Statistik

§1 Merkmale

Beispiel 1:

Bei allen Schülern einer Schulklasse wurden die Körpergrößen gemessen und die *Häufigkeitsverteilung* wurde in einer Häufigkeitstabelle festgehalten:

„Intervall“	I_1	I_2	$\dots$	I_n
relative Häufigkeit	h_1	h_2	$\dots$	h_n

$$\left(\sum_{r=1}^{n} h_r \;=\; 1\right)$$

$I_1, \dots, I_n$ sind hier Längenintervalle, z.B. $I_3 := [120 \text{ cm}; 121 \text{ cm}[$ und h_k gibt den Anteil der Schüler an, deren Körpergröße in das Intervall I_k fällt. Man nennt die Eigenschaft „Körpergröße“ ein *Merkmal* der Schüler, bezeichnet die Schüler als *Merkmalsträger* und die jeweils gemessene Länge „vom Scheitel bis zur Sohle“ als *Merkmalsausprägungen*.

Bei Häufigkeitsverteilungen dieses Typs sollte stets die Gesamtzahl N der Merkmalsträger angegeben werden, damit rückwirkend die Anzahlen der Merkmalsträger in den Klassen bestimmt werden können.

Bei Längen und Gewichten macht es Sinn, bei Vergleichen von Differenzen oder Messwertverhältnissen zu sprechen. Bei Temperaturen in Grad Celsius wird man keine multiplikativen Vergleiche mehr machen (allenfalls bei Temperaturerhöhungen in der gleichen Richtung). Bei Verhaltensmerkmalen, wie *Schulleistungen* ist es überhaupt erst einmal klärungsbedürftig, ob man hier von Ausprägungen sprechen kann, die sich mit Hilfe vob Maßzahlen beschreiben und numerisch vergleichen lassen:

Beispiel 2:

Angenommen, eine Lehrperson ist in der Lage, bei einer in Mathematik unterrichteten Schulklasse mit 20 Schülern bei je zwei Schülern a und b entscheiden zu können, wer über ein Halbjahr betrachtet „leistungsfähiger“ war (eine sehr realitätsferne Annahme!). Das Resultat aller Einschätzungen könnte bei konsistenten Entscheidungen in folgender Form angegeben werden:

$$s_1 \prec s_2 \prec \dots \prec s_{20} \quad (\text{mit } a \prec b \quad :\Longleftrightarrow \quad a \text{ war schwächer als } b)$$

Das „Merkmal“ *Mathematikleistung* besitzt wohl verschiedene Ausprägungen, ist aber damit noch nicht quantifiziert. Denkbar sind z.B. folgende „Quantifizierungen“:

Vorschlag 1: Man nehme den *Rangplatz*.

Vorschlag 2: Die Lehrperson gibt *Noten mit Dezimalzwischenwerten*.

Vorschlag 3: Die Lehrperson schätzt bei jedem Schüler, wieviel % der verlangten Leistung er durchschnittlich im Halbjahr gebracht hat.

Dann könnte durchaus folgende Tabelle aus der Gegenüberstellung der Varianten resultieren (die Schülerbezeichnungen sind dabei weggelassen):

Vorschl. 1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Vorschl. 2	5,5	5,1	4,8	4,4	4,2	4,0	3,7	3,5	3,2	3,1	3,0	2,9	2,8	2,7	2,4
Vorschl. 3	29	36	42	52	60	64	72	75	80	81	82	83	84	85	88

16	17	18	19	20
2,2	2,1	1,7	1,3	1,1
90	91	93	95	96

Trägt man hier die Bewertungen der Vorschläge 2 und 3 gegen die Rangziffern in einem Koordinatensystem auf, so liegen die Punkte jeweils auf einer gekrümmten Linie:

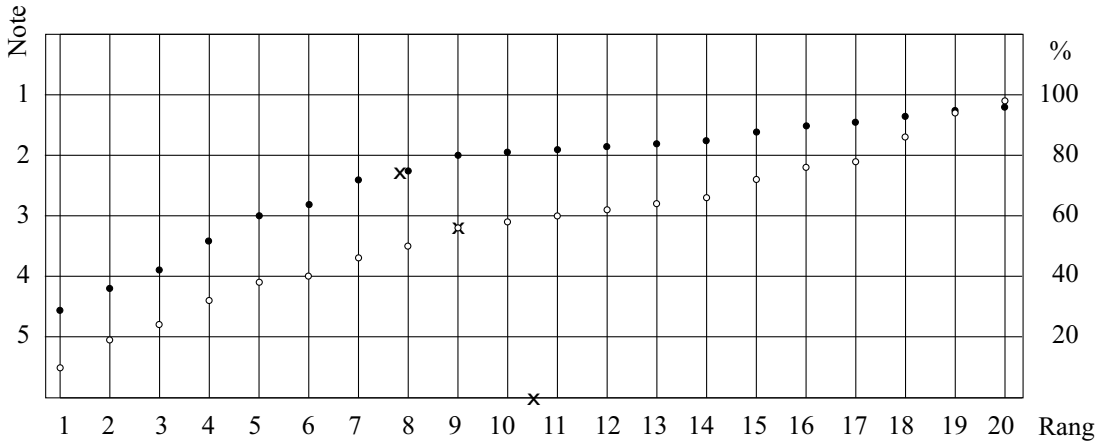


Fig. 1

Ein Vergleich des „mittleren Rangplatzes“ 10,5 mit dem gerundeten arithmetischen Mittelwert 3,2 der Noten und dem arithmetischen Mittelwert 74 der Prozentsätze zeigt, dass diese Werte ihre jeweilige Wertereihe nicht in der gleichen Weise unterteilen. Der mittlere Rangplatz zerlegt die Rangziffernmenge in eine untere und eine ober Hälfte, die arithmetischen Mittelwerte der beiden anderen Wertereihen sind gegenüber der „Tabellenmitte“ nach links verschoben. Sieht man alle drei Quantifizierungen als erlaubt an, so ist nur der sogenannte *Medianwert* zur Kennzeichung der „Merkmalsverteilungsmitte“ angemessen.

Die beiden Beispiele gehören zur Klasse der sogenannten *quantitativen Merkmale*. Bei solchen Merkmalen ist es möglich, für die Merkmalsausprägungen mindestens eine *Kleinerrelation* über einen Vergleich von Merkmalsträgern zu definieren und eine damit verträgliche Abbildung ψ der Menge $\mathcal{M}$ der Merkmalsträger in die Menge $\mathbb{R}$ der reellen Zahlen anzugeben. Man kennzeichnet quantitative Merkmale durch die Freiheiten, die man bei der Wahl von ψ hat, und unterscheidet im Wesentlichen folgende *Skalentypen*:

Definition 1.1 *(Ordinalskala)*

Ein Merkmal heißt genau dann ***ordinal skaliert****, wenn es eine definierende strenge Ordungsrelation $\prec$ in der Menge $\mathcal{M}$ aller Merkmalsträger gibt und mindestens eine Abbildung $\mu : \mathcal{M} \to \mathbb{R}$ mit*

$$a \prec b \iff \mu(a) < \mu(b) \text{ für alle } a, b \in \mathcal{M}$$

oder

$$a \prec b \iff \mu(a) > \mu(b) \text{ für alle } a, b \in \mathcal{M}$$

existiert.
*Man nennt μ eine **Quantifizierung** des Merkmals.*

Wenn die Relation $\prec$ *nicht* linear ist, gibt es mindestens zwei verschiedene Merkmalsträger a und b, für die weder $a \prec b$ noch $b \prec a$ gilt. Dann muss für die durch μ zugewiesenen Werte offensichtlich $\mu(a) = \mu(b)$ gelten. Also ist die durch

$$x \sim y \quad :\Longleftrightarrow \quad \mu(x) = \mu(y)$$

in $\mathcal{M}$ definierte Relation $\sim$ eine nichttriviale Äquivalenzrelation. Wenn $\prec$ linear ist, sind dagegen alle Äquivalenzklassen von $\sim$ einelementig.

Mit μ ist auch jede mit Hilfe einer streng monotonen Transformation $\tau : \mathbb{R} \to \mathbb{R}$ definierte Abbildung $\mu^* : \mathcal{M} \to \mathbb{R}$ mit $\mu^*(a) := \tau(\mu(a))$ eine Quantifizierung.

Werden Schulnoten auf der Basis von „Einschätzungen“ vergeben, so haben sie wohl nur Ordinalskalenqualität. Es sind zwar nicht beliebige streng monotone Transformationen erlaubt, trotzdem ist der Vorrat an Transformationen größer als bei den nun folgenden Skalentypen.

Definition 1.2 *(Intervallskala)*
*Ein Merkmal heißt genau dann **intervallskaliert**, wenn es neben einer strengen Ordungsrelation $\prec$ in der Menge $\mathcal{M}$ aller Merkmalsträger eine Relation $\lhd$ zum Vergleich von Paaren (a, b) und (c, d) gibt, eine Abbildung $\mu : \mathcal{M} \to \mathbb{R}$ mit*

$$a \prec b \iff \mu(a) < \mu(b) \text{ für alle } a, b \in \mathcal{M}$$

und

$$(a, b) \lhd (c, d) \iff \mu(b) - \mu(a) < \mu(d) - \mu(c) \text{ für alle } a, b, c, d \in \mathcal{M}\,.$$

existiert und nur Abbildungen des Typs $\mu^* : a \mapsto u \cdot \mu(a) + v$ mit $u, v \in \mathbb{R}$ und $u \neq 0$ als Quantifizierungen erlaubt sind.

Bei intervallskalierten Merkmalen sind also nur streng monotone **lineare** Transformationen zwischen Quantifizierungen zulässig. Beispiele für solche Skalen liefern Temperaturmessungen mit freier Wahl des Nullpunkts und der Einheit (Fahrenheit- und Celsius-Skala).

Mit viel Aufwand könnte man die Einschränkung der Menge aller Skalentransformationen auf lineare Transformationen durch explizite Angabe weiterer Relationen in $\mathcal{M}$ erzwingen, mit denen μ verträglich ist. Dies würde darauf hinauslaufen, dass für alle $a, b, c, d \in \mathcal{M}$ mit $a \prec b$ und $c \prec d$ der Quotient

$$\frac{\mu(b) - \mu(a)}{\mu(d) - \mu(c)}$$

im Sinne der Verträglichkeit von μ mit vierstelligen „Messrelationen“ in $\mathcal{M}$ interpretierbar ist.

Man kann sofort nachrechnen, dass bei Merkmalen mit Intervallskalenqualität die Berechnung des arithmetischen Mittelwerts Sinn macht, da die „relative Lage“ dieses Wertes in einer Wertereihe

nicht mehr von der Quantifizierung abhängt. Würden die zu erbringenden Schulleistungen im Sinne zählbarer Einzelbausteine von einer „Normierungsbehörde“ festgeschrieben und dürften Noten nur durch lineare Transformationen der Erfolgsquoten berechnet werden, so hätten diese Noten Intervallskalenqualität und würden die Interpretion von Durchschnittswerten auf der Merkmalsebene erlauben.

Definition 1.3 *(Verhältnisskala)*

*Ein Merkmal heißt genau dann **verhältnisskaliert**, wenn es neben einer strengen Ordungsrelation $\prec$ in der Menge $\mathcal{M}$ aller Merkmalsträger eine Relation $\lhd^*$ zum Vergleich von Paaren (a, b) und (c, d) gibt, eine Abbildung $\mu : \mathcal{M} \to \mathbb{R}^+$ mit*

$$a \prec b \iff \mu(a) < \mu(b) \text{ für alle } a, b \in \mathcal{M}$$

und

$$(a, b) \lhd^* (c, d) \iff \mu(b) : \mu(a) < \mu(d) : \mu(c) \text{ für alle } a, b, c, d \in \mathcal{M}\,.$$

existiert und nur Abbildungen des Typs $\mu^* : a \mapsto u \cdot \mu(a)$ mit $u \in \mathbb{R}^+$ als Quantifizierungen erlaubt sind.

Bei verhältnisskalierten Merkmalen sind also nur proportionale Transformationen zwischen Quantifizierungen zulässig. Beispiele für solche Skalen liefern Längenmessungen mit freier Wahl der Maßeinheit (z.B. Messung in Zoll und Zentimeter).

Definition 1.4 *(Absolute Skala)*

*Ein Merkmal heißt genau dann **absolut skaliert**, wenn es genau eine Abbildung $\mu : \mathcal{M} \to \mathbb{R}$ der Menge $\mathcal{M}$ aller Merkmalsträger gibt, die als Quantifizierung zulässig ist.*

Die Messung der Temperatur in Grad Kelvin erfolgt auf einer solchen Skala, da sowohl der Nullpunkt als auch die Einheit festgelegt sind. Auch Wahrscheinlichkeiten und Anzahlen werden auf absoluten Skalen gemessen.

Abschließend soll ein weiterer „Subtyp“ von Intervallskalen genannt werden, bei dem zwar die *Einheit*, nicht aber der *Nullpunkt* festgelegt ist:

Definition 1.5 *(Differenzenskala)*

*Ein Merkmal heißt genau dann **differenzenskaliert**, wenn es neben einer strengen Ordungsrelation $\prec$ in der Menge $\mathcal{M}$ aller Merkmalsträger eine Relation $\lhd$ zum Vergleich von Paaren (a, b) und (c, d) gibt, eine Abbildung $\mu : \mathcal{M} \to \mathbb{R}$ mit*

$$a \prec b \iff \mu(a) < \mu(b) \text{ für alle } a, b \in \mathcal{M}$$

und

$$(a, b) \lhd (c, d) \iff \mu(b) - \mu(a) < \mu(d) - \mu(c) \text{ für alle } a, b, c, d \in \mathcal{M}\,.$$

existiert und nur Abbildungen des Typs $\mu^* : a \mapsto \mu(a) + v$ mit $v \in \mathbb{R}$ als Quantifizierungen erlaubt sind.

Einen Überblick über die gebräuchlicheren Skalentypen bringt die folgende Tabelle, in der die Unterscheidung nach den den erlaubten Transformationen zwischen Quantifizierungen erfolgt. Dabei wurde noch die sogenannte *Nominalskala* aufgenommen, bei der die Zuweisung von Zahlen zu Merkmalsträgern nur der Klassenbildung dient und Größenvergleiche irrelevant sind.

Skalentyp	erlaubte Transformationen zwischen Quantifizierungen	Beispiele
Nominalskala	$\tau : \mathbb{R} \to \mathbb{R},\ \tau$ injektiv	Geschlecht, Muttersprache
Ordinalskala	$\tau : \mathbb{R} \to \mathbb{R}$, τ streng monoton (wachsend)	Schulnoten, Grad der Ängstlichkeit, Beliebtheitsgrad
Intervallskala	$\tau : \mathbb{R} \to \mathbb{R}$ mit $\tau(x) := u \cdot x + v$; $u \in \mathbb{R}^+, v \in \mathbb{R}$	Körpertemperatur in Grad Celsius und Fahrenheit
Verhältnisskala	$\tau : \mathbb{R} \to \mathbb{R}$ mit $\tau(x) := u \cdot x$; $u \in \mathbb{R}^+$	Alter, Gewicht, Taschengeld
absolute Skala	$\tau : \mathbb{R} \to \mathbb{R}$ mit $\tau(x) := x$	Anzahlen, Wahrscheinlichkeiten

II Persönlichkeitsmerkmale

§1 Ein probabilistisches Modell zur Schülerbewertung

Beispiel 2:
Ein Schüler habe in einem als Klassenarbeit gewerteten Test von 24 Teilaufgaben nur 10 gelöst. Da es bei der Prüfung nur um Routinefertigkeiten ging, war für das Bestehen an sich das 50%-Kriterium angekündigt worden.

Standpunkt 1: 10 Teilaufgaben sind weniger als die Hälfte von 24. Also gibt es bestenfalls die Note „mangelhaft“ für das Testergebnis.

Standpunkt 2: Zufallseinflüsse sind bei der Bearbeitung nicht auszuschließen. Daher sollte für das Bestehen der Prüfung nur verlangt werden, dass für die Anzahl X gelöster Teilaufgaben $E(X) \geq 12$ gilt. Wenn die Hypothese $H : E(X) \geq 12$ bei 10 gelösten Teilaufgaben noch glaubhaft ist, sollte kein „mangelhaft“ gegeben werden.

Bei Standpunkt 1 wird die Bewertung an die beobachtete „Leistung“ gekoppelt, womit sich weitere Diskussionen vermeiden lassen (so ist es in der Praxis an Schulen und Hochschulen weithin üblich). Interessanter sind die nötigen Betrachtungen bei Anhängern des zweiten Standpunkts.

Schreibt man dem Schüler für die Teilaufgaben jeweils im Moment der Inangriffnahme die (unbekannten!) Lösungswahrscheinlichkeiten $p_1, \ldots, p_{24}$ zu, so gilt für die Anzahl X gelöster Aufgaben:

$$E(X) = \sum_{i=1}^{24} p_i \quad \text{und} \quad V(X) = \sum_{i=1}^{24} p_i(1 - p_i)$$

Mit $\overline{p} := \dfrac{1}{24} \sum_{i=1}^{24} p_i$ ergibt sich daraus sowohl die Beziehung $E(X) = 24\overline{p}$ als auch die Abschätzung $V(X) \leq 24\overline{p}(1 - \overline{p})$. Da die größere Varianz für den Schüler zu einer „gutmütigeren“ Beurteilung führt, kann mit der Binomialverteilung mit den Parametern $p = \frac{1}{2}$ und $n = 24$ gearbeitet werden, wenn die Wahrscheinlichkeit von höchstens 10 Erfolgen abgeschätzt werden soll.

Verlangt wird nämlich nach Standpunkt 2 offensichtlich, dass die Hypothese $H : \overline{p} \geq \frac{1}{2}$. gilt. Da für eine mit den Parametern $p = \frac{1}{2}$ und $n = 24$ binomialverteilte Zufallsgröße Y die Wahrscheinlichkeit $P(Y \leq 10)$ noch über 10% liegt, sieht man den beobachteten Misserfolg noch nicht als ungewöhnlich an und gibt daher kein „mangelhaft“.

Wir übertragen die zweite Sichtweise aus dem Beispiel auf den allgemeinen Fall:

Definition 2.1 *Unter einem **probabilistisch** definierten Merkmal auf einer Menge $\mathcal{M}$ von Merkmalsträgern soll eine Menge $\mathcal{T}$ von Situationen, bei der es zu jeder Situation S genau eine erwünschte Reaktion R_S gibt, zusammen mit einer Funktion $p : \mathcal{M} \times \mathcal{T} \to [0; 1]$ verstanden werden.*

Der Funktionswert von p für ein Paar $(m, S) \in \mathcal{M} \times \mathcal{T}$ wird mit $p^{(m)}(S)$ bezeichnet und gilt als Wahrscheinlichkeit, dass der Merkmalsträger m in der Situation S die Reaktion R_S zeigt.

Folgerung:
Gilt bei einem derartigen Merkmal für alle $a, b \in \mathcal{M}$, dass aus der Existenz von $S^* \in \mathcal{T}$ mit $p^{(a)}(S^*) \leq p^{(b)}(S^*)$ für alle $S \in \mathcal{T}$ die Ungleichung $p^{(a)}(S) \leq p^{(b)}(S)$ folgt, so lässt sich das Merkmal mit Ordinalskalenqualität durch folgendes Verfahren quantifizieren:

① Wähle n Situationen $S_1, \ldots, S_n$ aus $\mathcal{T}$.

② Definiere eine Quantifizierung μ durch:

$$\mu : m \mapsto \overline{p}^{(m)} := \frac{1}{n} \sum_{k=1}^{n} p^{(m)}(S_k)$$

Besteht $\mathcal{T}$ nur aus endlich vielen Situationen und wird die Verwendung *aller* dieser Situationen in Schritt ① vorgeschrieben, so ist das mit dieser Einschränkung definierte Merkmal sogar absolut skaliert.

Anmerkung: Bereits für einfache mathematische Fertigkeiten ist die oben angegebene schwache Unabhängigkeitsforderung nicht mehr ganz erfüllt. So fällt zum Beispiel bei den 1×1-Aufgaben auf, dass es Schüler gibt, die unsicher bei der Multiplikation mit 0 sind. Ein solcher Schüler kann bei den restlichen Aufgaben durchaus „besser“ als ein Mitschüler m sein obwohl er m bei der Multiplikation mit 0 unterlegen ist.

§2 Binomialtests

Will man nicht andauernd mit Varianzabschätzungen arbeiten, so kann man zunächst einmal mit einer stark vereinfachten Variante probabilistischer Merkmale arbeiten. Wir schränken diese sofort auf die Vorlage von simultan präsentierten *Aufgaben* ein.

Definition 2.2 *Ein Prüfverfahren mit n dichotomen Items (=Bewertungseinheiten) $a_1, \ldots, a_n$ heißt binomial für eine Probandenpopulation $\mathcal{M}$*

$$:\Longleftrightarrow$$

(1) *Für jeden Probanden $m \in \mathcal{M}$ sind die Items $a_1, \ldots, a_n$ stochastisch unabhängige Zufallsversuche mit den möglichen Ergebnisse 0 und 1.*

(2) *Für die Lösungswahrscheinlichkeiten ${p_1}^{(m)}, \ldots, {p_n}^{(m)}$ jedes Prüflings m gilt* ${p_1}^{(m)} = \ldots = {p_n}^{(m)}$.

Offensichtlich ist jedes durch ein binomiales Prüfverfahren definierte Merkmal absolut skaliert, wenn man die *Lösungswahrscheinlichkeit* eines Probanden für ein Item als Quantifizierung vorschreibt.

Das Binomialmodell ist durch einen χ^2-Test prüfbar[1].

Wie lässt sich die Lösungswahrscheinlichkeit p aus dem Binomialmodell (bzw. eine mittlere Lösungswahrscheinlichkeit in einem allgemeineren Testmodell) in *Noten* umwandeln? Das Problem

[1] Die Mitteilung des Pr¨ufverfahrens erfolgt in einem sp¨ateren Abschnitt.

besteht offensichtlich in der dazu notwendigen Unterteilung des Intervalls $]0;1[$. Man kann trefflich darüber streiten, ob die Teilpunkte

$$0,4;\ 0,55;\ 0,7;\ 0,85$$

der einheitlichen Bewertungsempfehlungen der KMK oder eine Unterteilung des Typs

$$0,5;\ 0,75;\ 0,875;\ 0,9375$$

„angemessen“ sind. Dies liegt daran, dass solche Vorschläge von der Art der geprüften Fertigkeit abhängen.

Die verbindliche KMK-Empfehlung gilt für schriftliche Prüfungsleistungen, die in Form komplexer Aufgaben mit mehrstufiger Bewertung erbracht wurden. Würde man eine solche Aufgabe wie in unserer Modellvorstellung nur mit 0 oder 1 bewerten, so ist wohl eine „Lösungswahrscheinlichkeit“ von 0,4 in den meisten Fällen tatsächlich „gerade noch ausreichend“.

Für Routinefertigkeiten scheint der zweite Vorschlag eher passend, dürfte jedoch z.b. für das Beherrschen des kleinen 1×1 noch nicht extrem genug sein.

Um eine einheitliche Behandlung solcher Fragen zu ermöglichen, betrachten wir erst einmal das Problem, wie man Lösungswahrscheinlichkeiten „multiplikativ“ vergleicht.

Redundanzmodell:

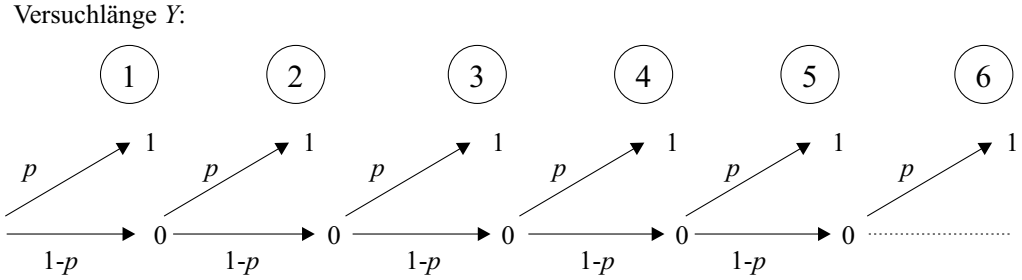
Angenommen, ein Schüler k wird folgendermaßen geprüft:

① k erhält eine zufällig gewählte Aufgabe a_1. Löst er sie, so wird als Versuchsergebnis eine 1 notiert und die Prüfung ist beendet.

② Löst k die Aufgabe a_i (beim ersten Mal ist i gleich 1) nicht, so wird eine Aufgabe a_{i+1} zufällig gewählt. Löst k diese Aufgabe, so wird die Nummer $i+1$ als Versuchsergebnis notiert und die Prüfung ist beendet. Ansonsten wird wieder eine Aufgabe präsentiert.

Je mehr Versuche ein Schüler braucht, desto leistungsschwächer erscheint er in der Prüfung. Da der Vorgang sich bei Fertigkeiten im Sinne der Definition 2.2 durch eine geometrische Verteilung mit der Lösungswahrscheinlichkeit p des Schülers als Parameter beschreiben lässt, kann aus der Versuchsdauer d die Wahrscheinlichkeit durch $\hat{p} := \frac{1}{d}$ geschätzt werden.

Unter dem Binomialmodell folgt aus dem Verfahren auch ein Zugang zum Vergleich von Lösungswahrscheinlichkeiten:



$P(Y=n) = p\,(1-p)^{n-1}$

Fig. 2

Hat nun ein Schüler k bei der ersten Aufgabe dieselbe Chance für den Prüfungserfolg, wie ein Schüler j bei bei insgesamt n Aufgaben, so kann man k „n-mal so sicher“ wie j nennen.

Definition 2.3 *Unter dem Binomialmodell heißt ein Schüler b* ***n-mal so sicher*** *wie ein Schüler a*

$$:\Longleftrightarrow$$

Für die Lösungswahrscheinlichkeiten $p^{(a)}$ und $p^{(b)}$ gilt $p^{(b)} = (1-p^{(a)})^n$.

Folgerung: Man erhält n in der Definition aus der Beziehung $1-p^{(b)} = (1-p^{(a)})^n$ durch Logarithmieren in der Form

$$n = \frac{\ln(1-p^{(b)})}{\ln(1-p^{(a)})}\,.$$

Setzt man die so definierte Funktion auf $]0;1[^{\times 2}$ fort, so erhält man ein Vergleichsmaß für Lösungswahrscheinlichkeiten:

Definition 2.4 *Haben Schüler a und b für eine Aufgabe i die Lösungswahrscheinlichkeiten p_a bzw. p_b mit $p_a, p_b \in]0;1[$, so heißt*

$$\varrho_i(a,b) := \frac{\ln(1-p_b)}{\ln(1-p_a)}$$

das ***Sicherheitsverhältnis*** *von b zu a bezüglich Aufgabe i.*

Offensichtlich gilt $\varrho_i(c,b)\cdot\varrho_i(b,a) = \varrho_i(c,a)$ für alle a, b, c einer Schülerpopulation.

Logarithmiert man das Sicherheitsverhältnis noch einmal, so erhält man ein additives Maß:

Definition 2.5 *Haben Schüler a und b für eine Aufgabe i die Lösungswahrscheinlichkeiten p_a bzw. p_b mit $p_a, p_b \in]0;1[$, so heißt*

$$d_i(a,b) := \ln(\frac{\ln(1-p_b)}{\ln(1-p_a)}) = \ln(-\ln(1-p_b)) - \ln(-\ln(1-p_a))$$

die ***Distanz*** *von b zu a bezüglich Aufgabe i.*

Während das Sicherheitsverhältnis von der Wahl der Logarithmenbasis *unabhängig* ist, würde sich in Definition 2.5 ein anderes Maß ergeben, wenn an Stelle des natürlichen Logarithmus z.b. der dekadische Logarithmus verwendet worden wäre.

In Definition 2.4 ist das Merkmal verhältnisskaliert, da man zur Definition einer Quantifizierung einen festen *Bezugsschüler* bzw. dessen Lösungswahrscheinlichkeit wählen muss.

Im Falle der Definition 2.5 ergibt sich ein Merkmal mit Differenzenskalenqualität, wenn ausschließlich die Logarithmenbasis e verwendet werden darf. Bei freier Wahl der Logarithmenbasis würden sich je zwei verwendete Skalen stets durch eine lineare Transformationsvorschrift des Typs $p \mapsto u \cdot p + v$ mit $u > 0$ ineinander überführen lassen. Das Distanzmerkmal wäre also wenigstens noch *intervallskaliert*.

In dem Beispiel mit der Unterteilung des Intervalls $]0;1[$ durch die Teilpunkte $0,5;\ 0,75;\ 0,875;\ 0,9375,\ \ldots$ ergibt sich für Schüler a, b und c mit $p_a = 0,5$, $p_b = 0,75$ und $p_c = 0,9375$ (bezüglich einer Aufgabe) aus der letzten Definition

$$d(b,a) = \ln 2 \approx 0,693,\ d(c,b) = \ln 2 \approx 0,693,\ d(c,a) = \ln 3 \approx 1,386\ .$$

Nach Wahl einer Lösungswahrscheinlichkeit p_0, die den Nullpunkt auf der Distanzskala festlegt, ergibt sich eine Transformation η des Intervalls $]0;1[$ nach $\mathbb{R}$, deren Vorschrift z.b. für $p_0 = \frac{1}{2}$ durch

$$\eta(p) := \ln(-\ln(1-p)) \ - \ \ln(\ln(2))$$

gegeben ist. Wie der Graf von η zeigt, ist diese Transformation im Teilintervall $]0,3;0,7[$ noch näherungsweise linear. In der Nähe der Intervallenden wird dagegen die *Spreizung* des Intervalls $]0;1[$ sehr deutlich:

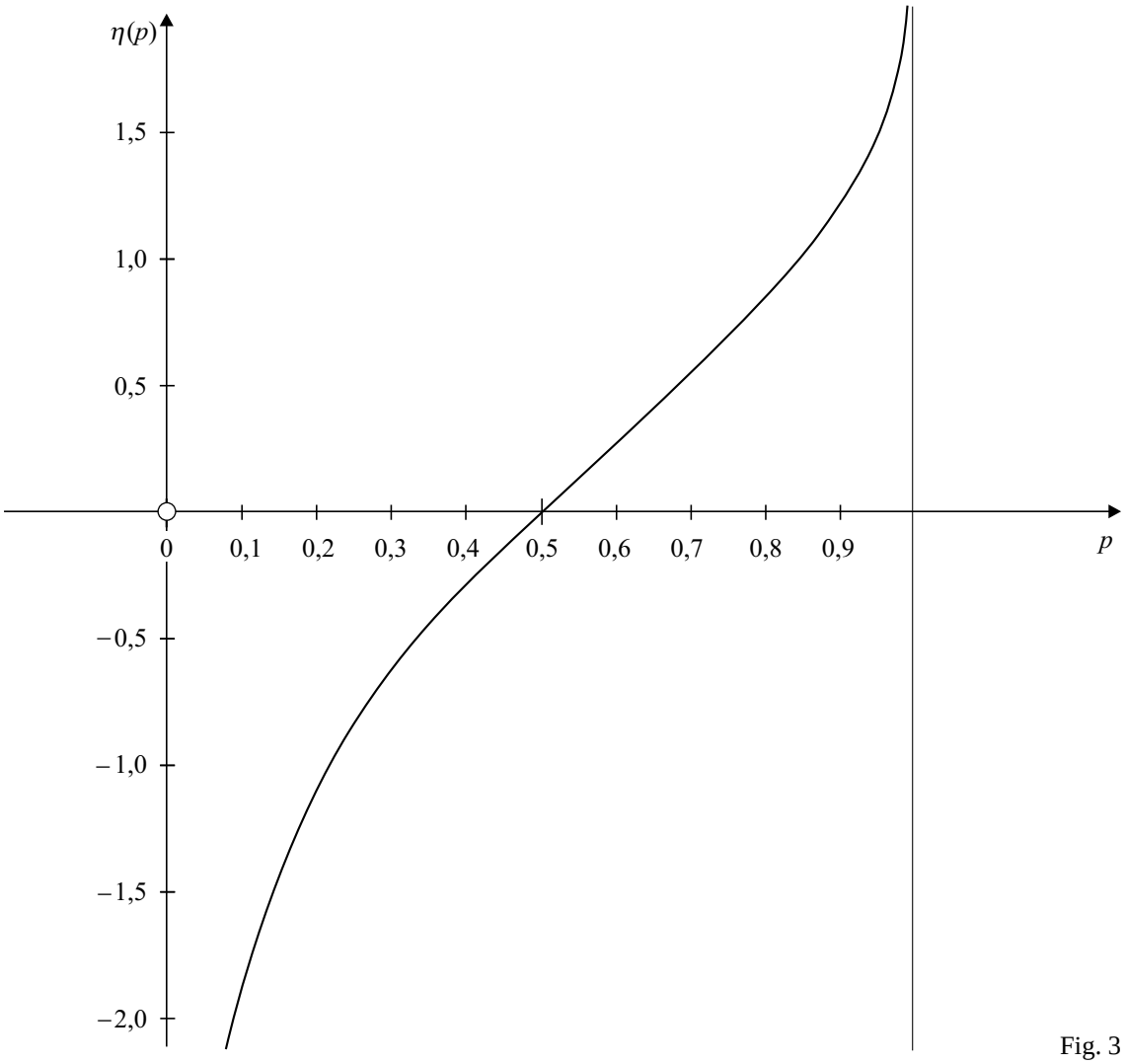


Fig. 3

Der Graf ist zwar „S-förmig“, jedoch nicht punktsymmetrisch bezüglich des Punktes (0,5;0). Er erinnert an die Transformation mit der Vorschrift

$$p \mapsto \frac{2}{\pi} \arcsin(\sqrt{p}) \,,$$

die von K. J. KLAUER 1982 zur Definition von Teilpunkten des Einheitsintervalls mit der Begründung vorgeschlagen worden war, dass damit in allen Leistungsbereichen in etwa die Schätzunsicherheit des transformierten Wertes gleich groß wird.

Da die Schätzunsicherheit wohl kein Maßstab für theoretische Leistungsunterschiede sein kann, diskutieren wir den KLAUERschen Vorschlag nicht weiter und wenden uns der Frage zu, wie man Distanzen linear in die üblichen Noten umrechnen kann. Offensichtlich reicht es dabei, (mittlere) Lösungswahrscheinlichkeiten p_u und p_o mit $p_u < p_o$ festzulegen, bei denen (vor der Rundung auf die üblichen Notenstufen) die Noten 4,49 bzw. 1,49 zu geben sind. Daraus resultiert dann eine Notenfunktion mit der Vorschrift

(2.4.1) $$\nu(p) := b - a \cdot \ln(-\ln(1-p)) \qquad (0 < p < 1)$$
$$\text{mit } b = 4,49 + \frac{3\ln(-\ln(1-p_u))}{\ln(-\ln(1-p_o)) - \ln(-\ln(1-p_u))}$$
$$\text{und } a = \frac{3}{\ln(-\ln(1-p_o)) - \ln(-\ln(1-p_u))}.$$

Beispiel (Routinefertigkeit):

Angenommen, für eine Routinefertigkeit wird ein „mittlerer“ Beherrschungsgrad (das heißt $p_u = 0,5$) als *noch ausreichend* angesehen. Einem Lerner u mit der Lösungswahrscheinlichkeit p_u soll also die Note 4,49 gegeben werden. Soll einem Lerner o mit der Lösungswahrscheinlichkeit $p = p_o := 0,9375$ noch die Note *sehr gut* gegeben werden, so muss die Notenfunktion mit der Vorschrift

$$\nu(p) := 3,697 - 2,164\ln(\ln(1-p))$$

verwendet werden. Würde man an Stelle des natürlichen Logarithmus mit dem dekadischen Logarithmus log arbeiten, so hätte die Vorschrift die Form

$$\nu(p) := 1,892 - 4,893\log(\log(1-p)) .$$

Im Falle $\nu(p) > 6$ ist die Note 6 zu erteilen. Falls sich $\nu(p) < 1$ ergibt, ist die Note 1 zu erteilen. Für die Notenstufen ergeben sich die (gerundeten!) Grenzen in folgender Tabelle:

Note	5,49	4,49	3,49	2,49	1,49
$p_{\min}$	0,35	0,50	0,67	0,83	0,94

Beispiel (KMK-Stufen):

Setzt man $p_u = 0,5$) und $p_o = 0,85$, so ergibt sich die Notenfunktion mit der Vorschrift

$$\nu(p) := 2,954 - 2,286\ln(\ln(1-p)) .$$

Für die Notenstufen ergeben sich gerundet nahezu die Grenzen des KMK-Schlüssels:

Note	(5,49)	4,49	3,49	2,49	1,49
$p_{\min}$	(0,28)	0,50	0,55	0,71	0,85

Die Grenze zwischen den Noten *mangelhaft* und *ungenügend* ist eingeklammert, da sie in den ursprünglichen Richtlinien nicht angesprochen war.

Modellprüfung bei Binomialtests: Wurde einer Schülerpopulation ein Test mit dichotom bewerteten Aufgaben vorgelegt, so lässt sich die Haltbarkeit der *Binomialmodellierung* unter der Unabhängigkeitsannahme von Definition 2.1 und der Zusatzannahme stochastisch unabhängiger Testbearbeitungen aller Schüler mit Hilfe des folgenden Satzes prüfen[2]:

[2] vgl. LIND, D.: Probabilistische Testmodelle. Mannheim 1994. S. 154

Satz 2.1 (Prüfung des Binomialmodells)

Voraussetzung: *Es sei $\mathcal{A}$ ein Test mit n Aufgaben ($n \geq 3$), die alle mit 0 oder 1 bewertet werden. Dieser Test sei für eine Schülerpopulation $\mathcal{M}$ mit N Schülern ein binomiales Prüfverfahren im Sinne der Definition 2.1 und die Testbearbeitungen aller Schüler seien stochastisch unabhängig.*

Für alle Testpunktzahlen t von 1 bis $n-1$ sei die Anzahl der bei Aufgabe i erfolgreichen Schüler mit Testpunktzahl t mit Y_{ti} bezeichnet.

Behauptung: *Unter der Voraussetzung, für $t = 1, \ldots, n-1$ jeweils N_t Schüler mit der Punktzahl t zu beobachten, nähert sich die bedingte Verteilung der Prüfgröße*

$$S_n^2 := \frac{n-1}{n-2} \sum_{t=1}^{n-1} \sum_{i=1}^{n} \frac{(Y_{ti} - \frac{t}{n} N_t)^2}{N_t \frac{t}{n}(1 - \frac{t}{n})}$$

für $N_1, \ldots, N_t \to \infty$ asymptotisch der χ^2-Verteilung mit $n(n-1)$ Freiheitsgraden.

Wir bezeichnen ab jetzt jeden Test $\mathcal{A}$, der die Voraussetzungen von Satz 2.1 in einer Probandenpopulation $\mathcal{M}$ erfüllt, als **Binomialtest** für $\mathcal{M}$.

Bevor wir die Anwendung des Satzes in einem Beispiel zeigen, soll noch der Fall eines Binomialtests dikutiert werden, der nur aus zwei Items besteht. Hier kann bei der Anwendung eines bedingten Tests nur die Gruppe der Probanden herangezogen werden, die lediglich eines der beiden Items erfolgreich bearbeitet haben:

Angenommen, ein Test mit nur 2 Items ist für eine Probandengruppe $\mathcal{M}$ mit N Mitgliedern ein Binomialtest. Dann gilt für die Anzahlen Y_{11} und Y_{12} der Probanden, die nur das erste bzw. das zweite Item erfolgreich bearbeiten: Unter der Voraussetzung, daß N_1 Probanden mit nur einem erfolgreich bearbeiteten Item beobachtet werden (d.h. man setzt $Y_{11} + Y_{12} = N_1$ voraus), hat die Zufallsgröße Y_{11} wegen der für jeden Probanden k geltenden Beziehung

$$\begin{aligned} P(X_1^{(k)} = 1 | X_1^{(k)} + X_2^{(k)} = 1) &= P(X_1^{(k)} = 1 \cap X_2^{(k)} = 0 | X_1^{(k)} + X_2^{(k)} = 1) \\ &= \frac{p_k(1-p_k)}{p_k(1-p_k) + (1-p_k)p_k} = \frac{1}{2} \end{aligned}$$

eine Binomialverteilung mit den Parametern $n = N_1$ und $p = \frac{1}{2}$ als bedingte Verteilung.

Für $N_1 \to \infty$ strebt die Verteilung von

$$\frac{Y_{11} - \frac{N_1}{2}}{\sqrt{\frac{1}{4}N_1}}$$

nach dem zentralen Grenzwertsatz gegen die Standardnormalverteilung. Damit strebt die Verteilung der Variablen

$$\frac{(Y_{11} - \frac{N_1}{2})^2}{\frac{1}{4}N_1} \quad \left(= \frac{(Y_{11} - Y_{12})^2}{N_1} \right) \tag{1}$$

für $N_1 \to \infty$ gegen die χ^2-Verteilung mit einem Freiheitsgrad.

Trägt man die Realisierungen x von Y_{11} und y von Y_{12} in eine Vierfeldertafel des Typs

	$1e$	$1ne$	
$2e$	u	y	N_b
$2nex$	v	$N - N_b$	
	N_a	$N - N_a$	N

mit den Abkürzungen:

$1e$: *bei Item 1 erfolgreich*
$1ne$: *bei Item 1 nicht erfolgreich*
$2e$: *bei Item 2 erfolgreich*
$2ne$: *bei Item 2 nicht erfolgreich*

ein, so gilt $N_1 = x + y$ und die Realisierung der Variablen in (1) läßt sich nach der bekannten Formel

$$s^2_{x;y} := \frac{(x-y)^2}{x+y} \tag{2}$$

berechnen. Falls $x+y$ größer als 40 ist, kann mit der χ^2_1-Verteilung als Näherung für die Verteilung der Prüfgröße gearbeitet werden. Ansonsten muß der Zähler des Quotienten in (2) durch $(|x-y|-1)^2$ ersetzt werden. Dies entspricht einer besseren Anpassung der Binomialverteilung von Y_{11} an die Normalverteilung. Mit dieser Korrektur darf die χ^2-Näherung bis herunter zu $x + y = 10$ verwendet werden. Die Anwendung des resultierenden Prüfverfahrens ist unter dem Namen χ^2-Test von MC NEMAR bekannt. Es kann auch verwendet werden, wenn lediglich zwei ausgewählte Items aus einem größeren Test verglichen werden sollen (eine übliche Prüfmethode in der klassischen Testtheorie). Wegen der dabei verschenkten Informationen sollte jedoch in solchen Fällen mit folgender Methode gearbeitet werden:

Bildet man bei einem Binomialtest $calA$ mit n dichotomen Items ($n > 2$) die Klassen der Probanden, die 1, 2, ..., $n-1$ Items gelöst (= erfolgreich bearbeitet) haben, so lässt sich für die Löseranzahlen Y_{tr} und Y_{ts} zweier verschiedener Items r und s in der Testwertkategorie t bei fest vorausgesetzten Kategoriengrößen $N_1, \ldots, N_{n-1}$ folgender Satz beweisen:

Satz 2.2 *Wird bei der Vorlage eines mindestens 3 Items enthaltenden Binomialtests in einer Probandenpopulation für $1 < t < n$ mit Y_{ti} die Anzahl der bezüglich Item i erfolgreichen Probanden mit dem Testwert t bezeichnet, so gilt für je zwei verschiedene Items s und t des Tests:*

Unter der Voraussetzung, für $t = 1, \ldots, n-1$ jeweils N_t Probanden mit Testwert t zu beobachten, ist die bedingte Verteilung der Prüfgröße

$$S^2_{rs} := \frac{n-1}{n-2} \sum_{t=1}^{n-1} \frac{(Y_{tr} - \frac{t}{n}N_t)^2 + (Y_{ts} - \frac{t}{n})^2}{N_t \frac{t}{n}(1-\frac{t}{n})} \tag{3}$$

für $N_1, \ldots, N_{n-1} \to \infty$ asymptotisch χ^2-verteilt mit $2 \cdot (n-1)$ Freiheitsgraden.

Die Prüfung der Binomialtesthypothese für eine Population $\mathcal{M}$ durch das χ^2-Verfahren ist natürlich insofern problematisch, als man dabei mit der Nullhypothese

H_0:= „Für jeden Probanden j aus $\mathcal{M}$ gilt $p_1^{(j)} = \ldots = p_n^{(j)}$.“

arbeitet. Da H_0 erst beim Überschreiten der zur gewählten Wahrscheinlichkeit α für den Fehler erster Art tabellierten Grenze $s_d(\alpha)$ verworfen wird, entspricht dieses Vorgehen der Philosophie, ein Modell solange beizubehalten, wie nicht massive Unverträglichkeiten mit Beobachtungsdaten auftreten. So arbeitet man bei kleineren Probandenzahlen trotz deutlicher Unterschiede zwischen Lösungshäufigkeiten mit der Binomialtestannahme, wenn diese noch nicht extrem genug sind.

Bei sehr großen Probandenzahlen tritt dieses Problem nicht auf. Dafür nimmt jetzt die Prüfgröße schon bei relativ geringen Unterschieden zwischen Lösungsquoten einen signifikanten Wert an. Hier muss eher überlegt werden, ob man die vermuteten tatsächliche Unterschiede im Schwierigkeitsverlauf der Items für so gering hält, daß man trotzdem mit dem einfacheren Binomialtestmodell arbeitet.

Beispiel

Wir demonstrieren für jeden der diskutierten Fälle die Prüfung der Binomialtesthypothese:

Der Fall n = 2: Angenommen, die Testauswertung eines 2-Item-Tests ergab die Tabelle

	$1e$	$1ne$	
$2e$	4	12	16
$2ne$	10	4	14
	14	16	30

mit der korrigierten Prüfgröße
$s^2_{10;12} = \frac{(|10-12|-1)^2}{23} = \frac{1}{23}$.

Auch bei einer sehr skeptischen Einstellung scheint hier zunächst die Binomialtestannahme haltbar zu sein, da noch nicht einmal der kritische Wert $s_1(0,50) \approx 0,455$ überschritten ist. Trotzdem ist das Überwiegen *verschiedener* Bearbeitungen verdächtig. Es könnte durchaus der Fall vorliegen, daß die Items nur im Populationsmittel ihrer Schwierigkeiten übereinstimmen und Bearbeitungsfertigkeiten erfordern, die bei vielen Probanden nur alternativ verfügbar sind.

Vergleich zweier Items aus einem größeren Test: Wir nehmen an, daß die gerade betrachteten beiden Items aus einem Test mit zwei weiteren Items 3 und 4 stammen, für den die vollständige Zusammenfassung der Testresultate folgendermaßen aussieht:

	Testwert t			
	1	2	3	
Y_1	8	6	0	14
Y_2	0	6	10	16
Y_3	0	5	10	15
Y_4	1	5	10	16
N_t	9	11	10	30
$\frac{t}{4}N_t$	2,25	5,5	7,5	
$N_t\frac{t}{4}(1-\frac{t}{4})$	1,688	2,750	1,875	

(Der Wert 1,688 ist auf drei Nachkommastellen gerundet.)

Die Prüfgröße S^2_{12} für die Items 1 und 2 hat den Wert

$$s^2_{12} = 1,5\left(\frac{(8-2,25)^2+(0-2,25)^2}{1,688}+\frac{(6-5,5)^2+(6-5,5)^2}{2,750} + \frac{(0-7,5)^2+(10-7,5)^2}{1,875}\right) \approx 84,2\,.$$

Falls der Test ein Binomialtest ist, kann S^2_{12} als näherungsweise χ^2_6-verteilt angesehen werden. Damit gilt

$$P(S^2_{12} \geq 84,2) < 0,01\,,$$

da eine nach χ^2 mit 6 Freiheitsgraden verteilte Zufallsgröße bereits Werte von etwa 17 mit weniger als der Wahrscheinlichkeit 0,01 überschreitet (eine übliche Sprechweise für diesen Sachverhalt ist, der beobachtete Wert sei auf dem 1%-Niveau *signifikant*). Damit läßt sich die Modellannahme gleicher Itemschwierigkeiten für die Items 1 und 2 mit einer niedrigeren Irrtumswahrscheinlichkeit als 0,01 ablehnen.

Prüfung eines Gesamttests mit mehr als 2 Items: Angenommen, die Vorlage eines 4-Item-Tests lieferte folgende Tabelle:

	Testwert t			
	1	2	3	
Y_1	3	10	35	48
Y_2	5	15	30	50
Y_3	6	15	30	51
Y_4	4	20	25	49
N_t	18	30	40	88
$\frac{t}{4}N_t$	$4,5$	15	30	
$N_t\frac{t}{4}(1-\frac{t}{4})$	$3,375$	$7,5$	$7,5$	

Die Prüfgröße S_4^2 hat den Wert

$$s_4^2 \;=\; 1,5\left(\frac{1}{3,375}(1,5^2+0,5^2+1,5^2+0,5^2)+\frac{1}{7,5}(5^2+0^2+0^2+5^2)\right.$$
$$\left.+\;\frac{1}{7,5}(5^2+0^2+0^2+5^2)\right)\;=\;22,22\,.$$

Bei 9 Freiheitsgraden beträgt die Überschreitungswahrscheinlichkeit für diesen Wert weniger als 0,01 (es gilt $s_9^2(0,01)\approx 21,7$). Daher würde man auch in diesem Fall die Binomialtesthypothese ablehnen.

Abschließend soll noch eine Empfehlung zur Handhabung des χ^2-Verfahrens bei Tests mit mindestens 3 Items angegeben werden (sie folgt aus Betrachtungen zur Anpassung von Binomialverteilungen an Normalverteilungen):

Faustregel: In jeder Testwertkategorie t sollte die Anzahl N_t der Probanden größer als $\frac{1}{4(\frac{t}{n})^2(1-\frac{t}{n})^2}$ sein. Ist dies nicht der Fall, so müssen Kategorien zusammengefaßt werden. Falls man dabei die entsprechenden Erwartungswerte $\frac{t}{n}N_t$ addiert und auch die Summe der Varianzen $N_t\frac{t}{n}(1-\frac{t}{n})$ bildet, resultiert bei m verbleibenden Kategorien eine Prüfgröße, die näherungsweise nach χ^2 mit $m(n-1)$ Freiheitsgraden verteilt ist.

Wird diese Regel z.b. auf den zuletzt betrachteten Fall des vorigen Beispiels angewendet, so muß die niedrigste Testwertkategorie mit einer der beiden anderen zusammengefaßt werden.

Werden die Klassen mit $t = 1$ und $t = 2$ vereinigt, so ergibt sich die Tabelle:

	Testwert t		
	1 oder 2	3	
Y_1	13	35	48
Y_2	20	30	50
Y_3	21	30	51
Y_4	24	25	49
N_t	48	40	88
$E(Y)$	$19,5$	30	
$V(Y)$	$10,875$	$7,5$	

Jetzt hat S_4^2 den Wert

$$s_4^2 \;=\; 1,5\left(\frac{1}{10,875}(6,5^2+0,5^2+1,5^2+4,5^2) + \frac{1}{7,5}(5^2+0^2+0^2+5^2)\right) \approx 18,97\,.$$

Bei 6 Freiheitsgraden ist dies immer noch auf dem 1%-Niveau signifi kant. Die Ablehnung der Binomialtesthypothese im Beispiel resultierte also nicht aus einer Unterschätzung der Irrtumswahrscheinlichkeit (bei starken Abweichungen von der verwendeten Grenzverteilung ist dieser Effekt möglich).

§3 Latent-Trait-Modelle mit dichotomen Items

Haben die Aufgaben eines Tests $\mathcal{A}$ nicht den gleichen *Komplexitätsgrad*, so ist die Annahme gleicher Lösungswahrscheinlichkeiten unrealistisch. Es gibt auch für diesen Fall Testmodelle, die mehr oder weniger hauptsächlich wegen ihrer *schätztheoretischen* Eigenschaften verwendet werden. Um eine Argumentationsbasis zur Einführung solcher Testmodelle zu bekommen, halten wir zunächst eine der Forderungen aus dem letzten Paragrafen in einer Definition fest:

Definition 2.6 *Ein n-Item-Test Γ_n mit dichotomer Bewertung heiße genau dann ein **verallgemeinerter Binomialtest** über einer Probandenpopulation $\mathcal{M}$, wenn gilt:*

(1) *Für jeden Probanden $k \in \mathcal{M}$ sind seine Antwortvariablen $X_1^{(k)}, \ldots, X_n^{(k)}$ stochastisch unabhängig.*

(2) *Die Antwortvektoren $(X_1^{(1)}, \ldots, X_n^{(1)}),\ \ldots\ , (X_1^{(N)}, \ldots, X_n^{(N)})$ aller Probanden sind stochastisch unabhängig.*

Offensichtlich sind diese Forderungen nur dann zur Definition eines *Fähigkeitsmerkmals* geeignet, wenn auch noch eine bereits im Anschluss von Definition 2.1 erwähnte Zusatzbedingung verlangt wird:

Definition 2.7 *Ein n-Aufgaben-Test Γ_n heiße genau dann ein **Fähigkeitstest** über einer Probandenpopulation $\mathcal{M}$, wenn er für diese Population ein verallgemeinerter Binomialtest ist und für alle $a, b \in \mathcal{M}$ gilt:*

(M) *Genau dann gibt es ein Item $i \in \Gamma_n$ mit $p_i^{(a)} \leq p_i^{(b)}$, wenn für alle Items $j \in \Gamma_n$ die Ungleichung $p_j^{(a)} \leq p_j^{(b)}$ erfüllt ist.*

In diese Klasse von Testmodellen fallen erst einmal alle Binomialtests. Hier kann man die Lösungswahrscheinlichkeit eines Probanden zum Fähigkeitsparameter erklären. Wenn die Items von Γ_n verschieden schwer sind, bietet sich die mittlere Lösungswahrscheinlichkeit $\overline{p}_k$ als Fähigkeitsparameter an. Ihre Verwendung führt zur sogenannten *True-Score-Testtheorie* und hat den Vorteil, dass der Anteil gelöster Aufgaben eines Probanden k ein erwartungstreuer Schätzer für $\overline{p}_k$ ist.

Für die Analyse von Teststrukturen und Aufgabenmerkmalen sind jedoch die Annahmen aus Definition 2.7 immer noch zu schwach.

Wir betrachten daher eine erste Beispielklasse von Tests, für sich mehr über den Schwierigkeitsverlauf von Testitems sagen lässt.

§4 Potenzmodelle

Beispiel (Kettenmodell):

Angenommen, ein Test besteht aus 5 Items, bei denen jeweils eine Routinetechnik mehrfach anzuwenden ist, um zur Lösung zu gelangen. Jedes Item i besitzt eine individuelle Zahl m_i von Bearbeitungsstufen und wird am Schluss der Bearbeitung lediglich nach den Kategorien *richtig* (1) und *falsch* (0) bewertet. Wir nehmen an, dass $m_1 = 1$, $m_2 = 2$, $m_3 = 3$, $m_4 = 4$ und $m_5 = 5$ gilt.

Es sei p die Wahrscheinlichkeit eines Probanden, bei einmaligem Einsatz der verlangten Technik keine Fehler zu machen. Dann hat die Wahrscheinlichkeit p_i des Probanden, ein Item i mit m_i-maliger Verwendung der Technik richtig zu bearbeiten, den Wert p^{m_i}:

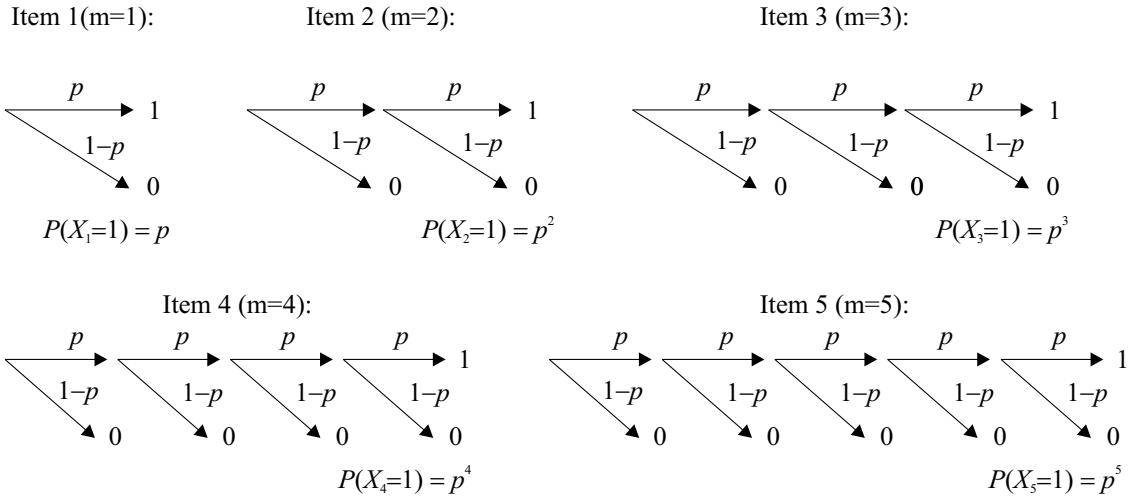


Fig. 4

Stellt man die Lösungswahrscheinlichkeiten der Items als Funktion von p dar, so erhält man folgende Grafen:

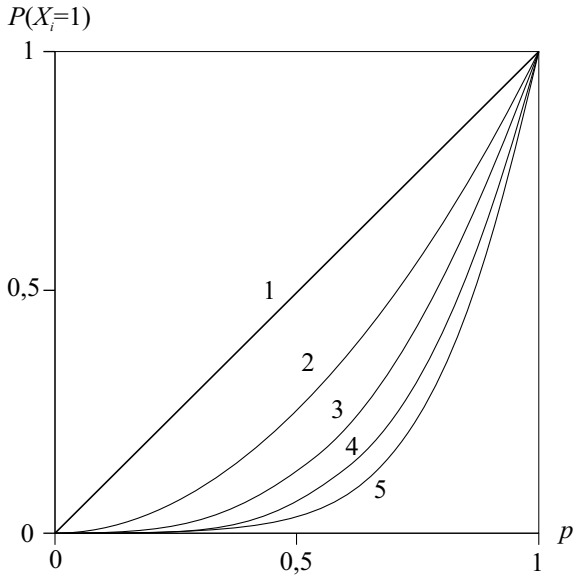


Fig. 5

Die Lösungswahrscheinlichkeit lässt sich für ein Item i mit der Vielfachheit m_i für $0 < p < 1$ auch in der Form

$$P(X_i = 1) = \exp\{-\exp(\ln m_i + \ln(-\ln p))\}$$

schreiben.
Bezeichnet man $\delta_i := \ln m_i$ als *Itemschwierigkeit* und $\theta := -\ln(-\ln p)$ als *Probandenfähigkeit*, so lassen sich die Lösungswahrscheinlichkeiten in der Form

$$P(X_i = 1) = \exp\{-\exp(\delta_i - \theta)\}$$

mit $-\infty < \theta < \infty$ schreiben.
Stellt man die Lösungswahrscheinlichkeiten in Abhängigkeit von θ dar, so gehen die Grafen für die Items 2, 3, 4 und 5 durch Paralellverschiebung längs der θ-Achse aus dem Grafen für Item 1 hervor:

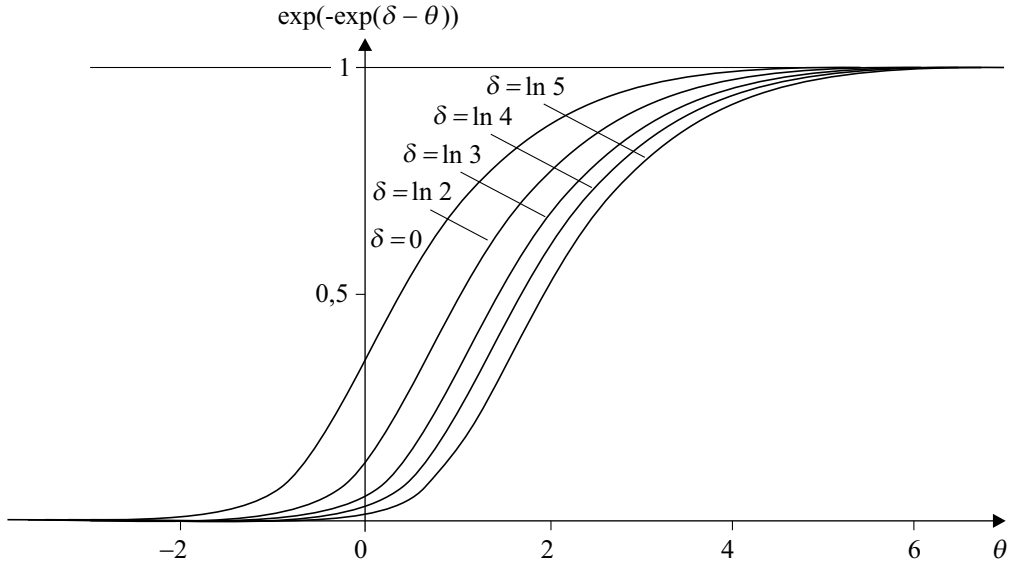


Fig. 6

Dies liegt daran, dass die Lösungswahrscheinlichkeit eines Items i bei dieser Parametrisierung nur von der *Differenz* $\delta_i - \theta$ abhängt. Die Kurven sind zwar alle S-förmig, sie sind jedoch nicht punktsymmetrisch zu ihrem jeweiligen Wendepunkt.

Wir verallgemeinern das Beispiel auf beliebige Itemzahlen und lassen die Forderung fallen, dass die Schwierigkeitparameter aller Items bis auf eine gemeinsame additive Konstante Logarithmen natürlicher Zahlen sind:

Definition 2.8 *Ein verallgemeinerter Binomialtest mit Γ_n mit n Items heißt genau dann ein **Potenzmodell** über einer Probandenpopulation $\mathcal{M}$, wenn sich jedem Item $i \in \Gamma_n$ eine reelle Zahl δ_i und jedem Probanden $k \in \mathcal{M}$ eine reelle Zahl θ_k so zuordnen lassen, dass für alle Items $i \in \Gamma_n$ und alle Probanden $k \in \mathcal{M}$ gilt:*

$$P(X_i = 1) \;=\; \exp\{-\exp(\delta_i - \theta_k)\}\;.$$

Der Name für ein solches Modell ist damit zu rechtfertigen, dass sich die Lösungswahrscheinlichkeit eines Items i für einen Probanden k mit den Setzungen $p_k := \exp\{-\frac{1}{\exp(\theta_k)}\}$ und $h_i := \exp(\delta_i)$ in der Form

$$P(X_i = 1) \;=\; p_k^{h_i}$$

schreiben lässt. Dabei liegt p_k stets zwischen 0 und 1 und h_i ist ein positiver reeller Exponent. Sieht man δ als fest gewählten Schwierigkeitsparameter an, so nennt man die auf $\mathbb{R}$ definierte Funktion f_δ mit der Vorschrift

$$f_\delta(\theta) := \exp\{-\exp(\delta - \theta)\}$$

die ***Itemcharakteristik*** eines Items mit ***Schwierigkeitsparameter*** δ. Die Wendepunkte der zugehörigen Funktionsgrafen liegen bei $(\delta; \frac{1}{e}$.

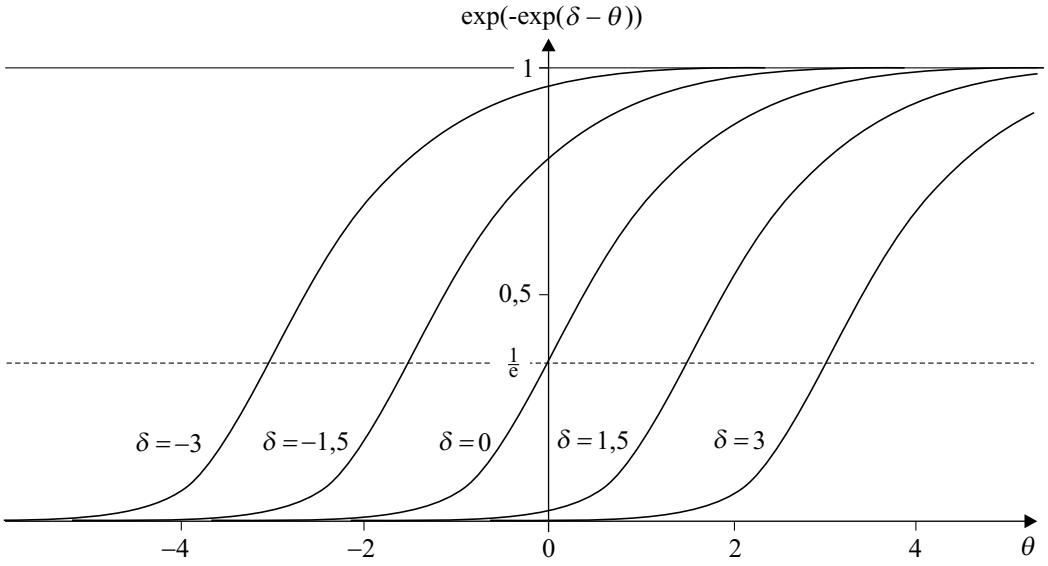


Fig. 7

Betrachtet man die Funktionsterme für große Werte von θ, so erhält man im Falle $\theta >> \delta$ folgende Näherung:

$$\begin{aligned} f_\delta(\theta) &= \exp\{-\exp(\delta - \theta)\} \\ &= (\exp\{\exp(\delta - \theta)\})^{-1} \\ &\approx (1 + \exp(\delta - \theta))^{-1} \end{aligned}$$

Verwendet man den Näherungsterm als Funktionsvorschrift für den gesamten Bereich von Fähigkeitswerten, so erhält man ein bekanntes Testmodell, das wegen seiner guten statistischen Handhabbarkeit weit verbreitet ist. Da die Funktion $x \mapsto (1 + \exp(-x))^{-1}$ den Namen ***logistische Funk-***

tion hat, nennt man es das *zweikategorielle logistische Testmodell*. Die Grafen der obigen Itemcharakteristiken gehen in diesem Modell in punktsymmetrische Grafen über, deren Wendepunkte jeweils an der Stelle $(\delta; 0,5)$ liegen.

III Logistische Testmodelle

In diesem Abschnitt sollen die wichtigsten Testmodelle vorgestellt werden, bei denen die „äußere Hüllfunktion“ bei der Definition von Itemcharakteristiken die logistische Funktion ist. Es kommen zwar auch andere streng monotone Funktionen von $\mathbb{R}$ nach $]0;1[$ (wie z.B. $\mathrm{x} \mapsto \exp(-\exp(-\mathrm{x}))$) in Frage, diese führen jedoch auf Modelle mit ungünstigen Eigenschaften bei der *Parameterschätzung*.

§1 Das Raschmodell für dichotome Items

Wir übernehmen die Näherung aus dem letzten Paragrafen von Abschnitt II und definieren:

Definition 3.1 *Ein Test Γ_n mit n Items heißt genau dann ein **(zweikategorielles) Raschmodell** über einer Probandenpopulation $\mathcal{M}$, wenn*

(1) *er ein verallgemeinerter Binomialtest über $\mathcal{M}$ ist,*

(2) *sich jedem Item $i \in \Gamma_n$ eine reelle Zahl δ_i und jedem Probanden $k \in \mathcal{M}$ eine relle Zahl θ_k so zuweisen lässt, so dass für alle Items $i \in \Gamma_n$ und alle Probanden $k \in \mathcal{M}$ gilt:*

$$P(X_i^{(k)} = 1) \;=\; (\; 1 + \exp(\delta_i - \theta_k)\;)^{-1}\,.$$

*Man nennt in diesem Fall Γ_n über $\mathcal{M}$ **Rasch-skalierbar**.*

Betrachtet man die Itemcharakteristiken dieses Testmodells, so wird sofort deren größere Symmetrie im Vergleich zum Potenzmodell deutlich:

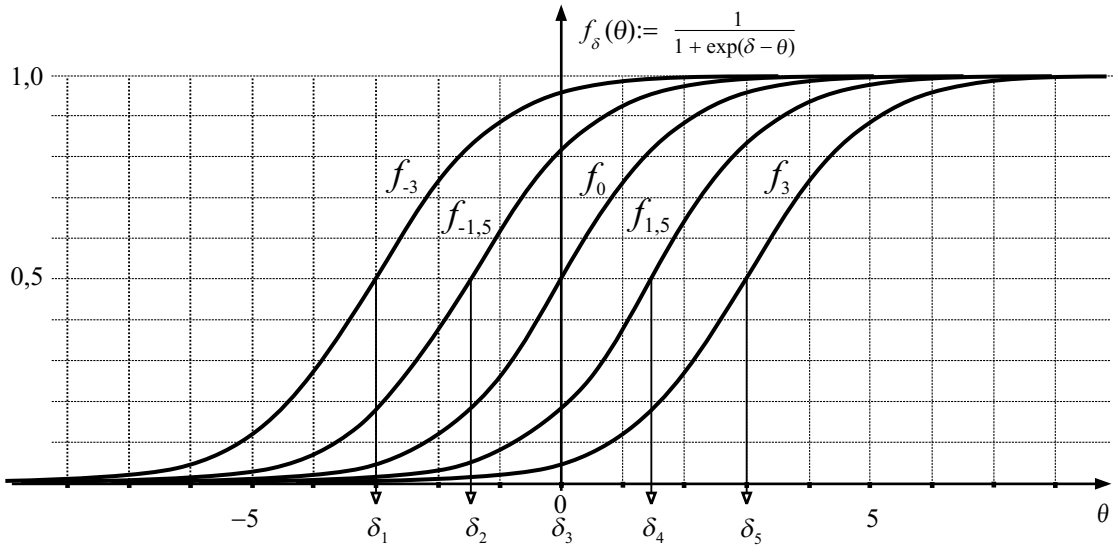


Fig. 8

Die wichtigste Eigenschaft dieses Modells ist die Möglichkeit, die Schwierigkeitsparameter der Items unabhängig von den Probandenfähigkeiten schätzen zu können. Eine zentrale Rolle spielt dabei die Betrachtung von Probanden mit gleicher *Testpunktzahl* t (=Anzahl gelöster Items). Es gilt nämlich:

Satz 3.1

Vor.: *Es sei Γ_n ein Test, der über einer Probandenpopulation $\mathcal{M}$ Rasch-skalierbar ist.*

Beh.: *Für jede Testpunktzahl t mit $0 < t < n$ und jeden Antwortvektor $\vec{x}$ mit $\sum_{s=1}^{n} x_s = t$ ist die bedingte Wahrscheinlichkeit $P_\theta(\vec{X} = \vec{x} \mid \sum_{s=1}^{n} X_s = t)$ eines Probanden aus $\mathcal{M}$ für diesen Antwortsvektor unabhängig vom Fähigkeitsparameter θ.*

Beweis:

1. Gegeben sei ein Proband aus $k \in \mathcal{M}$ mit dem Fähigkeitswert θ. Wir lassen den Probandenindex im Folgenden weg, um die Notation übersichtlicher zu halten, und schreiben zunächst den Term für die Antwort x_i von k auf das Items i in der Form

$$P_\theta(X_i = x_i) = \frac{\exp(x_i(\theta - \delta_i))}{1 + \exp(\theta - \delta_i)}\,.$$

   Man kann sich leicht davon überzeugen, dass dies sowohl für $x_i = 1$ als auch für $x_i = 0$ einen Term liefert, der zur ursprünglichen Definition der Lösungswahrscheinlichkeit bzw. der daraus abgeleiteten Versagenswahrscheinlichkeit äquivalent ist. Es sei t eine Testpunktzahl mit $0 < t < n$.

2. Dann gilt für jeden Antwortvektor $\vec{x}$ mit $\sum_{s=1}^{n} x_s = t$:

$$\begin{aligned} P_\theta(\vec{X} = \vec{x}) &= \prod_{i=1}^{n} \frac{\exp(x_i(\theta - \delta_i))}{1 + \exp(\theta - \delta_i)} \\ &= \exp(\sum_{j=1}^{n} x_j \cdot \theta) \prod_{i=1}^{n} \frac{\exp(-x_i\delta_i)}{1 + \exp(\theta - \delta_i)} \\ &= \exp(t \cdot \theta) \prod_{i=1}^{n} \frac{\exp(-x_i\delta_i)}{1 + \exp(\theta - \delta_i)}\,. \end{aligned}$$

3. Durch Summation über alle möglichen Antwortvektoren erhält man

$$\begin{aligned} P_\theta(\sum_{s=1}^{n} X_s = t) &= \exp(t \cdot \theta) \sum_{y_1+\ldots+y_n=t} \prod_{i=1}^{n} \frac{\exp(-y_i\delta_i)}{1 + \exp(\theta - \delta_i)} \\ &= \exp(t \cdot \theta) \prod_{i=1}^{n} \frac{1}{1 + \exp(\theta - \delta_i)} \cdot \sum_{y_1+\ldots+y_n=t} \prod_{i=1}^{n} \exp(-y_i\delta_i) \end{aligned}$$

4. Durch Quotientenbildung erhalten wir

$$P_\theta(\vec{X} = \vec{x} \mid \sum_{s=1}^{n} X_s = t) = \frac{\prod_{i=1}^{n} \exp(-x_i\delta_i)}{\sum_{y_1+\ldots+y_n=t} \prod_{i=1}^{n} \exp(-y_i\delta_i)}\,.$$

q.e.d.

Satz 1 bildet die Grundlage bedingter Itemparameterschätzungen. Dabei geht man von der bedingten Likelihood der beobachteten Datenmatrix aus, die sich als Produkt der im obigen Beweis hergeleiteten Terme für $P_\theta(\,\vec{X} = \vec{x}\,|\,\sum_{s=1}^{n} X_s \,=\, t)$ schreiben lässt. Wir führen zur Beschreibung des Verfahrens folgene Abkürzungen ein:

$$\begin{aligned} R_t &:= \{\vec{y} \in \{0;1\}^{\times n} \mid \sum_{s=1}^{n} y_s \,=\, t\} \qquad \text{für } t = 0, \ldots, n \\ \gamma_n(t;\vec{\delta}) &:= \sum_{\vec{y} \in R_t} \exp(-\sum_{s=1}^{n} y_s \delta_s) \\ p(\vec{x}|t;\vec{\delta}) &:= \frac{\exp(-\sum\limits_{j=1}^{n} x_j \delta_j)}{\gamma_n(t;\vec{\delta})} \qquad \text{für } t = 0, \ldots, n \text{ und } \vec{x} \in R_t\,. \end{aligned}$$

Es seien $\vec{x}_1, \ldots, \vec{x}_N$ die beobachteten Antwortvektoren mit den zugehörigen Testpunktzahlen $t_1, \ldots, t_N$. Die daraus gebildete *Datenmatrix* sei mit X bezeichnet. Dann hat die bedingte *Likelihood* des Datensatzes die Form

$$L^*(\mathrm{X};\vec{\delta}) := \prod_{k=1}^{N} p(\vec{x}_k|t_k;\vec{\delta})$$

und es ergeben sich durch partielles Differenzieren nach δ_1 bis δ_n und Nullsetzen der Ableitungen folgende **notwendige** Schätzgleichungen:

$$A_j \;=\; \sum_{t=1}^{n-1} N_t \exp -\delta_j \frac{\gamma_{n-1}(t-1;\vec{\delta}_{(j)})}{\gamma_n(t;\vec{\delta})} \qquad \text{für } j = 1, \ldots, n$$

mit

$$\begin{aligned} A_j &:= \text{Anzahl der Probanden, die Item } j \text{ gelöst haben,} \\ N_t &:= \text{Anzahl der Probanden mit Testpunktzahl } t, \\ \vec{\delta}_{(j)} &:= (\delta_1, \ldots, \delta_{j-1}, \delta_{j+1}, \ldots, \delta_n)\,. \end{aligned}$$

Dieses Gleichungssystem ist für große Probandenmengen im Allgemeinen lösbar und liefert dann auch ein globales Maximum der Likelihood. Wenn ein Item von allen Probanden in gleicher Weise beantwortet wurde, muss es allerdings vorab bei dem bedingten Schätzverfahren aus der Datenmatrix herausgenommen werden.

Nur wenn man die Probanden und Items so nummerieren kann, dass die Datenmatrix die Form

		Items					
		1	…	k	$k+1$	…	n
Probanden	1	Daten			0	…	0
	⋮				⋮		⋮
	J				0		0
	J+1	1	…	1	Daten		
	⋮	⋮		⋮			
	N	1	…	1			

annimmt, ist das Gleichungssystem unlösbar. Ein Sonderfall dieses Typs liegt vor, wenn ein Test nur Items mit der sogenannten GUTTMANN-Charak"teristik enthält. In diesem Fall besitzt jedes Item eine Schwelle σ_k, unterhalb der es überhaupt nicht und ab der es sicher lösbar ist. Hier lassen sich die Items nach ihren „Sprunggrenzen" ordnen und ein Proband, der ein Item mit höherer Schwelle richtig bearbeitet, löst auch alle Items mit niedrigeren Schwellen. Versagt er bei einem Item, so auch bei allen Items solchen mit noch höheren Schwellen.

Unter der Annahme, dass ein dichotomes RASCH-Modell unabhängige Testvariablen $\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_n$ beschreibt und mindestens eine Anzahl N_t mit $1 \leq t \leq n-1$ mit wachsendem Stichprobenumfang beliebig gross wird, ist die bedingte Parameterschätzung *konsistent*[3]. Von ANDERSEN stammt auch ein Likelihood-Quotienten-Test zur Modellprüfung, der auf dem Vergleich bedingter Itemparameterschätzungen beruht.

Viele Schätzprogramme gehen allerdings einen anderen Weg, der unter dem Namen E-M-Algorithmus bekannt ist. Dabei wird unterstellt, dass die Personenparameter einem Verteilungstyp angehören, dessen Parameter (z.b. bei der Normalverteilung: Erwartungswert μ und Standardabweichung σ) zu schätzen sind. Dabei wird der systematische Schätzfehler vermieden, den man bei einer simultanen Schätzung von Item- und Probandenparametern in Kauf nehmen muss.

Wir stellen die problematische Simultanschätzung trotzdem vor, da ihr zweiter Schritt durchaus Verwendung findet, wenn man mit Hilfe fester Itemparameter aus dem Antwortvektor eines Probanden dessen Fähigkeitsparameter nach der ML-Methode schätzen will:

Start: Alle Itemparameter werden auf einen Startwert z.b. einheitlich 0 gesetzt. Der Schrittzähler i wird auf 0 gesetzt. Es wird eine positive Abbruchschranke ε festgesetzt.

θ-Step: Mit den aus dem Vorgängerschritt resultierenden Itemparametern $\delta_1^{(i)}, \ldots, \delta_n^{(i)}$ werden Probandenparameter $\theta_1^{(i)}, \ldots, \theta_N^{(i)}$ aus den Gleichungen

$$\sum_{s=1}^{n} \frac{1}{1+\exp(\delta_s^{(i)} - \theta_k^{(i)})} = t_k \qquad \text{für } k = 1, \ldots, N$$

bestimmt. Diese werden anschließend durch Subtraktion einer passenden Konstante so normiert, dass ihr Mittelwert 0 beträgt.

δ-Step: Der Schrittzähler i wird um 1 erhöht und mit den aus dem Vorgängerschritt resultierenden Probandenparametern werden Itemparameter $\delta_1^{(i)}, \ldots, \delta_n^{(i)}$ aus den Gleichungen

$$A_j = \sum_{k=1}^{N} \frac{1}{1+\exp(\delta_j^{(i)} - \theta_k^{(i-1)})} = A_j \qquad \text{für } j = 1, \ldots, n$$

bestimmt. Falls die Ändereungen sowohl der Personenparameter als auch der Itemparameter kleiner als die vorher gewählte Abbruchschranke ε sind, endet der Algorithmus. Ansonsten wird zum δ-Step verzweigt.

[3] Dies hat E.B. Andersen in Psychometrika *48,* 1973, S. 123ff gezeigt.

In diesen Schätzgleichungen wird jeweils ein Parametersatz so bestimmt, dass beobachtete Anzahlen gleich einem Erwartungswert sind. Dass dies im Falle des Rasch-Modells aus der ML-Bedingung folgt, soll abschließend begründet werden:

Die (nichtbedingte!) Likelihood der Datenmatrix X beträgt bei bekannten Item- und Probandenparametern

$$L(\mathrm{X};\vec{\delta},\vec{\theta}) \;=\; \prod_{k=1}^{N} \exp(t_k \cdot \theta_k) \prod_{j=1}^{n} \frac{\exp(-x_j^{(k)}\delta_j)}{1+\exp(\theta_k-\delta_j)}\,.$$

Setzt man die Itemparameter als fest voraus, so ergibt sich für jeden Probanden k durch partielle Differentiation nach θ_k:

$$\begin{aligned}\frac{\partial L(\mathrm{X};\vec{\delta},\vec{\theta})}{\partial\theta_k} \;=\;& \Big(t_k \exp(t_k\theta_k - \sum_{s=1}^{n} x_s^{(k)}\delta_s) \prod_{j=1}^{n} \frac{1}{1+\exp(\theta_k-\delta_j)} \\ & - \exp(t_k\theta_k - \sum_{s=1}^{n} x_s^{(k)}\delta_s)(\sum_{s=1}^{n} \frac{\exp(\theta_k-\delta_s)}{1+\exp(\theta_k-\delta_s)})\Big) \\ & \cdot \prod_{s=1,s\neq k}^{N} \exp(t_s\cdot\theta_s) \prod_{j=1}^{n} \frac{\exp(-x_j^{(s)}\delta_j)}{1+\exp(\theta_s-\delta_j)} \cdot \prod_{j=1}^{n} \frac{1}{1+\exp(\theta_s-\delta_j)}\end{aligned}$$

Durch Nullsetzen der Ableitung ergibt sich die Schätzgleichung:

$$t_k = \sum_{s=1}^{n} \frac{\exp(\theta_k-\delta_s)}{1+\exp(\theta_k-\delta_s)}\,.$$

Ebenso zeigt man, dass sich aus den beobachteten Anwortvektoren bei als bekannt vorausgesetzten Probandenparametern für jeden Itemparameter δ_j die Schätzgleichung

$$A_j = \sum_{k=1}^{N} \frac{1}{1+\exp(\theta_j-\delta_k)}$$

ergibt. Verwendet man den beschriebenen Algorithmus, so muss man nach einer bewährten Faustregel anschließend die geschätzten Itemparameter alle mit $\frac{n-1}{n}$ multiplizieren, um die Werte aus der bedingten Schätzung zu erhalten.

Bei den Studien TIMSS und PISA wurde ein E-M-Algorithmus verwendet, da dort die Probanden verschiedene Testhefte erhielten, die nur durch gemeinsame „Ankeritems“ verbunden waren. So konnten in der PISA2000-Studie die Parameter von 117 Mathematikitems geschätzt werden, obwohl jeder Proband nur etwa 40 Items bearbeitete.

Wir geben einige kurze Hinweise zur Entstehungsgeschichte des Rasch-Modells:

Im Jahr 1958 schlug A. BIRNBAUM in einem Forschungsbericht[4] vor, an Stelle der Normalverteilungsogive Φ die logistische Funktion Ψ zur Erzeugung der Itemcharakteristiken von dreiparametrigen zweikategoriellen Testmodellen zu verwenden. In Kopenhagen erschien 1960 das bahnbrechende Buch von G. RASCH über probabilistische Modellierungen von Tests (Rasch, G.: Probabilistic models for some intelligence and attainment tests. Kopenhagen 1960, 2. Aufl.: Chikago 1980). Darin

[4] Series Report No. 15. Randolph Air Base: USAF School of Aviation Medicin, 1958

machte RASCH unter anderem auf eine Behandlung des Paarvergleichsproblems in Schachturnieren durch den Mathematiker (E. Zermelo, Mathematische Zeitschrift, 1929, *29*, S. 436-460) aufmerksam und zog Parallelen zur Bearbeitung von Testitems. So kam er dazu, auch Testmodelle mit logistischen Itemcharakteristiken zu betrachten und den gleichen Vorschlag wie BIRNBAUM zu machen. Dabei arbeitete er bereits deutlich die Vorteile eines Modells heraus, bei dem alle Diskriminationsparameter gleich sind (dieser Umstand erklärt, warum spätere Autoren dieses Testmodell kurz als RASCHmodell bezeichneten).

Die Ausarbeitung dieses Ansatzes durch G. RASCH, A. BIRNBAUM und E.B. ANDERSEN machte das Modell populär und löste eine Flut von Veröffentlichungen aus, die sich mit seiner Rechtfertigung, Anwendungsmöglichkeiten und Verallgemeinerungsfragen beschäftigten. Im deutschsprachigen Raum lässt sich der Beginn dieser Diskussion auf das Jahr 1968 datieren, in dem die erste Fassung von G.H. FISCHER (Hrsg.): Psychologische Testtheorie (Bern 1968) als Bericht über ein 1967 in Düsseldorf abgehaltenes Symposium zu neueren Entwicklungen in der Testtheorie erschien.

§2 Das 3-Parameter-Modell von Birnbaum

Herleitung des Modells: Betrachtet man für einen regulären verallgemeinerten Binomialtest mit der latenten Variablen θ, Itemcharakteristiken $\phi_1, \ldots, \phi_n$ und der Antwortvariablen $X := (X_1, \ldots, X_n)$ die bereits beim RASCH-Modell erwähnte logit-Transformation

$$\tau : p \mapsto \ln \frac{p}{1-p}, \qquad p \in]0;1[\,,$$

so sind alle transformierten Itemcharakteristiken $f_k : \theta \mapsto \tau(\phi_k(\theta))$ stetig nach θ differenzierbare Funktionen. Zeichnet man einen Probanden mit dem Fähigkeitswert θ_0 aus, so lassen sich die Vorschriften der Funktionen $f_2, \ldots, f_n$ nach dem Satz von TAYLOR in der Form

$$\begin{aligned} f_k(\theta) \;&=\; q_k(\theta_0) + (\theta-\theta_0)\cdot f_k'(\theta_0) + \frac{(\theta-\theta_0)^2}{2} r_k(\theta,\theta_0) \\ &\qquad \text{mit auf } \Theta\times\Theta \text{ stetigen Funktionen } r_1, \ldots, r_n\,. \end{aligned}$$

darstellen.

Da $\delta_k := f_k'(\theta_0) > 0$ gilt, ist jeweils die Grösse $\sigma_k := \theta_0 - q_k(\theta_0)/f_k'(\theta_0)$ definiert und in der Gruppe $\mathcal{M}_0$ der Probanden mit nahe bei θ_0 liegendem Parameterwert θ lässt sich jede der Funktionen f_k durch $\nu_k : \theta \mapsto (\theta - \sigma_k)\cdot\delta_k$ approximieren. Damit lassen sich in dieser Gruppe die Vorschriften der Itemcharakteristiken näherungsweise in der Form

$$\phi_k(\theta) \approx \frac{\exp\{\delta_k(\theta-\sigma_k)\}}{1+\exp\{\delta_k(\theta-\sigma_k)\}}$$

angeben. Ersetzt man $\phi_k(\theta)$ auch ausserhalb dieses Bereichs durch den Näherungsterm, so erhält man ein dreiparametriges Testmodell im, das 1958 von A. BIRNBAUM vorgeschlagen wurde und deshalb nach ihm benannt wird:

Definition 3.2 *Das durch die logistische Funktion Ψ erzeugte dreiparametrige 2-kategorielle Testmodell für einen verallgemeinerten Binomialtest mit n Items heisst* ***3-Parameter-Modell von Birnbaum***.

Eigenschaften: Die Vorschriften der Itemcharakteristiken des 3-Parameter-Modells lassen sich mit Hilfe eines Probandenparameters $\theta \in \mathbb{R}$, eines *Schwierigkeitsparameters* $\sigma \in \mathbb{R}$ und eines *Diskriminationsparameters* $\delta \in \mathbb{R}^+$ folgendermassen formulieren:

$$f_{\delta_k;\sigma_k}(\theta) := \frac{\exp\{\delta_k(\theta - \sigma_k)\}}{1 + \exp\{\delta_k(\theta - \sigma_k)\}}, \quad k = 1, \ldots, n \tag{4}$$

Diese Vorschriften lassen sich auch in der Form

$$f_{\delta_k;\sigma_k} = \frac{1}{1 + \exp\{\delta_k(\sigma_k - \theta_i)\}}, \quad k = 1, \ldots, n \tag{5}$$

schreiben. Zu beachten ist, dass nun δ_k die Steilheit der Charakteristik und σ_k den Wendepunkt des Grafen kennzeichnet.

Offensichtlich kann ein solches Modell reguläre verallgemeinerte Binomialtests auch dann noch akzeptabel approximieren, wenn das RASCH-Modell nicht mehr angebracht ist. F.M. LORD bevorzugte daher als Leiter des *Educational Testung Service* diesen Ansatz bei Itemanalysen und Testkonstruktionen und nahm die dabei auftretenden schätztechnischen Probleme in Kauf.

Die folgende Grafik demonstriert die Abhängigkeit der Itemcharakteristiken vom Diskriminationsparameter:

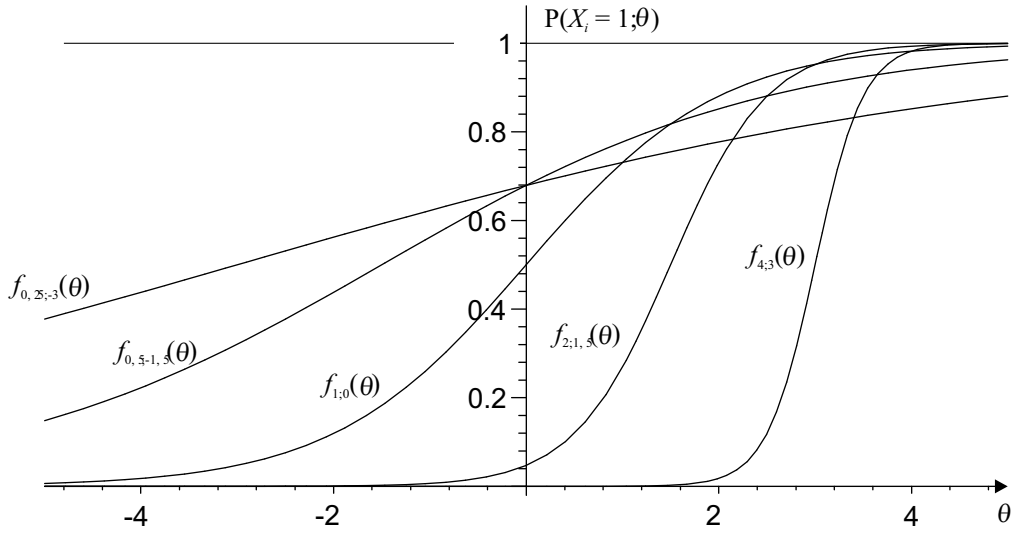


Fig. 9

Es lässt sich zeigen, dass die Wahrscheinlichkeit eines Antwortvektors $\vec{x}$ eines Probanden mit dem Fähigkeitswert θ sich in der Form

$$P(\vec{X} = \vec{x}) \quad = \quad \exp\{\theta \cdot T(x)\} \cdot \frac{\exp\{-\sum_{k=1}^{n} \delta_k x_k \sigma_k\}}{\prod_{k=1}^{n} (1 + \exp\{(\theta - \sigma_k)\delta_k\})}$$

$$\text{mit } T(x) := \sum_{k=1}^{n} \delta_k x_k$$

schreiben lässt. Daraus folgt dass T eine suffiziente Statistik bezüglich des Parameters θ ist und die bedingte Wahrscheinlichkeit für einen Antwortvektor $\vec{x}$ bei Fixierung eines Wertes von T unabhängig von θ ist. Da man jedoch die Diskriminationsparameter beim Schätzen nicht kennt, kann man diese Eigenschaft nicht für bedingte Schätzungen ausnutzen.

Mehrkategorielle Rasch-Modellierung

In einer Arbeit von G.N. MASTERS und B.D. WRIGHT aus dem Jahr 1984 werden neben dem dichtomen RASCH-Modell vier gebräuchliche mehrkategorielle Testmodelle mit geordneten Antwortkategorien unter dem Gesichtspunkt ihrer formalen Verwandtschaft diskutiert. Dabei wird gezeigt, dass bei *binomialer* Einstufungsverteilung, beim sogenannten *rating scale*-Modell von D. ANDRICH[5] und beim *partial credit*-Modell von G.N. MASTERS[6] die Wahrscheinlichkeit, ein m-stufig bewertetes Item j in der Kategorie k zu beantworten, sich mit einem Personenparameter θ und geeigneten Itemparametern $\delta_{j1}, \ldots, \delta_{jm}$ stets in der Form

$$P(x_{ij} = k; \theta, \delta_{j1}, \ldots, \delta_{jm}) = \begin{cases} \dfrac{1}{1 + \sum_{t=1}^{m} \exp\{k \cdot \theta_i - \sum_{s=1}^{t} \delta_{js}\}} & \text{für } k = 0 \\ \dfrac{\exp\{k \cdot \theta_i - \sum_{s=1}^{k} \delta_{js}\}}{1 + \sum_{t=1}^{m} \exp\{k \cdot \theta_i - \sum_{s=1}^{t} \delta_{js}\}} & \text{für } k = 1, \ldots, m \end{cases} \tag{6}$$

beschreiben lässt. Es sei darauf hingewiesen, dass in der obigen Darstellung die Notationsvereinbarung von MASTERS und WRIGHT zur Schreibweise der Wahrscheinlichkeit für die Kategorie 0 nicht übernommen wurde, da sie gegen die in der Mathematik üblichen Konventionen zur Summenschreibweise verstösst.

Hier ist die *Bewertungssumme* bezüglich θ eine minimalsuffiziente Statistik und das CML-Verfahren ist auf die Schätzung der Parameter δ_{jk} anwendbar.

Die Schätzung und Interpretation der Modellparameter $\delta_{j1}, \ldots \delta_{jm}$ ist natürlich davon abhängig, welche Hintergrundvorstellung man von dem Erreichen einer Bewertungskategorie hat. Wir demonstrieren dies in Anlehnung an die zitierte Arbeit für die *binomiale Bewertungsverteilung* und das *partial credit Modell*:

Binialmodellierung: Man nimmt an, dass ein Proband i in einem Item j insgesamt m stochastisch unabhängige Versuche hat, von denen jeder im Erfolgsfalle mit einer „Bewertungseinheit“ honoriert wird. Modelliert man nun die Erfolgswahrscheinlichkeit q_{ij} für den Einzelversuch in der Form

$$q_{ij} = \frac{\exp\{\theta_i - \sigma_j\}}{1 + \exp\{\theta_i - \sigma_j\}},$$

so gilt für $k = 1, \ldots, m$:

$$P(X_{ij} = k) = \binom{m}{k} q_{ij}^{k} (1 - q_{ij})^{m-k}$$

[5] D. ANDRICH, Educational and Psychological Measurement, 1978, *38*, S. 665-680

[6] G.N. MASTERS, Psychometrika 1982, *47*, S. 149-174

$$
\begin{aligned}
&= \binom{m}{k} \frac{\exp\{(\theta_i - \sigma_j)k\}}{(1+\exp\{\theta_i - \sigma_j\})^m} \\
&= \frac{\exp\{k\cdot\theta_i - k\cdot\sigma_j \;+\; \ln\binom{m}{k}\}}{(1+\exp\{\theta_i - \sigma_j\})^m} \\
&= \frac{\exp\{k\cdot\theta_i - \sum_{s=1}^{k}\delta_{jk}\}}{(1+\exp\{\theta_i - \sigma_j\})^m} \\
&\quad \text{mit } \delta_{js} := \sigma_j - \ln\frac{m-s+1}{s} \qquad \text{für } s = 1,\ldots,m
\end{aligned}
$$

Für die Kategorie 0 gilt:

$$P(X_{ij}=0) \;=\; \frac{1}{(1+\exp\{\theta_i-\sigma_j\})^m}$$

Da alle Terme im Nenner übereinstimmen und die Summe aller Bewertungswahrscheinlichkeiten 1 beträgt, ist die Summe aller Zählerterme für $k=0$ bis $k=m$ gleich dem gemeinsamen Nenner und es ergibt sich die Darstellung (6). Offensichtlich gibt es hier eigentlich nur *einen* echten Itemparameter, da für $1 \le s \le m$ gilt:

$$\delta_{js} \;=\; \delta_{j1} \;+\; \ln\frac{m\cdot s}{m-s+1}$$

Der Schwierigkeitsparameter σ_j für den Teilversuch ergibt sich aus der Beziehung $\sigma_j = \delta_{j1} + \ln\, m$.

partial credit Modell: Hier wird angenommen, dass sich für $k \ge 1$ die bedingte Wahrscheinlichkeit, unter der Annahme

$$\text{„}X_{ij} = k-1 \;\;\vee\;\; X_{ij} = k\text{“}$$

in Kategorie k zu antworten, in der Form

$$P(X_{ij}=k \mid X_{ij}=k-1 \;\vee\; X_{ij}=k) \;=\; \frac{\exp\{\theta_i-\delta_{jk}\}}{1+\exp\{\theta_i-\delta_{jk}\}}$$

modellieren lässt. Dieser Ansatz lässt sich im Gegensatz zum vorherigen Modell kaum *prozesstechnisch* begründen, da die verwendeten bedingten Wahrscheinlichkeiten eigentlich nicht auf der Prozessebene interpretierbar sind.

Durch eine Rekursion nach k folgt ebenfalls die Darstellung in Gleichung (6). Dabei sind jetzt *alle* Variablen δ_{js} echte Itemparameter (sie müssen wegen der verlangten Identifizierbarkeit des Testmodells noch Normierungsbedingungen unterworfen werden).