

# Verteilte Schulnoten sind normal – aber nicht notwendig normalverteilt

U. Mortensen,

01. 07. 2016

überarbeitete Version: 01. 09. 2018

## Inhaltsverzeichnis

<b>1 Leistungen, Verteilungen und Kriterien</b>	<b>2</b>
<b>2 Die klassischen Modelle</b>	<b>7</b>
2.1 Repräsentation von Merkmalsausprägungen . . . . .	7
2.2 Das Gauß-Modell . . . . .	12
2.3 Die Log-Normalverteilung . . . . .	18
2.3.1 Das Shockley-Modell . . . . .	18
2.3.2 Das Gibrat-Modell . . . . .	21
2.3.3 Illustration des log-Normal-Modells . . . . .	21
<b>3 Das Rasch-Modell</b>	<b>24</b>
3.1 Das Modell . . . . .	24
3.2 Spezifische Objektivität . . . . .	30
3.3 Die Schätzung der Modellparameter . . . . .	31
3.4 Die Bewertung von Schulleistungen . . . . .	33
<b>4 Anhang</b>	<b>35</b>
4.1 Zur Gauß-Verteilung . . . . .	35
4.2 Simulation log-normalverteilter zufälliger Variablen . . . . .	37
<b>Literatur</b>	<b>38</b>
<b>Index</b>	<b>40</b>

**Zusammenfassung:** Schulnoten werden in einigen Bundesländern nicht nach Maßgabe von festgelegten Kriterien, sondern nach Maßgabe der Repräsentation einer Leistung auf einer Normal- oder Gauß-verteilten Skala vergeben. Dabei werden die Parameter der Verteilung (ihr Erwartungswert  $\mu$  und ihre Varianz  $\sigma^2$ ) so aus den Punktwerten, wie sie etwa auf der Basis einer Klassenarbeit von den Schülerinnen und Schülern erworben wurden, bestimmt, dass die mittlere Note in der Nachbarschaft einer "3" ("befriedigend") liegt, und Einsen und Fünfen bzw. Sechsen einem bestimmten Prozentsatz entsprechen. Es wird eine Reihe von Gründen aufgeführt, die die Benotung nach diesem Schema als unsinnig erscheinen lassen. Zum einen ist keineswegs klar, dass die Normalverteilung das Leistungsspektrum beschreibt, andere Verteilungen wie die log-Normalverteilung, die logistische oder die Weibull-Verteilung sind plausiblere Alternativen, zumal es bei der Anwendung der Normalverteilung zu einer Konfundierung der Einschätzung der Schwierigkeit von Aufgaben in einer Arbeit und der Leistung der Schüler kommt. Es werden alternative Modelle zur Benotung aufgezeigt.

## 1 Leistungen, Verteilungen und Kriterien

Die Bewertung schulischer Leistungen kann für Schülerinnen und Schüler Rückmeldung über den Leistungsstand, aber auch Selektion bedeuten. Als Rückmeldung sollte sich die Bewertung an als allgemeingültig definierten Kriterien orientieren, und dementsprechend hat bereits 1968 die Kultusministerkonferenz (KMK) eine kriterienbezogene Benotung von schulischen Leistungen vorgeschrieben. Gleichwohl wirkt insbesondere in Bayern die Idee fort, dass Schulleistungen Gaußverteilt (normalverteilt) sind und die Noten deshalb ebenfalls Gaußverteilt zu sein haben. Es wird vermutet (Behler (2010)), dass auf der Basis dieser Annahme die Selektion guter, für weiterführende Schulen geeignete Schülerinnen und Schüler erleichtert werden könne. Die Annahme der Gauß-Verteilung wird im Folgenden kritisch diskutiert.

Die erste Frage ist, wie diese Annahme überhaupt begründet werden kann. Die Gauß-Verteilung ist die Verteilung einer zufälligen Veränderlichen  $X$  auf  $(-\infty, \infty)$ , d.h. es soll  $-\infty < X < \infty$  gelten. Schulische Leistungen können aber keine negativen Werte annehmen – weniger als nichts kann man nicht können –, und dementsprechend sind Schulnoten stets größer als Null. Es wird argumentiert, dass die mittlere Leistung im Vergleich zur Streuung der Messwerte groß ist, so dass negative Werte eine Wahrscheinlichkeit von praktisch Null haben, die Gauß-Verteilung also eine sehr gute Approximation darstellen könne. Das Argument ist richtig, macht aber bereits in einer grundsätzlichen Weise deutlich, dass die Annahme der Gauß-Verteilung eben nur eine Approximation sein kann. Es gibt aber weitere Eigenschaften der Gauß-Verteilung, denen die Verteilung der Schulleistungen genügen sollte, damit die Gauß-Verteilung eine vernünftige Approximation sein kann: die Gauß-Verteilung ist symmetrisch in Bezug auf den Punkt der Leistungsskala, an dem sie ihren maximalen Wert annimmt (vergl.

Abbildung 1: Gauß-Modell für Noten auf einer 1 - 5 - Skala; Standardabweichung links ist = .9, rechts ist = .6

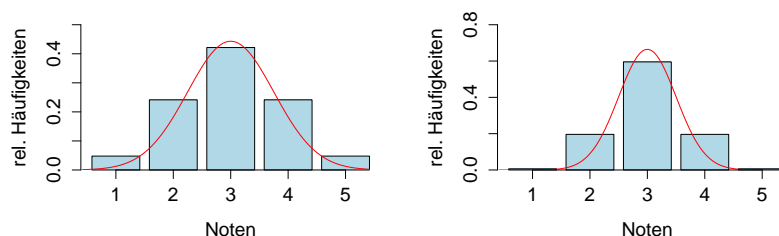
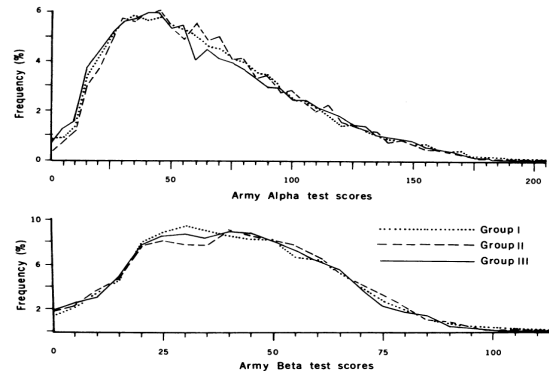


Abbildung 1). In Abschnitt 3 wird diese Theorie kurz vorgestellt; wie in der Abbildung 1 illustriert wird, ist nicht nur der Mittelwert (hier gleich 3) ein festzulegender Parameter, sondern auch die Standardabweichung; die Häufigkeiten der Noten hängen von diesen Parametern ab. Probleme für diese Approximation ergeben sich z.B. dann, wenn eine Klausur relativ zum Klassenstandard entweder schwierig oder leicht ist: im ersteren Fall ergibt sich eine Häufung der Punktwerte am unteren Ende der Punkteskala, im zweiten Fall ergibt sich eine Häufung am oberen Ende der Punkteskala. Da die Gauß-Verteilung symmetrisch in Bezug auf den Erwartungswert ist muß möglicherweise eine Varianz postuliert werden, die zu klein im Vergleich mit der beobachteten Varianz der Punktwerte ist, und überdies wird der empirischen Schiefe der Verteilung nicht Rechnung getragen. In Abschnitt 2.2 werden diese Aspekte der Gauß-Approximation näher erläutert.

Man kann natürlich argumentieren, dass letztlich jede Verteilung eine Approximation an die jeweils betrachteten realen Daten darstellt. Die Frage ist nun, warum man gerade auf die Gauß-Verteilung als Approximation fokussiert, zumal andere Verteilungen im allgemeinen genauso gut an die Daten angepasst werden können und darüber hinaus sinnvollere Interpretationen der Daten (Punktwerte in Klassenarbeiten etc) erlauben. Vielleicht ist einer der Gründe dafür die suggestive Wirkung des Ausdrucks 'Normalverteilung'; dieser geht auf den französischen Mathematiker und Astronomen Lambert Adolphe Jacques Quetelet (1796 – 1874) zurück, der auf der Basis von biologischen Messungen (z.B. der Brustumfänge französischer Rekruten) behauptete, die Gauß-Verteilung sei der Normalfall bei biologischen und sozialen Variablen, weshalb er den Ausdruck 'Normalverteilung' prägte. Spätere genauere Messungen zeigten aber, dass biologische Größen keineswegs normalerweise Gauß-verteilt sind; man kann argumentieren, dass die *log-Normalverteilung* die bessere Approximation ist (Aitchinson & Brown (1963), Koch (1966), Limpert, Stahel & Apt (2008), Limpert & Stahel (2011)). Diese ergibt sich, wenn eine Variable  $X$  gemessen wird und die Logarithmen  $Y = \log_e X$  der Messwerte normalverteilt sind. Dann gilt  $X = e^Y$  und  $Y$  ist Gauß-verteilt.  $e^Y$  kann nicht negativ werden, also ist  $X$  nur für positive Größen definiert. In Ab-

Abbildung 2: log-normalverteilte Army Alpha Scores (aus Dorfman 1978)



schnitt 2.3 werden Begründungen für die Annahme dieser Verteilung aufgeführt, Abbildung 2 zeigt die Häufigkeitsverteilung von Testwerten im *Army-Alpha-Test*, einem vom US-Militär entwickelten Intelligenztest; die Häufigkeiten der Testwerte ("Scores") entsprechen einer log-Normalverteilung. Dorfman (1978), p. 1180, zitiert Yule und Kendall (1937) – beide renommierte Statistiker – mit der Bemerkung: "The normal curve was, in fact, to the early statisticians what the circle was to the Ptolemaic astronomers" (der Kreis galt als "perfekte" Figur, woraus man folgerte, dass Gott die Planetenbahnen als Kreisbahnen eingerichtet habe; Kepler hatte Mühe, sich auf der Basis seiner Analyse der Daten Tycho Brahes von dieser Vorstellung zu lösen). McNemar (1942) merkt an, dass "the ease with which the shape of a distribution can be altered by a change in test difficulty would also have served as a warning to those who were out to demonstrate the normal law for psychological traits" (zitiert nach Dorfman (1978), p. 1180). Bereits der Ausdruck *normal law* ist verstörend, denn die Gauß-Verteilung repräsentiert kein Gesetz in dem Sinne, dass biologische, psychologische und welche Daten auch immer Gauß-verteilt sein müssen. Die Gauß-Verteilung ergibt sich höchstens dann als ein zwingend anzunehmendes Wahrscheinlichkeitsmodell, wenn sie aus den Annahmen folgt, die über den jeweils betrachteten Prozess gemacht werden, – aber dann ist die Gauß-Verteilung eben eine Folgerung aus diesen Annahmen, und ob die zwingend sind, ist wieder eine andere Frage. Viele Wahrscheinlichkeitsverteilungen können sich in ihrer Form sehr ähneln, obwohl sich die Formeln für diese Verteilungen sehr unterscheiden, die im Folgenden näher betrachteten Gauß-Verteilungen, log-Normalverteilungen und die logistische Verteilungen sind ein Beispiel für diesen Sachverhalt. Diese impliziert, dass an ein und denselben Datensatz ganz verschiedene Modelle angepasst werden können<sup>1</sup>.

<sup>1</sup> "Any statement can be held true come what may, if we make drastic enough adjustments elsewhere in the system." Quine (1951) (p. 40) Quines Arbeit gehört zu den zentralen Arbeiten zur Wissenschaftstheorie im 20-ten Jahrhundert, ihre Bedeutung hält bis heute an. In Bezug auf

Abbildung 10 auf Seite 27 illustriert diesen Sachverhalt für die Gauß- und die logistische Verteilung.

Ein wichtiges Desiderat der Benotung schulischer Leistungen ist die Objektivität der Benotung. Ein Test ist *objektiv*, wenn er zur selben Beurteilung durch verschiedene Personen führt. Das Ausmaß an Objektivität eines Tests, hier z.B. einer Klassenarbeit, kann durch bestimmte statistische Verfahren bestimmt werden, worauf an dieser Stelle aber nicht weiter eingegangen werden muß. Pädagogen sind oft der Meinung, ihre Schülerinnen und Schüler 'ganzheitlich' beurteilen zu können; die Note soll dann eine solche ganzheitliche Beurteilung repräsentieren. Allerdings ist es bemerkenswert, in welchem Ausmaß sich die Beurteilungen derselben Leistungen durch verschiedene LehrerInnen voneinander unterscheiden können; ein und derselbe Schulaufsatz kann mit sämtlichen Schulnoten von "1" bis "6" beurteilt werden. Selbst wenn verschiedene Lehrer zum selben Urteil kommen ist damit noch nicht nachgewiesen, dass die Beurteilung vorurteilsfrei ist<sup>2</sup>.

Schließlich gibt es noch die Frage nach der Skaleneigenschaft der Schulnoten. Von einem formalen Standpunkt aus gesehen sind Schulnoten Messungen. Der Begriff der Messung wird in der Statistik sehr allgemein gefaßt und wird in Zusammenhang mit dem der Skala definiert. Allgemein bedeutet Messen die Zuordnung von Zahlen zu Objekten. Man unterscheidet vier Haupttypen von Skalen, auch Messniveaus genannt:

- (1) *Kategoriale Messung – die Kategorial- oder Nominalskala*, bei der ein Objekt nur kategorisiert und dementsprechend mit einem Namen versehen wird. Zahlen spielen hier nur die Rolle von Namen und reflektieren keine quantitativen Unterschiede zwischen den Kategorien (etwa: weiblich – männlich, blond – brünett, etc),
- (2) *Ordinale Messungen – die Ordinalskala*, bei der Objekte bezüglich der Ausprägung eines Merkmals in eine Rangordnung gebracht werden. Die Zahlen repräsentieren nur die Position in der Rangreihe. Gleiche Differenzen zwischen Rangzahlen bedeuten nicht auch gleiche Differenzen bezüglich der Ausprägungen des betrachteten Merkmals,
- (3) *Intervallmessungen – die Intervallskala*, bei der die Maßeinheit sowie der Nullpunkt der Skala frei wählbar sind, und gleiche Differenzen der Skalenwerte bedeuten auch gleiche Differenzen der Merkmalsausprägungen (Beispiel: Tempe-

---

die Wahl der Normalverteilung zur Interpretation schulischer Leistungen heißt die Quinesche Folgerung (es handelt sich Quines Aussage nicht um eine Behauptung, sondern um das Resultat einer längeren Analyse), dass man diese Verteilung in den meisten Fällen an die Daten anpassen kann, – wenn man nur bereit ist, die dadurch erzwungenen Implikationen zu akzeptieren. Aber wozu diese Übung, wenn inhaltlich bessere Modelle zur Verfügung stehen?

<sup>2</sup><http://www.presse.uni-oldenburg.de/mit/2010/319.html>. Eine Zusammenfassung der Psychologie der Urteilsfehler findet man in Tversky & Kahneman(1974) (dieser Artikel kann mit anderen hier erwähnten Arbeiten von der Seite <http://www.uwe-mortensen.de/SkriptenStatistik.html> heruntergeladen werden. Dort findet man auch ein ausführliches Skriptum zur Konstruktion psychometrischer Tests, in dem einige der hier angesprochene Begriffe ausführlicher behandelt werden: <http://www.uwe-mortensen.de/TestkonstruktionNeuB.pdf>

raturskalen wie die Celsius- und die Fahrenheitskala);

(4) *Verhältnismessung – die Verhältnisskala*, bei der nur die Maßeinheit frei gewählt werden kann, der Nullpunkt liegt fest. Beispiele sind Längen- und Gewichtsskalen, und für die Temperatur ist es die Kelvin-Skala, bei der der absolute Nullpunkt das absolute Nichtvorhandensein von Wärme repräsentiert: Da Wärme ein kinetisches Phänomen ist, repräsentiert der Nullpunkt den Fall, bei dem absolut keine Bewegung mehr existiert.

Es wird oft postuliert, dass die Annahme einer Gauß-Verteilung automatisch die Konstruktion einer Intervallskala impliziert. Dies ist nicht so, die Konstruktion einer Intervallskala beruht stets – d.h. auch bei anderen Verteilungen – auf zusätzlichen Annahmen.

Schulnoten haben dem Anspruch nach die Qualität von Werten auf einer Intervallskala, d.h. der *Unterschied* von Leistungen, die mit der Note "1" bewertet wurden, und Leistungen, die mit der Note "2" bewertet wurden, entspricht dem Unterschied von mit der Note "2" bzw. mit der Note "3" bewertet wurden, etc. Diese Hypothese läßt sich überprüfen, – und wird im Allgemeinen *nicht* bestätigt, so dass viele Schulforscher der Ansicht sind, dass Noten bestenfalls die Qualität einer Ordinalskala (also einer Rangskala) haben. Dies sollte beim Kampf um die Noten, bei dem es um die Zahl nach dem Dezimalpunkt gehen kann (1.5 oder 1.7?, noch 4.5 oder doch schon die 5?), bedacht werden.

Die Beantwortung von Fragen bzw. die Lösung von Aufgaben wird oft mit Punkten bewertet, und die Note wird aus der Summe aller Punkte bestimmt, wobei die Punktwerte für die verschiedenen Fragen noch gewichtet werden können. Die Gewichtung soll die möglicherweise unterschiedlichen Schwierigkeiten der Fragen oder Aufgaben widerspiegeln. Die Frage ist nun, wie der Begriff der Schwierigkeit so definiert werden kann, dass er quantifiziert in die Bewertung durch Punkte eingehen kann. Die Schwierigkeit wird in der (statistischen) Theorie der psychologischen Tests<sup>3</sup>, durch den Anteil  $\pi_j$  der Personen, die die  $j$ -te Frage beantworten bzw. die  $j$ -te Aufgabe ( $j = 1, \dots, N$ ,  $N$  die Gesamtzahl der Fragen bzw. Aufgaben eines Tests) lösen können definiert: Je mehr Personen eine Aufgabe lösen bzw. eine Frage beantworten können, desto leichter ist die Aufgabe, und je weniger Probanden die Aufgabe lösen können, desto schwieriger ist sie. Auf eine analoge Definition von 'Schwierigkeit' wird i. a. bei der Bewertung schulischer

---

<sup>3</sup>Gemeint ist die *Klassische Testtheorie* (KKT), die in der ersten Hälfte des 20-ten Jahrhunderts entwickelt wurde. Man ging von der Annahme aus, dass einerseits die "wahre" Fähigkeit einer beliebig gewählten Person durch einen Wert  $\tau$  repräsentiert werden kann, andererseits der Messwert  $X$  dieser Fähigkeit durch zufällige Effekte, die insgesamt durch eine zufällige Veränderliche  $\varepsilon$  repräsentiert werden können, beeinflusst wird, und dass diese Effekte *additiv* in den Messwert  $X$  eingehen, so dass man  $X = \tau + \varepsilon$  schreiben kann.  $\varepsilon$  wird wiederum als Summe voneinander (mehr oder weniger) unabhängigen zufälligen Effekten aufgefasst:  $\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \dots$ . Wie in der mathematischen Statistik gezeigt wird, strebt diese Summe gegen eine Gauß-Verteilung, wenn die Anzahl der  $\varepsilon_j$  nur "hinreichend groß" wird; dies ist die Aussage des *Zentralen Grenzwertsatzes*. Dann ist auch  $X$  approximativ Gauß-verteilt. Der Zentrale Grenzwertsatz wird zur Begründung der Annahme,  $X$  sei Gauß-verteilt, herangezogen.

Leistungen zurückgegriffen. Diese Definition führt aber in Probleme, wenn es um den Vergleich verschiedener Populationen geht. Im schulischen Bereich ergeben sich bereits Probleme, wenn verschiedene Klassen vergleichbar gemacht werden sollen; wie sich zeigt, kann es zu systematischen Über-, aber auch Unterbewertungen kommen. In der diagnostischen Psychologie hat man nach Alternativen für die Definition des Schwierigkeitsbegriffs gesucht. Es läßt sich zeigen, dass die Schwierigkeit in einem bestimmten Sinn unabhängig von der Population durch einen Parameter ausgedrückt werden kann, wenn man *nicht* von der Gauß-Verteilung der Leistungen ausgeht, sondern statt dessen die logistische Verteilung annimmt; diese Verteilung wird in Abschnitt 3.1 vorgestellt. Hier sei angemerkt, dass die logistische Verteilung der Gauß-Verteilung sehr ähnlich sieht, allerdings ist sie durch einen anderen mathematischen Ausdruck definiert. Die auf der Annahme der logistischen Verteilung basierende Theorie psychometrischer Tests (eine Klassenarbeit kann als ein solcher Test angesehen werden). Die auf der Annahme der logistischen Verteilung beruhende sogenannte *probabilistische Testtheorie* – auch Item-Response-Theorie bekannt – vermeidet die mit dem Parameter  $\pi_j$  verbundene Zirkularität der Definition der 'Schwierigkeit', konfligiert aber mit Betrachtungen, denen zufolge die log-Normalverteilung die angemessene Verteilung ist. Die drei Verteilungen – die Gauß-Verteilung, die log-Normalverteilung und die logistische Verteilung können sich, für entsprechend gewählte Parameter, sehr ähneln, so dass es schwierig sein kann, anhand von empirischen Daten zu entscheiden, welche von ihnen nun die "wahre" Verteilung ist, und ob überhaupt eine von ihnen die wahre Verteilung ist. Man könnte die Entscheidung zwischen ihnen als rein akademische und insofern als vernachlässigbare Übung betrachten, würde sie nicht potentiell zu verschiedenen Interpretationen der Daten führen, und dieser Sachverhalt kann durchaus Folgen für die beurteilten SchülerInnen haben. Dieser Befund entspricht der wissenschaftstheoretischen Einsicht, dass jede Interpretation von Daten metaphysische, d.h. nicht direkt aus den Daten ableitbare Aspekte hat, über die sich streiten läßt. Diesen sehr allgemeinen Sachverhalt sollte man vielleicht bedenken, wenn es zu einem Streit darüber, ob eine Leistung noch mit 1.7 oder vielleicht besser "nur" mit 2.0 zu bewerten ist, kommt.

Die probabilistische Testtheorie wird in den PISA-Studien angewandt, wo es ja um den Vergleich verschiedener Länder geht. In Abschnitt 3 wird diese Theorie kurz vorgestellt.

## 2 Die klassischen Modelle

### 2.1 Repräsentation von Merkmalsausprägungen

Das Ausmaß der Kompetenz in einem bestimmten Fach kann von Person zu Person variieren. Die Benotung der Kompetenz bedeutet, das jeweilige Ausmaß von Kompetenz auf einer Skala zu repräsentieren. Die so vertraute Notenskala

suggeriert, dass Kompetenz auf einer einzelnen Skala abgebildet werden kann. Tatsächlich ist Kompetenz im Allgemeinen ein mehrdimensionales Merkmal. Im Fach Geschichte kann man etwa annehmen, dass (i) historisches Faktenwissen und (ii) die Fähigkeit, historische Entwicklungen zu interpretieren zwei verschiedene Dimensionen sind; vermutlich sind weitere Dimensionen denkbar. In der Mathematik können reines Faktenwissen (Lehrsätze und Formeln) und die Fähigkeit, Probleme zu lösen (Beweise zu finden) ebenfalls zwei verschiedene Dimensionen sein. Analog kann man in anderen Fächer argumentieren. Kompetenz setzt sich dann aus den Ausprägungen auf diesen, im allgemeinen Fall  $r \geq 1$  Dimensionen zusammen, im einfachsten Fall additiv:

$$X = a_1 D_1 + a_2 D_2 + \dots + a_r D_r \quad (2.1)$$

wobei die  $a_j$ ,  $j = 1, \dots, r$  Gewichte repräsentieren, mit denen die Kompetenzdimensionen in das Maß  $X$  eingehen. Nun kann man in der Mathematik bei aller Kreativität keine Probleme lösen, wenn man kein entsprechendes Grundwissen hat; dieser Sachverhalt kann bedeuten, dass ein linearer Ansatz wie (2.1) inadäquat ist; eine bessere Repräsentation kann durch

$$X = a_1 D_1 + a_2 D_2 + a_3 D_1 D_2 \quad (2.2)$$

gegeben sein, wobei der Einfachheit halber  $r = 2$  angenommen wurde. Das Produkt  $D_1 D_2$  repräsentiert den Fall, dass die Kompetenz durch eine "Interaktion" der beiden Dimensionen mitdefiniert wird:  $D_2$  geht proportional zu  $D_1$  in die Kompetenz ein, oder  $D_1$  proportional zu  $D_2$ . In welcher Form für einen gegebenen Beurteiler Kompetenz durch die Dimensionen definiert wird ist nicht bekannt, so dass man auch einfach

$$X = f(D_1, \dots, D_r) \quad (2.3)$$

schreiben kann;  $f$  ist hier eine unbekannt Funktion der  $r$  Dimensionen, wobei auch nicht bekannt ist, wieviele Dimensionen in ein Urteil eingehen. Hinzu kommt, dass auch nicht kompetenzbezogene Größen in das Urteil eingehen. Die Leistungen von Schülern und Schülerinnen können auf Grund solcher Faktoren von einer Stunde zur nächsten, also auch von einer Klassenarbeit zur nächsten schwanken. Diese Faktoren sind im Einzelnen nicht notwendig bekannt und in diesem Sinne zufällig. Aber auch die Lehrerin oder der Lehrer sind keine konstant und deterministisch funktionierenden Automaten. Deswegen wird man in den Ausdrücken für  $X$  einen Fehlerterm  $e$  (von englisch *error*, es handelt sich also nicht um die Eulersche Zahl  $e$ , der Basis des natürlichen Logarithmus) hinzufügen müssen:

$$X = f(D_1, \dots, D_r) + e. \quad (2.4)$$

$e$  bildet alle nicht systematischen Effekte in  $X$  ab, weshalb  $X$  eine zufällige Veränderliche Sinne der Wahrscheinlichkeitstheorie und mathematischen Statistik ist. Man kann die Funktion  $f$  als die Beurteilungsregel auffassen, die den Urteilen einer Lehrerin oder eines Lehrers zugrunde liegt;  $f$  repräsentiert den systematischen



Aspekt der Leistungsbeurteilung. Man muß hier hinzufügen, dass die Konzeption der Beurteilungsregel eigentlich noch in Rechnung stellen muß, dass auch die Anzahl der jeweils berücksichtigten Dimensionen, die Abschätzung ihres jeweiligen Wertes und ihre Gewichtung letztlich zufällige ("stochastische") Prozesse sind. Die Repräsentation (2.4) stellt dementsprechend eine Vereinfachung dar,  $f(D_1, \dots, D_r)$  bildet nur ab, wie die Beurteilung gewissermaßen im Mittel verläuft. Empirisch zeigt sich oft, dass der Ansatz (2.1), erweitert um den Term  $e$ , eine gute Näherung darstellt.

$X$  kann im Prinzip als eine stetige zufällige Veränderliche betrachtet werden, d.h. sie ist auf einem Intervall von reellen Zahlen definiert. Für den Fall, dass  $X$  als Gauß-verteilt angenommen wird ist das Intervall  $(-\infty, +\infty)$ , für den Fall, dass eine log-normale Verteilung angenommen wird, ist das Intervall  $[0, \infty)$ .

$X$  ist nicht direkt beobachtbar, der Wert für eine(n) Probandin/en muß erschlossen werden. Es werden – z.B. in einer Klassenarbeit –  $n$  Aufgaben bzw. Fragen gestellt, und für die Antworten werden Punkte vergeben, für die Frage 1 die Punktzahl  $X_1$ , für die Frage 2  $X_2$ , etc. Die Gesamtzahl der Punkte ist die Summe  $S = X_1 + \dots + X_n$ . Natürlich ergibt sich sofort die Frage, wie die  $X_j$  definiert werden, d.h. wieviele Punkte für die Antworten auf die  $j$ -te Frage vergeben werden sollen. Es genügt, für die Zwecke dieses Aufsatzes den einfachen binären Fall anzunehmen, bei dem eine Antwort oder Aufgabenlösung entweder richtig oder falsch ist.  $X_j$  ist dann als Indikatorvariable

$$X_j = \begin{cases} 0, & \text{Frage nicht oder falsch beantwortet} \\ 1, & \text{Frage korrekt beantwortet} \end{cases} \quad (2.5)$$

definiert. Es bleibt dann noch die Frage, ob alle Fragen gleich gewichtet werden sollen oder können, oder ob im Falle unterschiedlich schwieriger Fragen eine Gewichtung vorgenommen werden soll, so dass  $X_j = b_j X_j$ .  $b_j$  ist das Gewicht für die  $j$ -te Frage. Dann ergibt sich das Problem, die  $b_j$  zu bestimmen. Im einfachsten Fall setzt man  $b_j = 1$  für alle  $j = 1, \dots, n$  und nimmt damit an, dass alle Fragen oder Aufgaben dieselbe Schwierigkeit haben. Eine zweite Möglichkeit ist,  $b_j = 1/\pi_j$  zu setzen, wobei  $\pi_j$  die Schwierigkeit der  $j$ -ten Frage ist. Dabei ist  $\pi_j$  gleich dem Anteil von Personen in der Stichprobe (etwa der Schüler in einer Klasse), die die  $j$ -te Frage korrekt beantwortet haben. Je größer der Wert von  $\pi_j$ , desto weniger geht die  $j$ -te Frage in den Gesamt-Score  $S$  ein; die Problematik dieses Schwierigkeitsbegriffs ist bereits in der Einleitung angedeutet worden. Besser wäre eine empirische Bestimmung der  $b_j$  als Regressionskoeffizienten in dem Ansatz

$$S = b_1 X_1 + \dots + b_n X_n + e \quad (2.6)$$

nur benötigt man dafür eine von den speziellen Aufgaben unabhängige Schätzung für  $S$ , – und die ist nicht gegeben, wenn man nicht einfach ein globales Rating für  $S$ , wie ein Lehrender sie für einen Schüler abgeben könnte, für diese Schätzung annehmen will. Der Nachteil eines solchen Vorgehens liegt wegen der

in der Einleitung genannten Nachteile einer "ganzheitlichen" Beurteilung auf der Hand. Man ist gezwungen, jeweils eine ad hoc-Setzung für die  $b_j$  zu wählen, – etwa  $b_j = 1$  für alle  $j$ .

Es sei  $S_i$  der Gesamt-Score für den Probanden  $Pb_i$ .  $S_i$  ist eine Schätzung für  $X_i$ , der Kompetenz von  $Pb_i$ . Von  $S_i$  muß auf die Note  $N_i$  für  $Pb_i$  geschlossen werden. Nun ist aber nach den obigen Ausführungen  $X_i$  eine zufällige Veränderliche, so dass man davon ausgehen kann, dass auch der Punktwert  $S_i$  eine zufällige Veränderliche ist. Zufälligen Veränderlichen sind Wahrscheinlichkeitsverteilungen zugeordnet, in deren Definition bestimmte Parameterwerte auftauchen. Zwei wichtige Parameter sind der Erwartungswert  $\mathbb{E}(X_i)$  bzw.  $\mathbb{E}(S_i)$  und die Varianz  $Var(X_i)$  bzw.  $Var(S_i)$ . Der Erwartungswert ist das arithmetische Mittel über *alle möglichen* Werte der zufälligen Veränderlichen, und die Varianz ist das arithmetische Mittel über alle möglichen Abweichungen  $(X_i - \mathbb{E}(X_i))^2$  (analog für  $S_i$ ). Es gibt nun zwei Möglichkeiten:

1. Man definiert anhand gewisser Kriterien die Intervalle der Gesamt-Scores, die den Noten entsprechen. Ist der maximal erreichbare Punktwert  $S = 50$ , so könnte man festsetzen, dass allen Punktwerten zwischen 45 und 50 "1" zugeordnet wird, allen Punktwerten zwischen 38 und 44 eine "2", etc.
2. Man bestimmt die zu den Noten korrespondierenden Intervalle nach Maßgabe einer Wahrscheinlichkeitsverteilung. Beliebige ist hier die Gauß-Verteilung, – wohl aufgrund der Annahme, dass die Gauß-Verteilung, den Argumenten Quetelets folgend, als "natürliche" und daher "normale" Verteilung gilt.

Es muß eine Häufigkeitsverteilung für die  $S_i$ -Werte erstellt werden, an die die Verteilung einer stetigen zufälligen Veränderlichen angepasst werden kann (dies ist üblicherweise die Gauß-Verteilung). Die  $S_i$  sind selbst zufällig verteilt. (2.6) entsprechend könnte man für  $S_i$  ebenfalls eine stetige Veränderliche annehmen, sofern die  $b_j$  irgendwelche reellen Zahlen sein dürfen. Andererseits hat man kaum eine Möglichkeit, die  $b_j$  zu schätzen. Nimmt man vereinfachend an, dass  $b_j = 1$  für alle  $j$ , so ist  $S_i$  diskret, d.h.  $S_i$  kann nur ganzzahlige Werte annehmen. Kann man annehmen, dass die Aufgaben unabhängig voneinander beantwortet werden, so wäre die verallgemeinerte Binomialverteilung eine vernünftige Annahme für die Verteilungsfunktion  $F_i(S)$ . Allerdings ist die Annahme der Unabhängigkeit nicht automatisch gerechtfertigt und bedarf der Überprüfung. xxx

Aus den  $S_i$ -Werten läßt sich eine Häufigkeitsverteilung der Punkte erstellen. Es ist oben gesagt worden, dass die individuellen Scores  $S_i$  selbst zufällig verteilt sind. Es sei  $I_j = [g_j, g_j + \Delta g_j)$  ein Intervall dieser Häufigkeitsverteilung; für  $\Delta g_j = \Delta g$  für für alle  $j$  sind die Intervalle für die Häufigkeitsverteilung gleich groß. Wenn  $S_i$  diskret ist, müssen die  $g_j$ - und  $\Delta g_j$ -Werte ebenfalls Werte aus dem Bereich der  $S_i$ -Werte sein. Dann ist

$$P(S_i \in I_j) = F_i(g_j + \Delta g) - F_i(g_j), \quad i = 1, \dots, N; j = 1, \dots, k \quad (2.7)$$

die Wahrscheinlichkeit, dass der Score der  $i$ -ten Person in die  $j$ -te Kategorie fällt.  $F_i = F_i(s) = P_i(S \leq s)$  ist die Verteilungsfunktion für die Scores bei der  $i$ -ten

Person. Dann ist

$$p_j = \sum_{i=1}^N P(S_i \in I_j) = \sum_{i=1}^N (F_i(g_j + \Delta g) - F_i(g_j)) \quad (2.8)$$

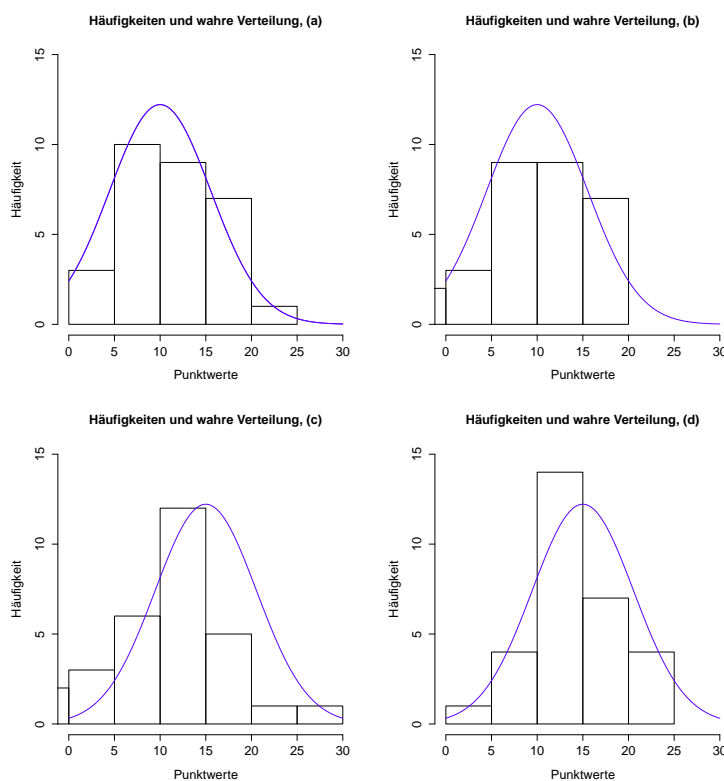
Die Wahrscheinlichkeit, dass überhaupt einer der Scores in das  $j$ -te Intervall fällt;  $Np_j = \hat{n}_k$  ist dann die erwartete Häufigkeit von Scores in  $I_j$ , und  $\sum_k n_k = N$ . Jeder Score fällt mit Sicherheit in eines der Intervalle  $I_j$ , da ja  $I_1 \cup \dots \cup I_k$  die gesamte Notenskala abdecken. Dann ist

$$\sum_{j=1}^k p_j = \sum_{j=1}^k \sum_{i=1}^N (F_i(g_j + \Delta g) - F_i(g_j)) = 1. \quad (2.9)$$

Sind die  $F_i$  Gauß-Verteilungen, so entspricht die Wahrscheinlichkeit  $p_j$  *nicht* den entsprechenden Wahrscheinlichkeiten einer Gauß-Verteilung. Die  $p_j$ -Werte korrespondieren zu einer Verteilung, die im Vergleich zu einer Gauß-Verteilung im Allgemeinen eine größere Wölbung zeigt, d.h. sie ist nicht so "spitzgipfelig" wie eine Gauß-Verteilung. Will man für die Häufigkeitsverteilung der  $S_i$ -Werte eine Gauß-Verteilung annehmen, so müssen die  $F_i$  entsprechend anders definiert sein. Sie müssen im Vergleich zur durch die  $p_j$  definierten Verteilung sehr spitzgipfelig sein; möglicherweise können Gaußsche  $F_i$ -Verteilungen mit sehr kleinen Varianzen zu einer akzeptablen Gauß-Approximation für die  $p_j$ -Verteilung führen.

Entscheidet man sich für die Möglichkeit 2., so hat man die Frage nach den Intervallgrenzen für die Notenintervalle noch nicht beantwortet. Wie Thomas (1982) ausführt scheint sich die Auffassung festgesetzt zu haben, dass die Annahme einer Gauß-Verteilung eine Intervallskala für die Noten (bzw. für die von Thomas diskutierten Intelligenzquotienten) impliziert; dies ist die Annahme, dass die Gauß-Verteilung aus rein mathematischen Gründen stets eine Intervallskala für die jeweils gemessene Größe impliziert. Diese Annahme ist natürlich nicht gerechtfertigt. Die Annahme einer Gauß-Verteilung impliziert nur die Annahme, dass das betrachtete Merkmal kontinuierlich variiert. Diese Annahme ist für ein Merkmal wie Kompetenz plausibel; würde man sie nicht machen, so müßte man eben von der Annahme ausgehen, dass die Kompetenz in diskreten Abschnitten ausgeprägt ist, und es ist einigermaßen schwer, zu erklären, wie es zu diskreten Ausprägungen kommen soll und wie man sie repräsentieren soll. Die Konstruktion einer Intervallskala für die Noten bedeutet, dass den Intervallen zwischen den Noten gleich große Unterschiede zwischen den Kompetenzen entsprechen. Aber eine derartige Wahl von Intervallgrenzen ist nur eine von mehreren möglichen, wie sich leicht verdeutlichen läßt. Es wird zunächst die Gauß-Verteilung betrachtet, im Anschluß daran die log-Normalverteilung.

Abbildung 3: Gauß-Dichten und Zufallsstichproben für Punkteverteilungen ( $n = 30$ ): (a) und (b) –  $\mu = 10, \sigma = 5.55$ , (c) und (d)  $\mu = 15, \sigma = 5.55$ .



## 2.2 Das Gauß-Modell

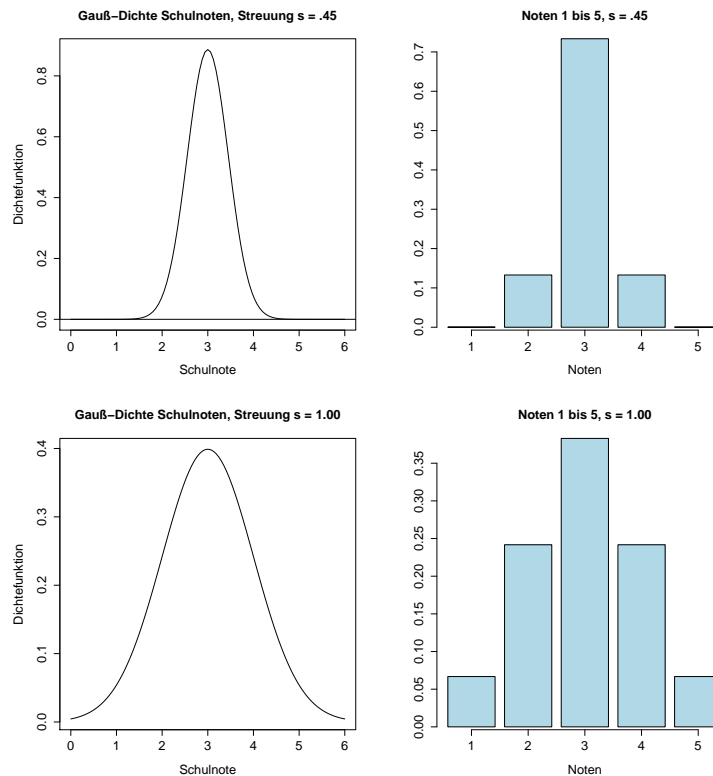
Die Gauß-Dichte ist durch

$$f_X(x) = \frac{1}{\sigma_x \sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(x - \mu_x)^2}{\sigma_x^2}\right), \quad -\infty < x < \infty \quad (2.10)$$

gegeben. Wie auf Seite 6 (Fußnote) bereits angemerkt wurde, wird als Begründung für die Annahme der Gauß-Verteilung üblicherweise der Zentrale Grenzwertsatz<sup>4</sup> angenommen. Abbildung 3 zeigt mögliche Stichprobenverteilungen von Punktwerten in einer Klasse mit 30 Schülerinnen und Schülern. In den Fällen (a) und (b) ist die Klausur "zu schwer" – die niedrigeren Punktwerte kommen

<sup>4</sup>Man kann  $X = \mu + \varepsilon(m)$  setzen, mit  $\varepsilon(m) = \varepsilon_1 + \varepsilon_2 + \dots + \varepsilon_m$ . Die  $\varepsilon_j$  seien unabhängig und identisch verteilt, jeweils mit dem Erwartungswert  $c$  und endlicher Varianz  $\sigma_\varepsilon^2$ . Dann strebt die Verteilung von  $(\varepsilon(m) - mc)/\sigma_\varepsilon \sqrt{m}$  mit wachsendem  $m$  gegen die Standardnormalverteilung, d.h. gegen (2.10) mit  $\mu = 0$  und  $\sigma = 1$ . Der Satz kann auf den Fall schwacher Abhängigkeiten zwischen den  $\varepsilon_j$  verallgemeinert werden, vergl. Billingsley (1979), Abschnitt 27.

Abbildung 4: Gauß-Dichten für Schulnoten: verschiedene Streuungen; rechts die zur Dichte korrespondierenden Verteilungen der Häufigkeiten der Noten (ohne Teilnoten).



häufiger vor als die größeren. Die Daten wurden für einen "wahren" Wert  $\mu = 15$  und einer "wahren" Streuung  $\sigma = 5.55$  generiert<sup>5</sup>. Für die Verteilung der Noten muß nun eine Unterteilung der Punktwerte um den Mittelwert  $\bar{x}$  vorgenommen werden; Schwierigkeiten bereitet aber die Asymmetrie der Verteilung der Punktwerte, die durch die relativ große Standardabweichung  $\sigma = 5.55$  bedingt wird. Um eine "richtige" Normalverteilung zu erhalten müßte die Streuung deutlich kleiner sein, was eine andere Verteilung der Aufgabenschwierigkeiten erfordern würde, d.h. die Unterschiede zwischen den Aufgabenschwierigkeiten müßten kleiner sein. Hat man aber eine Klassenarbeit schreiben lassen, muß man die Daten so nehmen, wie sie sind ... Man beachte die Unterschiede zwischen den Häufigkeitsverteilungen (a) und (b) sowie zwischen (c) und (d) für die die Populationsparameter  $\mu$  und  $\sigma^2$  jeweils identisch sind.

Nimmt man an, dass die Punkteverteilung eine Stichprobe aus einer Gauss-

<sup>5</sup>Die Stichproben wurden mit dem R-Paket `rnorm` erzeugt, das auch gleich die bestpassenden Gauß-Dichten mitliefert.

verteilten Grundgesamtheit ist, so korrespondiert zu  $x$  in (2.10) ein Punktwert  $S$ . Auf die Details der Schätzung von  $\mu$  und  $\sigma$  muß hier nicht eingegangen werden. Von der geschätzten Dichte (2.10) muß man nun zur korrespondierenden Gauß-Verteilung für die Noten übergehen. Dazu kann man annehmen, dass die Noten Werte auf einem Kontinuum von  $y$ -Werten zwischen 0 und 7 liegen. Die Noten 1 bis 6 markieren bestimmte Abschnitte auf diesem Kontinuum. Die Frage ist nun, welchen Erwartungswert  $\mu_y$  und welche Varianz  $\sigma_y^2$  die  $y$ -Werte haben sollen. Nimmt man an, dass  $\mu_y$  der mittleren Note (dem Mittelwert der Noten 1 bis 6) entsprechen soll, so wird man  $\mu_y = 3.5$  setzen.  $\sigma_y$  muß so gewählt werden, dass die Wahrscheinlichkeit, einen  $y$ -Wert kleiner als 0 oder größer als 6.5 oder 7 *praktisch* gleich Null ist. Für die Gauß-Verteilung gilt, dass eine lineare Transformation der Gauß-verteilten zufälligen Veränderlichen  $X$  wieder Gauß-verteilt ist; setzt man also  $Y = aX + b$ , so erhält man

$$f_Y(y) = \frac{1}{\sigma_y \sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(y - \mu_y)^2}{\sigma_y^2}\right), \quad \mu_y = a\mu_x + b, \quad \sigma_y = a\sigma_x a. \quad (2.11)$$

(s. Anhang für die Herleitung dieser Formel). Man kann nun  $\mu_y = 3.5$  setzen, weil 3.5 als arithmetische Mittel der natürlichen Zahlen 1 bis 6 ist, – aber natürlich kann auch ein anderer Wert für  $\mu_y$  angenommen werden.  $\sigma_y$  und damit der Parameter  $a$  in der Transformationsgleichung  $Y = aX + b$  muß in jedem Fall so bestimmt werden, dass die Wahrscheinlichkeit von  $y$ -Werten kleiner als 0 und größer als 6.5 *praktisch* gleich Null ist.

Durch die Wahl der Parameter  $a$  und  $b$  kann man sowohl den Mittwert (in der Abbildung 4 ist er die Note "3") als auch die Varianz relativ frei wählen (man muß nur die Einschränkung beachten, dass keine Werte kleiner als Null und größer als etwa 6.5 eine substantiell von Null abweichende Wahrscheinlichkeit haben), und mit dieser Wahl legt man die Häufigkeiten der Noten fest. Aus dieser Verteilung lassen sich die Intervallgrenzen für die Punkteverteilung bestimmen. Der Note "1" entspricht das Intervall  $Y \leq 1.5$  und ein Punktwert  $S \geq s_1$ , und es gilt  $p_1 = P(Y \leq 1.5) = P(S \geq s_1) = q_1$ , und wegen der Symmetrie der Gauß-Dichte gilt  $q_1 = 1 - p_1$ . Ist  $Z$  standardnormalverteilt, so gilt mit  $F(z) = P(Z \leq z)$  die Beziehung  $p_1 = F(z_1)$ , so dass  $F^{-1}(p_1) = z_{01} = (s_{01} - \bar{s})/s_s$ , d.h.  $s_{01} = sz_1 + \bar{s}$ , und  $s_1 = -s_{01}$  wegen  $q_1 = 1 - p_1$ . Analog dazu bestimmt man  $s_2$  mit  $P(Y \leq s_2) = P(Y \geq s_2)$  etc. Für  $S_i \geq s_1$  bekommt der Proband  $Pb_i$  eine "1", für  $S_i \in (s_1, s_2)$  eine "2", etc.

Zur Konstruktion der Notenverteilung sind die Anmerkungen

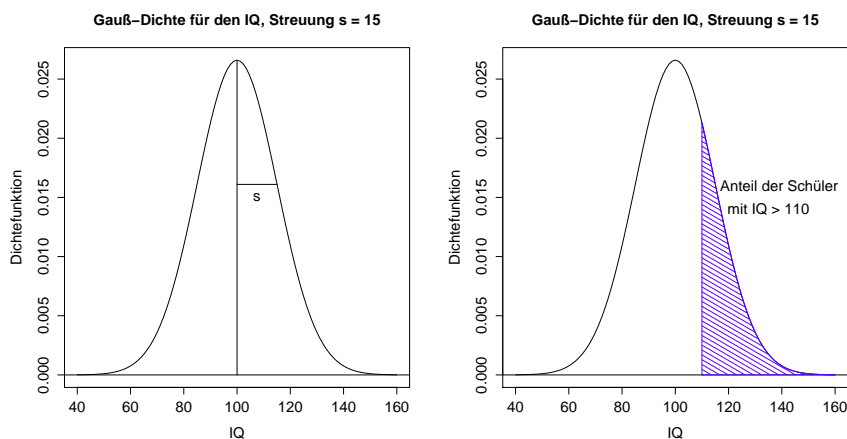
<http://matheplanet.com/default3.html?call=viewtopic.php?topic=192487&ref=https>

von Interesse. In Abbildung 4 sind die Intervalle für die Noten als gleich groß angenommen worden, d.h. man hat eine Intervallskala für die Noten konstruiert. Es gibt Alternativen: an der Universität Münster wurden z.B. Diplomnoten auf der Basis eines Kommissionsbeschlusses wie folgt ermittelt: In den einzelnen Fächern des Diploms wurden die Noten nach Gusto der jeweiligen Prüfer vergeben. Dann wird für eine(n) KandidatIn zunächst das arithmetische Mittel der Noten

in den Einzelfächern gebildet. Anschließend wurden Mittelwerte bis 1.7 als "sehr gut" (um-)gewertet, d.h. es wurde die Note "1" vergeben. Mittelwerte im Bereich  $1.7 < \bar{x} \leq 2.5$  wurden als "gut" umskaliert, d.h. es wurde die Note "2" vergeben, etc. Man kann eine derartige Regelung auch auf der Basis der Punkteverteilung einführen, denn den Kategoriengrenzen für die Noten entsprechen ja bestimmte Kategoriengrenzen für die Punkteverteilung. Natürlich wird damit die Symmetrie der Notenvergabe ("gute" Noten haben dieselbe Wahrscheinlichkeit wie die entsprechenden "schlechten" Noten) aufgehoben. Man kann die Symmetrie wieder herstellen, indem man die Intervalle für die "1" und "6" gleichgroß, aber größer als für die Noten "2" und "5" definiert und die für die Noten "3" und "4" wiederum gleichgroß, aber als kleinste Intervalle definiert. Das mag dem Geist der Annahme einer Normalverteilung widersprechen, aber das Münsteraner Beispiel zeigt, dass die Wahl der Intervallgrößen letztlich eben doch auf Annahmen beruht. Ein Argument für derartige ungleich große Intervalle ist zum Beispiel, dass ein Proband der Gauß-Verteilung entsprechend "sehr gut" sein kann, dass aber kleine Fehler sich nur negativ auswirken können, also notwendig häufiger zu einer Verschlechterung der Repräsentation der wahren Leistungsfähigkeit des Probanden führen. Zum Ausgleich müsse man dann, so die Argumentation, den Bereich, in dem eine "1" vergeben wird, vergrößern. Ob bei geringer Kompetenz Fehler nur eine Verbesserung der Repräsentation der Leistungsfähigkeit bewirken, sei dahin gestellt. Es ist bekannt, dass bei freier, also nicht auf einer Punkteverteilung beruhenden Benotung im Allgemeinen keine Intervallskala der Noten zugrunde liegt: Lehrende teilen den Notenbereich nach subjektiven Kriterien in Intervalle auf, wobei oft die mittlere Note, etwa "befriedigend" häufiger gewählt wird, als es einer Gauß-Verteilung mit gleich großen Intervallen entspricht. Dies soll ein Grund dafür sein, dass die Note "6" überhaupt noch gewählt werden kann, obwohl doch schon eine "5" ein "ungenügend" repräsentiert: die mittlere Note ist dann 3.5, und die Lehrenden müssen sich dann zwischen einer "3" und einer "4" entscheiden; der Neigung, eine "central tendency"-Note zu vergeben, soll damit entgegengewirkt werden.

**Relativierung durch Normierung** Im Prinzip läßt sich jede Punkteverteilung auf eine vorgegebene Notenverteilung transformieren. Aber genau deswegen stellt sich die Frage nach dem Sinn einer solchen Transformation; die Frage nach der Sinnhaftigkeit der Annahme der Gauß-Verteilung ist nicht von dieser Frage abhängig. Das Leistungsspektrum in zwei Parallelklassen kann aus verschiedenen Gründen unterschiedlich sein; die durchschnittliche Begabung ("Intelligenz") in den Klassen kann verschieden sein, verschiedene LehrerInnen können einen unterschiedlichen Effekt auf die Leistungen haben, etc. Zur Illustration betrachte man die möglichen Begabungsunterschiede. Es werde der Einfachheit halber angenommen, dass die SchülerInnen in einem Gymnasium einen Mindest-IQ haben müssen, um das Gymnasium erfolgreich abschließen zu können, es gelte etwa  $IQ > 110$ , vergl. Abbildung 5, wo die übliche Annahme einer Gauß-Verteilung der Intelligenzquotienten adoptiert wurde. Die Verteilung der Intelligenzquoti-

Abbildung 5: Beispiel für eine Gauß-Dichte: die Verteilung von Intelligenzquotienten. Links: Illustration des  $s$ -Werts, rechts: Anteil der Schüler mit einem IQ > 110 (gestrichelte Fläche unter der Dichtefunktion)



enten an einem Gymnasium ist dann mit großer Wahrscheinlichkeit nicht mehr Gauß-verteilt, da die IQen der SchülerInnen eben aus einer Teilpopulation mit einem IQ > 110 kommen. Man spricht von einer *trunkierten*<sup>6</sup> Gauß-Verteilung. Abbildung 6 illustriert diesen Sachverhalt. Die Stichprobengröße  $n = 30$  bezieht sich auf die Anzahl von SchülerInnen in einer Klasse. Der Vergleich des Falles B mit den Fällen C und D zeigt, wie unterschiedlich die Begabungen in zwei Klassen verteilt sein können, obwohl sie Stichproben aus derselben Grundgesamtheit sind. Statt des Merkmals Intelligenz kann man auch ganz allgemein den Begriff "Leistungsfähigkeit" einsetzen, die ja auch von den jeweiligen LehrerInnenn mit beeinflusst sein kann.

Der Begriff der Intelligenz, repräsentiert durch den IQ, ist ein *Konstrukt*, d.h. unter 'Intelligenz' wird das verstanden, was ein Intelligenztest mißt und was als Intelligenzquotient (IQ) angegeben wird<sup>7</sup> ist in der Gesamtpopulation normalverteilt, mit einem Populationsmittelwert von 100 und einer Streuung  $\sigma = 15$ .

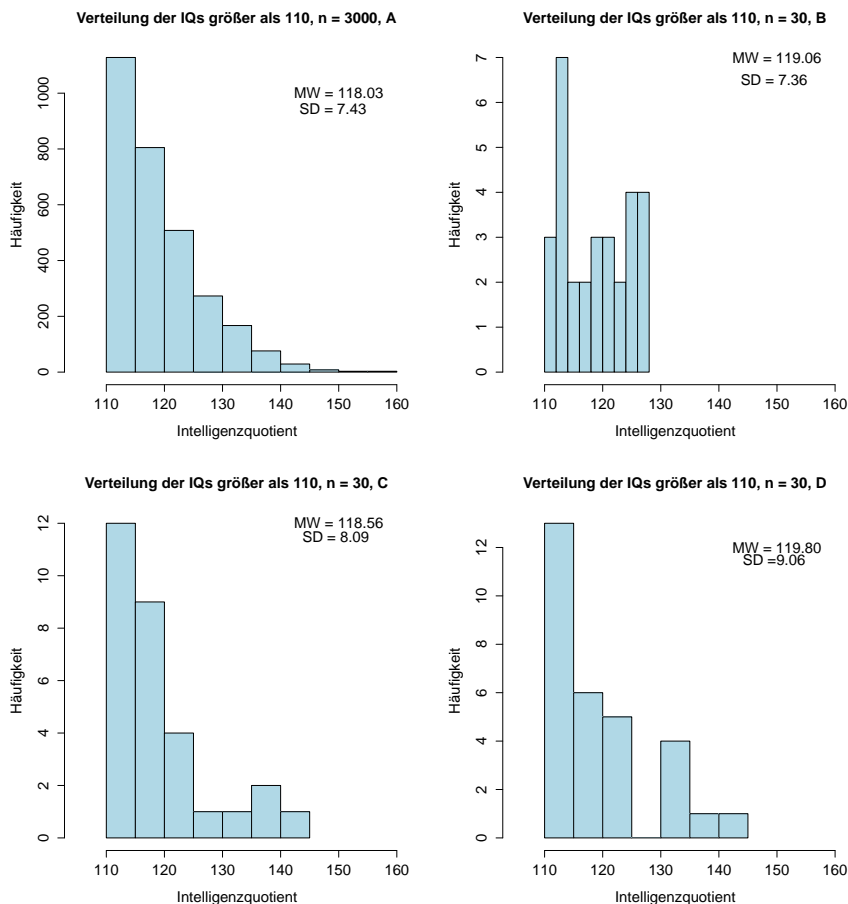
Weiterführende Schulen, insbesondere das Gymnasium, setzen i. A. einen Mindest-IQ voraus. Für ein Gymnasium kann ein Mindest-IQ von 110 angenommen werden; werden fast 50% eines Jahrgangs zum Gymnasium zugelassen, so wird man eher einen Mindest-IQ von 105 annehmen müssen. Die folgenden Be-

<sup>6</sup>wohl aus dem Englischen *to truncate* - kürzen, abschneiden etc

<sup>7</sup>Der IQ ist ein gewogener Mittelwert von Punktwerten ("Scores") für verschiedene kognitive Fähigkeiten; rechnerisches Denken, analytisches Denken, sprachliche Ausdrucksfähigkeit, etc gehen mit unterschiedlichen Gewichtungen in das IQ-Maß ein. Ob der IQ ein sinnvolles Maß ist, ist umstritten, viele Diagnostiker betrachten statt des IQ das Profil der Leistungen in den Untertests, weil es mehr Information über spezifische Begabungen bzw. Defizite liefert.



Abbildung 6: Stichprobenverteilungen der trunkierten Gauß-Verteilung. Fall A:  $n = 3000$  Fälle, die Verteilung zeigt in guter Näherung die Verteilung der Populationswerte; sie entsprechen der rechten Seite der Dichte in Abb. 5. Die Fälle B bis D zeigen mögliche Verteilungen für *kleine* Stichproben ( $n = 30$ ); die Häufigkeiten können stark von denen in der Population abweichen.



trachtungen beziehen sich beispielhaft auf einen Mindest-IQ = 110. Die Annahme eines Mindest-IQ bedeutet in jedem Fall, dass die Verteilung der IQen an einem Gymnasium *nicht* mehr normalverteilt ist, denn die Schüler stellen dann eine Stichprobe aus einer Grundgesamtheit dar, die aus der Teilmenge aller Personen mit einem IQ größer als der, mindestens aber gleich dem Mindest-IQ ist. Die Verteilung der IQen in dieser Teilmenge folgt einer *trunkierten Normalverteilung*, d.h. einer Verteilung, deren unteres Ende ( $IQ < \text{Mindest-IQ}$ ) abgeschnitten ist. Abbildung 6, (A) zeigt die Verteilung der IQen in einer Stichprobe vom Umfang 3000, – die Verteilung repräsentiert in guter Näherung die Population. Man

sieht, dass der durchschnittliche IQ in dieser Teilpopulation bei 118 liegt; dieser Wert entspricht in guter Näherung dem Durchschnittswert der studentischen Population. Die Abbildungen B, C und D zeigen mögliche Verteilungen des IQ in einzelnen Klassen, also Verteilungen von Stichproben aus dieser Population von im Durchschnitt  $n = 30$  Schülern.

Nimmt man (vernünftigerweise) an, dass der  $\mu$ -Wert einer Person hauptsächlich durch den IQ der Person bestimmt wird, so wird man eine Häufung von  $\mu$ -Werten im unteren Bereich haben und eher wenige  $\mu$ -Werte im oberen Bereich, d.h. die Verteilung der  $\mu$ -Werte wird i.A. asymmetrisch sein. Diese Eigenschaft widerspricht der Annahme einer Normalverteilung, die grundsätzlich symmetrisch ist, wie die Glockenkurve ja schon indiziert. Selbst wenn man von einer Normalverteilung der Gesamt"fehler"  $\varepsilon$  ausgehen kann (sie können symmetrisch um Null verteilt sein), macht es vermutlich wenig Sinn, von der Normalverteilung der individuellen Leistungen  $X_i$  auszugehen, – eben wegen der Nicht-Normalverteiltheit der  $\mu_i$ .

Aus der Tatsache, dass der belgische Astronom und Statistiker Adolphe Quetelet fand, dass die Brustumfänge französischer Rekruten ebenso wie andere biologische Größen in sehr guter Näherung Gauß-verteilt sind (s. den kurzen historischen Abriss zur Gauß-Verteilung im Anhang), so folgt daraus noch lange nicht, dass auch kognitive Leistungen "normalverteilt" sind. Eine wesentlich plausiblere Verteilung der Leistungsfähigkeit der Schüler ist die Log-Normalverteilung, die im Folgenden kurz besprochen wird.

## 2.3 Die Log-Normalverteilung

Die Log-Normalverteilung entsteht, wenn  $y$ -Werte gemessen werden, die über die Exponentialfunktion mit normalverteilten  $X$ -Werten verbunden sind, d.h. wenn  $y = e^x$  gilt; offenbar ist dann  $\log y = x$  normalverteilt. Man findet diese Verteilung in vielen Bereichen, u.a. in der Biologie, der Ökonomie, etc. Die Frage ist, welchen Mechanismus man hinter dieser Verteilung vermuten kann. Es werden die beiden hauptsächlich diskutierten Modelle vorgestellt. Verschiedene Log-Normalverteilungen über einem Leistungsbereich werden in Abbildung 7 gezeigt. Man sieht, dass für bestimmte Parameterwerte eine Log-Normalverteilung einer Normalverteilung ähnlich wird. Charakteristisch bleibt allerdings eine Asymmetrie: kleinere Punktwerte sind häufiger als größere Punktwerte.

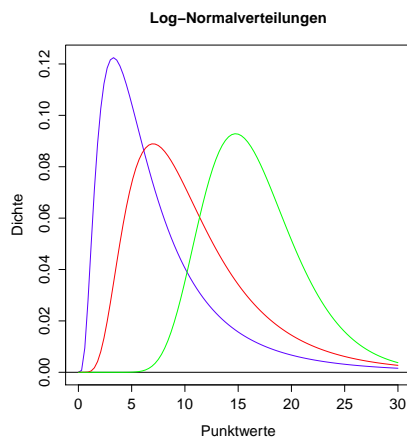
### 2.3.1 Das Shockley-Modell

Shockley<sup>8</sup> wunderte sich darüber, dass viele Leistungen log-normalverteilt sind: Die Anzahl von Publikationen von Wissenschaftlern ebenso wie die Größe von Firmen, die Häufigkeit einer Species in einer biologischen Umgebung, die Häufig-

---

<sup>8</sup>William Bradford Shockley (1910 – 1989), Physiker, Erfinder des Transistors (Nobelpreis)

Abbildung 7: Verschiedene Log-Normaldichten über einem Bereich möglicher Punktwerte



keit von Buchstaben in einzelnen Wörtern ebenso wie die Häufigkeit von Wörtern in verschiedenen Sätzen, ja, auch die Häufigkeiten von Poren in Filtermembranen sind log-normalverteilt. Die Körpergröße von Menschen scheint normalverteilt zu sein, aber die Verteilung von Körpergrößen kann ebenso gut als Spezialfall einer Log-Normalverteilung beschrieben werden. Die Beispiele kommen aus allen Wissenschaften (Limpert et al, 2008).

Shockley nahm an, dass sich die Komponenten  $K_j$  einer Leistung nicht additiv (wie bei der Normalverteilung) überlagern, sondern dass sie als Konjunktion zusammenwirken müssen, um die Gesamtleistung zu erzeugen, d.h. es muß die Kombination

$$K_1 \wedge K_2 \wedge \dots \wedge K_n$$

der Leistungen  $K_j$ ,  $j = 1, \dots, n$ , erbracht werden ( $\wedge$  steht für "und"). Will ein Schüler eine gute Klassenarbeit in Mathematik schreiben, so muß er zunächst den im Unterricht behandelten Stoff beherrschen ( $K_1$ ), er muß imstande sein, das Erlernte auf neue Aufgabenstellungen anzuwenden ( $K_2$ ), dazu muß er u.U. imstande sein, gewisse Kopfrechenleistungen zu erbringen ( $K_3$ ), auch sollte er zu gewissen logischen Operationen imstande sein ( $K_4$ ), etc. Dass er dann während der Klassenarbeit die Leistung  $K_j$  erbringt, habe die Wahrscheinlichkeit  $p_j$ . Ist  $p_j = 1$ , so liefert er die Leistung (fast) mit Sicherheit<sup>9</sup>. Weiter wird angenommen, dass diese Kompetenzen statistisch unabhängig voneinander sind. Diese Annahme stellt vermutlich eine Vereinfachung dar, es läßt sich aber zeigen, dass es hinreichend ist, wenn die Unabhängigkeitsannahme zumindest angenähert gilt (die Rede ist

<sup>9</sup>Tritt ein Ereignis mit Siherheit ein, so hat es die Wahrscheinlichkeit 1. Ein Ereignis muß aber nicht eintreten, nur weil man ihm die Wahrscheinlichkeit 1 zugeordnet hat ...

dann von *asymptotischer Unabhängigkeit*). Die Unabhängigkeitsannahme impliziert dann, dass die Wahrscheinlichkeit, dass die Prüfung gemeistert wird, durch das Produkt

$$p = p_1 \times p_2 \times \cdots \times p_n = \prod_{j=1}^n p_j, \quad (2.12)$$

gegeben. Für gute Schüler sind die  $p_j$ -Werte eher größer, für schlechte sind zumindest einige der  $p_j$ -Werte eher kleiner. Da wegen der Zufallsauswahl der Schüler in einer Klasse auch die individuellen Fähigkeiten zufällig variieren, sind auch die  $p_j$ -Werte und damit auch der  $p$ -Wert eines Schülers zufällige Veränderliche.

Der natürliche Logarithmus ist die Umkehrung der  $e$ -Funktion: ist  $y = e^x$ , so ist  $x = \log_e x = \log x$  (zur Vereinfachung wird im Folgenden  $\log$  statt  $\log_e$  geschrieben). Dann gilt die zunächst etwas triviale Gleichung

$$x = e^{\log x},$$

denn dann ist ja  $\log x = \log e^{\log x} = \log x$ . Weiter gilt allgemein, dass der Logarithmus eines Produkts  $a \times b$  gleich der Summe der Logarithmen ist, d.h.

$$\log(a \times b) = \log a + \log b.$$

Angewendet auf (2.12) folgt

$$p = e^{\log p} = e^{\log p_1 + \log p_2 + \cdots + \log p_n}. \quad (2.13)$$

Da die  $p_i$  zufällige Veränderliche sind, ist auch der Logarithmus

$$x = \log p = \log p_1 + \log p_2 + \cdots + \log p_n$$

eine zufällige Veränderliche. Auf die Summe  $x$  ist aber wieder der Zentrale Grenzwertsatz anwendbar, so dass angenommen werden kann, dass  $x$  in guter Näherung normalverteilt ist, und das heißt, dass  $p$  eine Zufallsverteilung hat, die durch

$$p = e^x \quad (2.14)$$

mit normalverteiltem  $x$  gegeben ist. Die Verteilung der  $p$  ist dann die *Logarithmische Normalverteilung*, auch *log-Normalverteilung* genannt. Die Formel findet man im Anhang. Hier ist nur von Bedeutung, dass die entsprechende Funktion (die Wahrscheinlichkeitsdichte) im Unterschied zur Normalverteilung *nicht* symmetrisch ist. Die häufigsten Leistungswerte finden sich im unteren Bereich, große und sehr große Werte treten im Vergleich dazu selten auf. Bei der Normalverteilung treten große und sehr große Werte mit der gleichen Wahrscheinlichkeit auf wie die entsprechenden kleinen und sehr kleinen Werte.

Die log-Normalverteilung hat mindestens die gleiche Bedeutung in den Anwendungen wie die Normalverteilung. Praktisch alle Leistungen – ökonomische, wissenschaftliche (gemessen an der Zahl der Veröffentlichungen), künstlerische,

sportliche etc etc folgen der log-Normalverteilung; ebenso folgen viele biologische Größen der log-Normalverteilung. Eine der bekanntesten Anwendungen dieser Verteilung ist die auf die Größe von Einkommen bzw Vermögen: das Einkommen ist praktisch in allen Nationen log-normalverteilt, – zumindest im Bereich des ”normalen” Einkommens. Für die ganz großen Einkommen bzw. Vermögen geht die Verteilung in die Pareto-Verteilung über, so dass die Einkommensverteilung insgesamt eine *Mischung von Verteilungen* ist. Analoge Betrachtungen können für Schulleistungen gelten.

### 2.3.2 Das Gibrat-Modell

Ein alternativer Vorschlag über den der log-Normalverteilung zugrundeliegenden Mechanismus kommt von dem französischen Ingenieur Robert Gibrat (1904–1980), der von der Annahme

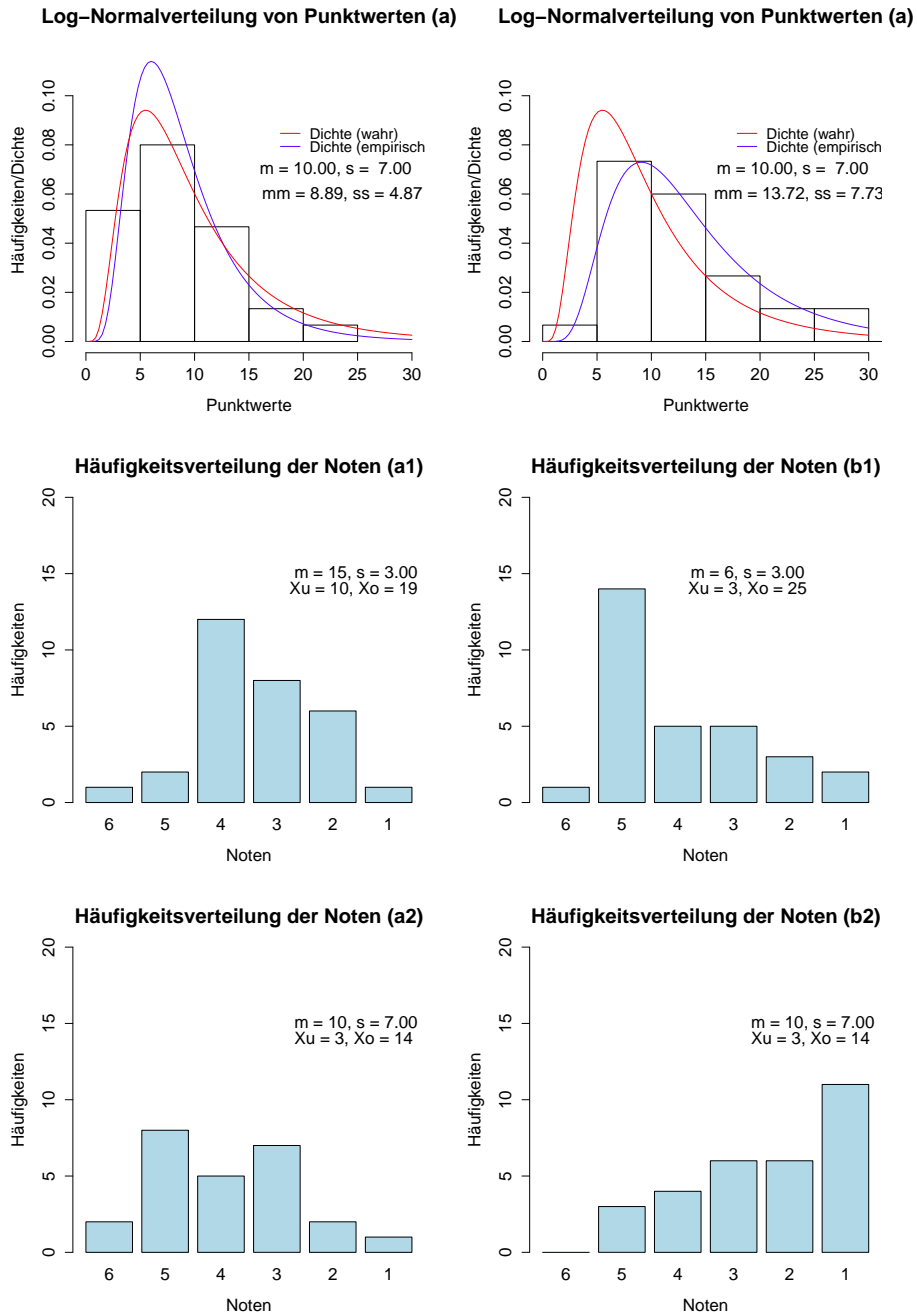
$$x_t - x_{t-1} = \varepsilon_t x_{t-1} \quad (2.15)$$

ausging; hier bedeuten  $x_t$  und  $x_{t-1}$  die Größen einer Firma zum Zeitpunkt  $t$  bzw.  $t - 1$  und  $\varepsilon_t$  ist die zufällig variierende Wachstumsrate zum Zeitpunkt  $t$ . Es läßt sich zeigen, dass die Firmengrößen dann – wie in der Realität – log-normalverteilt sind. Bei einer Anwendung des Modells auch schulisches Lernen würde man (2.15) zufolge annehmen, dass der Wissens- oder Kompetenzunterschied zwischen zwei Zeiten (sinnvollerweise 1 Monat, 1 halbes Jahr, oder so ähnlich) proportional zu  $x_{t-1}$  ist, mit der zufälligen Variablen  $\varepsilon_t$  als Proportionalitätsfaktor. Ob dieses Modell ein plausibles Modell für das Erlernen von Fähigkeiten in der Schule ist müßte noch diskutiert werden. Das Modell wird hier in erster Linie erwähnt, um den Sachverhalt zu illustrieren, dass es für Wahrscheinlichkeitsverteilungen oft verschiedene ”Mechanismen” gibt, die der Verteilung zugrunde liegen können.

### 2.3.3 Illustration des log-Normal-Modells

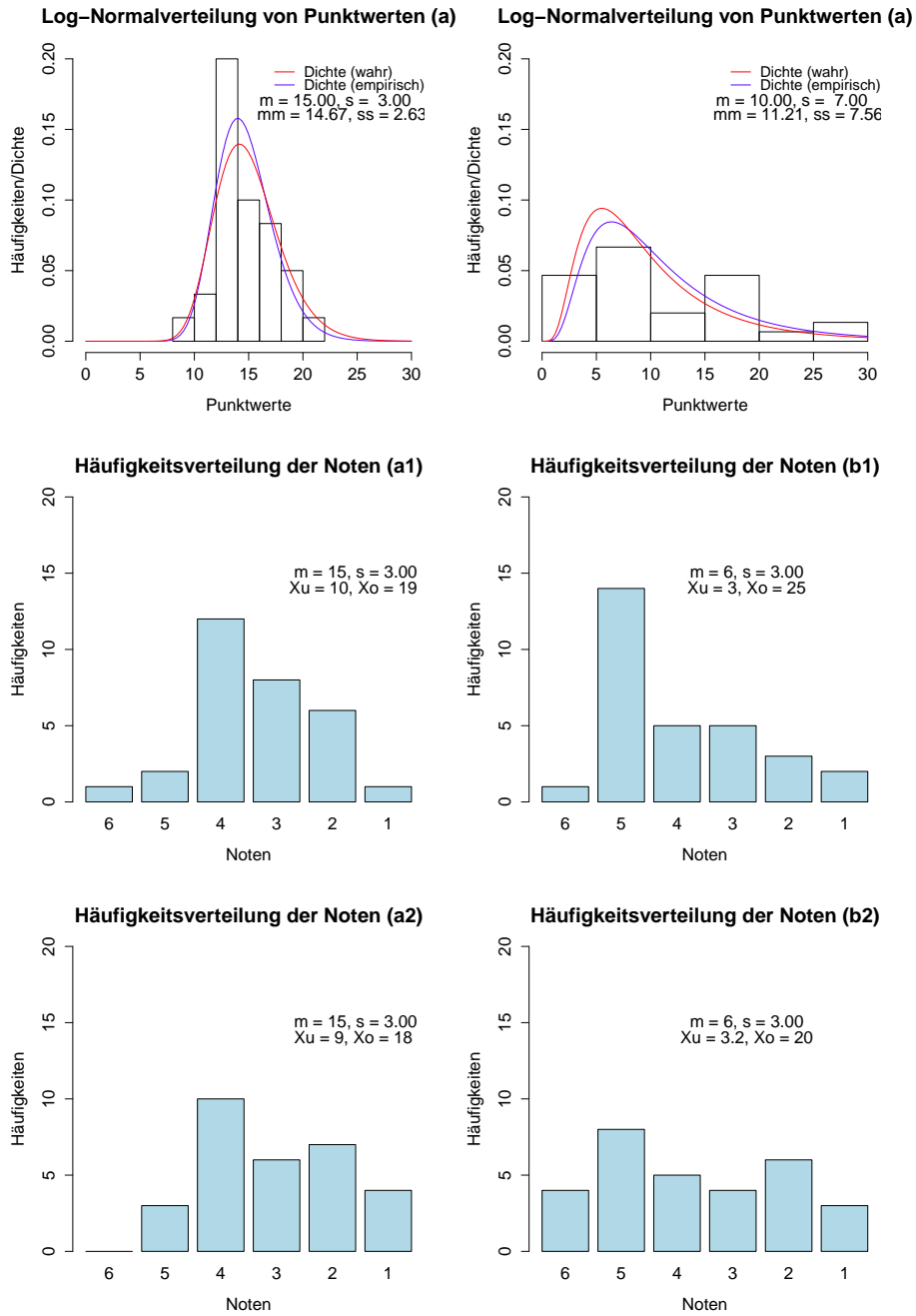
Die Abbildungen 8 und 9 zeigen in den obersten zwei Grafiken (i) eine ”wahre” log-Normalverteilung (rot), mit den Parametern  $\mathbb{E}(X) = m$  und der Standardabweichung  $\sigma = s$  (ii) eine Häufigkeitsverteilung, die sich als Zufallsstichprobe vom Umfang  $N = 30$  aus der durch die wahre Verteilung definierten Population ergibt, und eine log-Normalverteilung (blau), deren Parameter  $mm$  (Erwartungswert) und Streuung  $ss$  (Standardabweichung) der Stichprobenmittelwert und die Stichprobenstreuung sind.  $\sigma$  wird so gewählt, dass die Anzahl der Punkte den Wert 30 nicht übersteigt; dies entspricht einer Klausur mit maximal 30 Fragen. Die Übereinstimmung mit dem Stichprobenumfang ist Koinzidenz. Die roten und blauen Kurven zeigen, wie groß die Unterschiede zwischen der wahren und der Stichprobenverteilung sind. Die beiden oberen Grafiken zeigen zwei verschiedene Stichproben aus ein und derselben Population. Die Verteilungen in Abbildung 8 entsprechen einer ”schwierigen” Klausur – der mittlere Punktwert ist eher niedrig,

Abbildung 8: Log-normale Punkteverteilung I neu



während die Abbildung 9 eine mittelschwere Klausur repräsentiert; hier würde ei-

Abbildung 9: Log-normale Punkteverteilung II



ne Gauß-Verteilung ebenfalls gut passen.

Unter jeder dieser Grafiken werden zwei Notenverteilungen gezeigt, die sich beide aus derselben Stichprobenverteilungen ergeben. Die Notenverteilungen ergeben sich, indem ein unterer  $X_u$ ) und ein oberer  $X_o$ ) Punktwert gewählt wird und eine Einteilung in vier Unterabschnitte  $d = (X_o - X_u)/4$  berechnet wird. Einem Score  $X < X_u$  wird die Note "6", zu geordnet, einem Score  $X \in [x_u, X_u + d)$  die Note "5", für  $X \in [X_u + d, X_u + 2d)$  ergibt sich die Note "4", etc, und für  $X \geq X_o$  ergibt sich die Note "1". Die beiden Notenverteilungen ergeben sich durch verschiedene Wahlen von  $X_u$  und  $X_o$ .

Die wesentliche Aussage der beiden Abbildungen ist der Befund, dass sich die Notenverteilungen in Abhängigkeit von Stichprobeneffekten einerseits und der mehr oder weniger willkürlich gewählten  $X_u, X_o$ -Werte stark unterscheiden. Natürlich kann man auch hier den Eindruck erzwingen, die Noten seien Gauß-verteilt: dazu müssen aber die Intervallgrenzen geeignet gewählt werden und möglicherweise eine Gewichtung der einzelnen Aufgaben vorgenommen werden, wie anfangs schon angemerkt. Eine alternative Wahl der Intervallgrenzen würde aber die Eigenschaft der Intervallskala zerstören, d.h. die einzelnen Noten würden verschieden große Fähigkeitsbereiche definieren; man kann dann argumentieren, dass die Noten kaum interpretierbar sind. Würde man von einer Gauß-Verteilung der Scores ausgehen, würde man zu einem analogen Befund kommen. Die Log-Normalverteilung hat den Vorteil, auch für "schwierige" Klausuren eine vernünftige Grundverteilung zu sein, da sie nur für Scores größer oder gleich Null definiert ist. Damit die Anwendung der trunkierten Gauß-Verteilung umgangen werden kann erfordert eine schwierige Klausur korrespondierend kleine Standardabweichungen, um den Bereich der erreichten Scores abzudecken, – empirisch ist eine kleine Standardabweichung aber oft nicht zu rechtfertigen. Die log-Normalverteilung erklärt leicht den Fall hauptsächlich kleinerer Scores und relativ wenigen großen Scores.

### 3 Das Rasch-Modell

#### 3.1 Das Modell

Ein zentrales Problem der klassenspezifischen Renormierung der Ergebnisse etwa einer Klassenarbeit besteht in der impliziten Neubewertung von Aufgaben von "leicht" bis "schwer" nach Maßgabe der jeweils vorgefundenen Punkteverteilung. Dadurch werden die unvermeidlichen Zufälligkeiten in den vergebenen Punkten überbewertet: im Extremfall sind alle Schüler gleich kompetent, werden aber auf der Basis der zufälligen Unterschiede zwischen den Punktwerten mit den Noten 1 bis 5 bedacht. Darüber hinaus wird die Vergleichbarkeit der Noten zerstört: so reflektiert eine 2 ("gut") eben nicht mehr eine "gute" Kompetenz in allen Klassen und Schulen, – ein "gut" in einer Klasse kann im Prinzip einem "ausreichend" in einer anderen Klasse entsprechen.



Das Problem ist schon lange in der psychologischen Diagnostik bekannt. In der in den ersten Jahrzehnten des 20-ten Jahrhunderts entwickelten Klassischen Testtheorie (KTT), der Theorie psychometrischer Tests, gelten Aufgaben als "schwer", wenn nur wenige in der Population sie korrekt beantworten können, dagegen als "leicht", wenn viele die richtige Antwort geben können. Vergleiche z.B. der durchschnittlichen Intelligenz in verschiedenen Bevölkerungsgruppen (z.B. sozioökonomische Gruppen) führen dann schnell zu Fehleinschätzungen, zumal auch hier die Gauß-Verteilung als angeblich natürliche Verteilung der Ausprägungen von Merkmalen für deren Bewertung zugrunde gelegt wurde.

Ideen wie die der Gauß-Verteilung als natürliche Verteilung für Merkmalsausprägungen können sich in einer Forschergemeinde lange halten, auch wenn sie keine guten Ideen sind. Die meisten Forscher werden in ihre jeweiligen Gebiete über das Studium des Faches, in dem das Gebiet liegt eingeführt und übernehmen die stillschweigend gemachten Annahmen eben als natürliche Annahmen. Erst wenn jemand aus einem anderen Fach sich dem Gebiet – hier dem Gebiet der Messung von Persönlichkeitsmerkmalen – zuwendet kann es geschehen, dass sie oder er die Standardannahmen nicht mehr als gottgegeben einschätzt. Diese Rolle übernahm der dänische Statistiker Georg Rasch (1901 – 1980), als er 1960 und 1961 forderte, dass Messungen unabhängig von den jeweiligen Bedingungen und Messinstrumenten sein sollten. Er schlug ein *probabilistisches* Modell der Messung vor, das dieser Forderung genügt, wobei angemerkt werden muß, dass die Bezeichnung 'probabilistisches' Modell insofern irreführend ist, als auch der klassische Ansatz der KTT probabilistisch ist, denn dort wird ja ebenfalls angenommen wird, dass die Leistung eine zufällige Komponente enthält ( $x = \mu + \varepsilon$ ). Bezeichnungsfragen müssen hier aber nicht weiter diskutiert werden, statt dessen soll kurz der Rasch-Ansatz vorgestellt werden.

Der Einfachheit halber werden *dichotome Aufgaben* betrachtet; das sind Aufgaben, deren Lösung entweder mit "richtig" oder "falsch" beurteilt werden kann. Zur Kennzeichnung wird eine *Indikatorvariable* eingeführt:  $X_g \in \{0, 1\}$ , wobei der Index  $g$  anzeigt, dass  $X_g$  die Antwort auf die  $g$ -te Aufgabe anzeigt:  $X_g = 0$ , wenn die  $g$ -te Aufgabe nicht oder falsch beantwortet wurde, und  $X_g = 1$ , wenn sie korrekt beantwortet wurde, und  $g = 1, \dots, n$ , wenn es  $n$  Aufgaben gibt. Weiter sei  $\theta$  ein Maß für die Kompetenz oder Fähigkeit einer oder eines Probanden.

Rasch geht davon aus, dass Aufgaben mit einer bestimmten, vom  $\theta$ -Wert eines Probanden abhängenden Wahrscheinlichkeit korrekt beantwortet werden, d.h. es soll  $0 \leq P(X_g = 1|\theta) \leq 1$  gelten. Weiter muß Rasch annehmen, dass die Beantwortung der verschiedenen Aufgaben für fixen Wert von  $\theta$  stochastisch unabhängig sind. Nach der elementaren Wahrscheinlichkeitstheorie bedeutet dies, dass die Wahrscheinlichkeit von einer Folge von Werten  $X_1, X_2, \dots, X_n$  (wobei ein beliebiges  $X_g$  den Wert 0 oder 1 annehmen kann) durch das Produkt der Wahrscheinlichkeiten  $P(X_g|\theta)$  gegeben ist:

$$P(X_1, X_2, \dots, X_n|\theta) = P(X_1|\theta)P(X_2|\theta) \cdots P(X_n|\theta). \quad (3.1)$$

Dies bedeutet, dass die Aufgaben dieser Forderung entsprechend ausgewählt werden müssen, d.h. die Beantwortung einer Aufgabe darf nicht die Beantwortung einer anderen voraussetzen.

**Definition 3.1** *Gilt für  $n$  Aufgaben die Gleichung (3.1), so erfüllen die Aufgaben die Bedingung der lokalen stochastischen Unabhängigkeit.*

Man spricht auch kurz von *lokaler Unabhängigkeit*. "Lokal" heißt dabei, dass die Unabhängigkeit jeweils für einen bestimmten Ort auf der Fähigkeitsskala, eben  $\theta$ , bezieht.

Rasch benötigt nun noch eine Funktion, die die Wahrscheinlichkeit für eine korrekte Antwort auf eine Aufgabe definiert, wenn die Fähigkeit des Schülers den Wert  $\theta$  hat, d.h. für  $P(X_g = 1|\theta)$ . Dieser Ausdruck muß so definiert sein, dass die Schätzung von  $\theta$  unabhängig von der Schätzung von  $\kappa_g$ , der Schwierigkeit der  $g$ -ten Aufgabe ist. Er findet ihn in der *logistischen Verteilung*<sup>10</sup>

$$F(x) = P(X \leq x) = \frac{1}{1 - \exp\left(-\frac{(x-\mu)\beta}{\sigma}\right)}, \quad \beta = \pi/\sqrt{3}. \quad (3.2)$$

und mit  $\gamma = \beta/\sigma$  erhält man

$$P(X \leq x) = \frac{1}{1 + \exp(-(x - \mu)\gamma)} \quad (3.3)$$

$\mu$  ist der Erwartungswert von  $X$ , d.h. der Mittelwert über alle möglichen Realisierungen von  $X$ , und  $\sigma$  ist die korrespondierende Standardabweichung. Interpretiert man  $X$  als die Leistungsfähigkeit eines Probanden zur Zeit der Aufgabenstellung, so kann man  $\theta = \mu$  setzen, d.h. als den Erwarteten Wert der Leistungsfähigkeit eines Probanden setzen. Die Wahrscheinlichkeit, dass der Proband die Aufgabe korrekt beantwortet, ist durch  $P(X > \kappa_g)$  gegeben, wobei  $\kappa_g$  die Schwierigkeit der  $g$ -ten Aufgabe ist; die Schwierigkeit ist damit als die minimale Fähigkeit definiert, die ein Proband haben muß, um die Aufgabe korrekt zu beantworten;  $\kappa_g$  ist also ein Wert auf derselben Skala wie  $\theta$ . Da

$$P(X > \kappa_g) = 1 - P(X \leq \kappa_g)$$

folgt

$$P(X > \kappa_g) = 1 - \frac{1}{1 + \exp(-(\kappa_g - \theta)\gamma)} = \frac{\exp(-(\kappa_g - \theta)\gamma)}{1 + \exp(-(\kappa_g - \theta)\gamma)}.$$

---

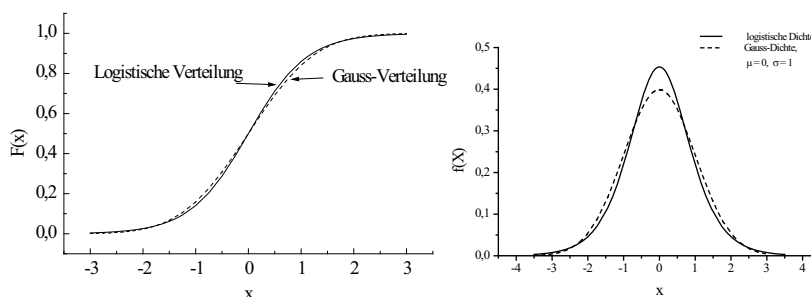
<sup>10</sup>Der Ausdruck 'logistisch' bezieht sich auf einen stochastischen Wachstumsprozess, den der belgische Mathematiker Pierre Verhulst (1801 – 1849) konzipiert hat, um im Auftrag der Pariser Stadtverwaltung den wegen des Bevölkerungswachstums benötigten Zuwachses an Wohnungen (frz. logis = Wohnung) abzuschätzen. Die Dichtefunktion der logistischen Verteilung ist durch  $f(x) = \gamma F(x)(1 - F(x))$  gegeben, und damit entspricht die rechte Seite formal dem Ausdruck  $N(t)(K - N(t))$  in Verhulsts Modell,  $N(t)$  die Anzahl der Wohnungen zur Zeit  $t$ ;  $\gamma$  ist eine Normierungskonstante.

Setzt man  $\kappa'_g = \gamma\kappa_g$ ,  $\theta' = \gamma\theta$  und benennt dann  $\kappa'_g$  und  $\theta'$  wieder in  $\kappa_g$  und  $\theta$  um (Reparametrisierung), so ergibt sich

$$P(X > \kappa_g) = \frac{\exp(\theta - \kappa_g)}{1 + \exp(\theta - \kappa_g)}. \quad (3.4)$$

**Anmerkung:** Für geeignet gewählte Parameter zeigen die Gauß- und die logistische Verteilung eine bemerkenswerte Ähnlichkeit; s. Abbildung 10. Sollte also die

Abbildung 10: Verläufe der standardisierten Gauß- und der logistischen Verteilung; Verteilungsfunktionen links, Dichten rechts.



logistische Funktion die "wahre" Verteilung sein, kann i.A. ebensogut eine Gauß-Verteilung an die Daten angepasst werden, – allerdings hat die Gauß-Funktion nicht die hier betrachteten Eigenschaften der logistischen Funktion.

**Definition 3.2** Es sei  $F_g(\theta) = P(X_g = 1|\theta) = P(X > \kappa_g)$ ;  $F_g(\theta)$  heißt Itemfunktion für die  $g$ -te Aufgabe.

Der Ausdruck 'Item' steht in der Theorie psychometrischer Tests für 'Aufgabe', 'Frage'. Abbildung 11 zeigt Itemfunktionen für verschiedene  $\kappa$ -Werte; die Itemfunktionen sind parallel, d.h. sie werden auf der  $\theta$ -Achse nur parallel verschoben. Die Definition des Rasch-Modells ist dann

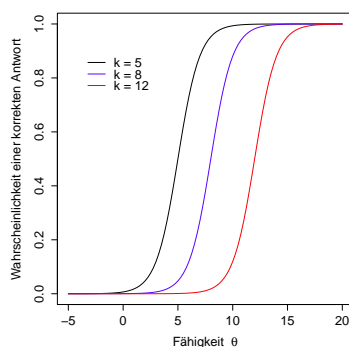
**Definition 3.3** Es sei  $X_g = \{0, 1\}$  eine Indikatorvariable definiert:  $X_g = 1$ , wenn die Aufgabe gelöst oder der Meinungsaussage zugestimmt wurde, und  $X_g = 0$  sonst. Die Itemfunktion sei durch

$$F_g(\theta) = P(X_g = 1|\theta) = \frac{e^{\theta - \kappa_g}}{1 + e^{\theta - \kappa_g}}, \quad (3.5)$$

gegeben und es gelte lokale Unabhängigkeit. Das Testmodell heißt dann Rasch-Modell.

$F_g(\theta)$  ist nur eine abkürzende Schreibweise für  $P(X_g = 1|\theta)$ . Die Wahrscheinlichkeit, die  $g$ -te Aufgabe, zu korrekt zu beantworten, hängt also von den Parametern

Abbildung 11: Itemfunktionen (Wahrscheinlichkeiten einer korrekten Antwort als Funktion der Fähigkeit  $\theta$ ) für verschiedene Schwierigkeiten  $\kappa$ ; je größer der Wert von  $\kappa$ , desto schwieriger die Aufgabe.



$\kappa_g$  und  $\theta$  ab. Dividiert man Zähler und Nenner der rechten Seite von (3.5) durch  $e^{\theta-\kappa_g}$ , d.h. multipliziert man Zähler und Nenner mit  $e^{-(\theta-\kappa_g)} = e^{\kappa_g-\theta}$ , so wird (3.5) zu

$$F_g(\theta) = \frac{1}{1 + e^{\kappa_g-\theta}}. \quad (3.6)$$

Man sieht leicht, dass  $F_g(\theta) \rightarrow 1$  für  $\theta \rightarrow \infty$  und  $F_g(\theta) \rightarrow 0$  für  $\kappa_g \rightarrow \infty$ . Hat man  $m$  Probanden (Schüler), so muß man  $\theta_1, \dots, \theta_m$  Fähigkeitsparameter schätzen, und zusätzlich  $n$  Schwierigkeitsparameter  $\kappa_1, \dots, \kappa_n$ .

Ein wichtiger Begriff, der in Bezug auf (3.5) eingeführt werden muß, ist der der *Rasch-Homogenität*: intuitiv ist damit gemeint, dass die Aufgaben (Items), die (3.5) genügen, alle nur ein und dasselbe Merkmal messen, dessen Ausprägung bei einer Person  $a$  eben durch  $\theta$  bzw.  $\theta_a$  repräsentiert wird. Die Rasch-Homogenität läßt sich auch formal definieren, und aus dieser Definition folgt dann (3.5):

**Definition 3.4** *Es seien  $X_g, X_h$  Indikatorvariablen,  $g, h = 1, \dots, n$ , für dichotome Aufgaben, d.h. für Aufgaben mit nur zwei möglichen (richtig, falsch) Antworten. Irgendzwei Aufgaben  $I_g$  und  $I_h$  sind Rasch-homogen, wenn*

$$\theta - \kappa_g = \log \left[ \frac{P(X_g = 1|\theta)}{P(X_g = 0|\theta)} \right] = \log \left[ \frac{P(X_h = 1|\theta)}{P(X_h = 0|\theta)} \right] + \delta_{gh}. \quad (3.7)$$

**Anmerkung:** Formal bedeutet (3.7), dass sich die Logits für irgendzwei Aufgaben nur um eine additive Konstante  $-\delta_{gh}$  unterscheiden sollen; für fixen  $\theta$ -Wert bedeutet dies, dass sich die additive Konstante nur auf eine Differenz der Schwierigkeiten beziehen kann, was wiederum impliziert, dass  $\theta$  und die Schwierigkeiten sich auf das gleiche Merkmal beziehen müssen.

**Nachweis der Äquivalenz von (3.7) und (3.5):** Gilt nun (3.5), so folgt leicht, dass auch (3.7) gilt, mit  $\delta_{gh} = \kappa_h - \kappa_g$ .

Um zu sehen, dass die Beziehung (3.7) auch (3.5) impliziert, betrachte man zunächst

$$\theta - \kappa_g = \log \left[ \frac{P(X_g = 1|\theta)}{P(X_g = 0|\theta)} \right].$$

Dies ist gleichbedeutend mit

$$e^{\theta - \kappa_g} = \frac{P(X_g = 1|\theta)}{P(X_g = 0|\theta)} = \frac{P(X_g = 1|\theta)}{1 - P(X_g = 1|\theta)},$$

woraus sofort

$$P(X_g = 1|\theta) = \frac{e^{\theta - \kappa_g}}{1 + e^{\theta - \kappa_g}}$$

folgt. Weiter folgt in Bezug auf die rechte Seite von (3.7)

$$e^{\theta - \kappa_g} = \frac{P(X_h = 1|\theta)e^{\delta_{gh}}}{P(X_h = 0|\theta)} = \frac{P(X_h = 1|\theta)e^{\delta_{gh}}}{1 - P(X_h = 1|\theta)},$$

woraus sich

$$P(X_h = 1|\theta) = \frac{e^{\theta - \kappa_g - \delta_{gh}}}{1 + e^{\theta - \kappa_g - \delta_{gh}}} = \frac{e^{\theta - \kappa_h}}{1 + e^{\theta - \kappa_h}}$$

ergibt. □

**Parametrisierungen:** Die Form (3.5) des Rasch-Modells heißt *subtraktive Parametrisierung*, da die Wahrscheinlichkeit einer Lösung oder Beantwortung ( $X_g = 1$ ) von der Differenz  $\theta - \kappa_g$  abhängt. Es sei noch einmal darauf hingewiesen, dass sowohl die Person (mit dem Parameter  $\theta$ ) wie auch das Item (die Aufgabe)  $I_g$  (mit dem Parameter  $\kappa_g$ ) auf der gleichen Skala positioniert werden. Nun gilt aber auch

$$\frac{e^{\theta - \kappa_g}}{1 + e^{\theta - \kappa_g}} = \frac{e^\theta e^{-\kappa_g}}{1 + e^\theta e^{-\kappa_g}},$$

d.h. das Modell kann in der Form

$$F_g(\theta) \rightarrow G_g(\vartheta) = \frac{\vartheta \sigma_g}{1 + \vartheta \sigma_g}, \quad \vartheta = e^\theta, \quad \sigma_g = e^{-\kappa_g} \quad (3.8)$$

geschrieben werden; diese Form des Modells heißt *multiplikative Parametrisierung*. Für manche Betrachtungen ist diese Art der Parametrisierung vorteilhaft.

**Zulässige Transformationen:** Die  $\theta$ - und der  $\kappa$ -Skalen sind offenbar nur eindeutig bis auf Translationen. Dies folgt eben aus

$$\theta' = \theta + \alpha, \quad \kappa'_g = \kappa_g + \alpha \Rightarrow \theta - \kappa_g = \theta' - \kappa'_g, \quad (3.9)$$

d.h. die Skala, auf der  $\theta$  und  $\kappa_g$  Werte annehmen, ist eindeutig bis auf Addition einer beliebigen, endlichen Konstanten  $\alpha$  (die positiv oder negativ sein kann).

Hat man also Skalenwerte  $\theta$  und  $\kappa_g$  gefunden, die dem Modell genügen, so kann man

$$\alpha = - \sum_{g=1}^n \kappa_g \quad (3.10)$$

setzen und zu Werten  $\theta' = \theta + \alpha$ ,  $\kappa'_g = \kappa_g + \alpha$  übergehen. Anschließend kann man  $\theta'$  und  $\kappa'_g$  wieder in  $\theta$  und  $\kappa_g$  umbenennen, und für die  $\kappa_g$ -Werte gilt dann

$$\sum_{g=1}^n \kappa_g = 0. \quad (3.11)$$

**Wettquotienten und Logits:** Aus (3.5) folgt sofort

$$P(X_g = 0|\theta) = 1 - P(X_g = 1|\theta) = \frac{1}{1 - e^{\theta - \kappa_g}}.$$

Daraus ergibt sich die als *Wettquotient* bekannte Größe

$$\frac{P(X_g = 1)}{P(X_g = 0)} = e^{\theta - \kappa_g}, \quad (3.12)$$

und durch Logarithmieren<sup>11</sup> erhält man das *Logit*

$$\text{Logit}P(X_g = 1|\theta) = \log\left(\frac{P(X_g = 1)}{P(X_g = 0)}\right) = \theta - \kappa_g. \quad (3.13)$$

Der Wettquotient gibt an, wie die Chancen stehen, dass z.B. die Person  $a$  das Item  $I_g$  löst. Es sei etwa  $P(X_{ag} = 1) = .75$ . Dann ist  $P(X_{ag} = 0) = .25$  der Wettquotient beträgt  $.75/.25 = 3/1$ , d.h. die Chancen stehen 3 zu 1, dass  $a$  die Aufgabe löst. Beträgt die Wahrscheinlichkeit, das Item zu beantworten oder zu lösen, nur  $.65$ , so ist  $.65/.35 \approx 1.8 \approx 9/5$ ; - Wettquotienten werden üblicherweise als Quotienten ganzer Zahlen ausgedrückt und sind deswegen im Allgemeinen Näherungswerte.

Die folgende Eigenschaft des Modells macht seine wirkliche Bedeutung für die Evaluation von Leistungen aus: die Personenparameter  $\theta$  und die Schwierigkeitsparameter  $\kappa_g$  können *unabhängig voneinander geschätzt werden*:

### 3.2 Spezifische Objektivität

Gegeben seien zwei Probanden  $a, a' \in \mathcal{P}$  mit den Parameterwerten  $\theta_a$  und  $\theta_{a'}$ . Für ein gegebenes Item  $I_g$  werde nun die Differenz der Logits für die beiden Probanden betrachtet:

$$\log\left(\frac{P(X_{ag} = 1)}{P(X_{ag} = 0)}\right) - \log\left(\frac{P(X_{g a'} = 1)}{P(X_{g a'} = 0)}\right) = \theta_a - \theta_{a'} - \kappa_g + \kappa_g = \theta_a - \theta_{a'}. \quad (3.14)$$

<sup>11</sup>Es ist hier stets der *natürliche Logarithmus*, also der Logarithmus zur Basis  $e$ , gemeint.

Die Differenz der Logits ist also nur durch die Differenz der Probandenparameter definiert und unabhängig vom Itemparameter  $\kappa_g$ . Dieses Ergebnis gilt natürlich für ein beliebiges Item  $I_g$ , also für alle  $I_g$ . Die Gleichung (3.14) erlaubt den Vergleich zweier Personen unabhängig von den Items, mit denen sie getestet werden. Diese Eigenschaft eines Testmodells, also der Vergleich von Personen unabhängig von den im Test verwendeten Items, wird die *spezifische Objektivität* eines Tests genannt.

Auf analoge Weise sieht man, dass die Differenz der Logits für zwei Items unabhängig von den Personen ist:

$$\log\left(\frac{P(X_{ag} = 1)}{P(X_{ag} = 0)}\right) - \log\left(\frac{P(X_{g'a} = 1)}{P(X_{g'a} = 0)}\right) = \theta_a - \theta_a - \kappa_g + \kappa_{g'} = \kappa_{g'} - \kappa_g. \quad (3.15)$$

Diese Gleichung bedeutet, dass man die Differenz der Schwierigkeit zweier Items unabhängig von irgendwelchen Personenparametern bestimmen kann, d.h. die Items können unabhängig von der Stichprobe oder gar Population von Probanden hinsichtlich ihrer Schwierigkeit miteinander verglichen werden. Die spezifische Objektivität und die Populationsunabhängigkeit der Itemparameter sind charakteristisch für das Rasch-Modell und machen einen großen Teil seiner Attraktivität aus.

### 3.3 Die Schätzung der Modellparameter

Werden Personen befragt oder getestet, so möchte man i. A. etwas über die Ausprägung der untersuchten Merkmals bei den Personen erfahren. Darüber hinaus kann es von Interesse sein, zu erfahren, ob die Items das Merkmal in unterschiedlichem Maße erfassen oder nicht, ob sie also gleich schwierig sind oder nicht. Dazu müssen im Falle des Rasch-Modells die Parameter  $\theta_a$  und  $\kappa_g$  geschätzt werden.

Das Problem bei Modellen wie denen von Rasch ist, dass die Parameter nicht in direkter Beziehung zu beobachtbaren Größen stehen. Denn es werden nur Wahrscheinlichkeiten "vorausgesagt", um im Jargon der Regressionsanalyse zu bleiben: Die Gleichung

$$P(X_{ag} = 1 | \theta_a, \kappa_g) = \frac{e^{\theta_a - \kappa_g}}{1 + e^{\theta_a - \kappa_g}}$$

besagt ja, dass die rechte Seite eben nur die Wahrscheinlichkeit  $P(X_{ag} = 1 | \theta_a, \kappa_g)$  bestimmt. Die wird aber nicht beobachtet, sondern nur die manifeste Variable  $X_{ag}$ , die bei der Person  $a$  und der Aufgabe  $I_g$  bei einer Testvorgabe entweder den Wert 1 oder 0 annimmt. Man könnte die Wahrscheinlichkeit durch wiederholte Vorlage abschätzen, nur wird das nicht funktionieren: wenn die Person etwa die Lösung erinnert, wird die manifeste Variable stets den Wert 1 annehmen. Abgesehen davon wird die Anzahl der Vorlagen des Tests kaum jemals hinreichend groß sein können, um eine vernünftige Schätzung von  $P(X_{ag} = 1)$  anhand relativer Häufigkeiten zu erlauben.

Die Methode der Wahl ist demnach die *Maximum-Likelihood-Methode*. *Likelihood* heißt zunächst einfach nur – wie 'probability' – 'Wahrscheinlichkeit', meint hier aber speziell die Wahrscheinlichkeit der beobachteten Daten oder Messungen, unter der Bedingung, dass die Parameter der entsprechenden Wahrscheinlichkeitsverteilung bestimmte Werte haben. Es seien die Daten einer Stichprobe von  $m$  Personen gegeben, die auf  $n$  dichotome Items antworten. Die Daten können in einer Matrix zusammengefasst werden; man spricht auch von einer *Datenstruktur*:

$$D = \begin{array}{c|ccc|c|c} \text{Person} & \text{Aufgaben} & & & \Sigma_a \\ \hline 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 1 \\ 3 & 0 & 1 & 0 & 0 & 0 & 1 \\ 4 & 1 & 1 & 0 & 1 & 0 & 3 \\ 5 & 0 & 0 & 1 & 1 & 1 & 3 \\ 6 & 1 & 1 & 0 & 0 & 0 & 2 \\ 7 & 1 & 0 & 1 & 1 & 1 & 4 \\ 8 & 0 & 0 & 0 & 0 & 1 & 1 \\ 9 & 1 & 1 & 0 & 1 & 1 & 4 \\ 10 & 0 & 0 & 1 & 1 & 0 & 2 \\ \hline \Sigma_g & 4 & 4 & 3 & 5 & 5 & \end{array} \quad (3.16)$$

Hier gibt es  $m = 10$  befragte oder getestete Personen  $a = 1, \dots, m$ , und fünf dichotome Aufgaben,  $g = 1, \dots, n$ . Eine 1 oder eine 0 repräsentiert die Realisierung der zufälligen Veränderlichen  $X_{ag} = \{0, 1\}$ .  $n_g = \sum_a X_{ag}$  ist die Häufigkeit, mit der die  $g$ -te Aufgabe gelöst bzw. positiv beantwortet wurde,  $m_a = \sum_g X_{ag}$  ist die Anzahl der Aufgaben, die von der Person  $a$  gelöst wurden.

Man kann nun die Likelihood dieser Daten definieren, vorausgesetzt, die  $X_{ag}$  sind alle stochastisch unabhängig. Die Likelihood der Daten ist durch

$$L(X_{11}, \dots, X_{mn}) = \prod_{a=1}^m \prod_{g=1}^n \frac{e^{X_{ag}(\theta_a - \kappa_g)}}{1 + e^{\theta_a - \kappa_g}}, \quad X_{ag} = \{0, 1\} \quad (3.17)$$

gegeben.

Bei der Likelihood-Funktion (3.17) sind die Parameter  $\theta_a$  und  $\kappa_g$  die freien Parameter, deren Werte bestimmt werden sollen. Der Bestimmung dieser Werte liegt die folgende 'Philosophie' zugrunde: Es folgt ja aus dem Begriff der Wahrscheinlichkeit, dass ein zufälliges Ereignis  $E$ , dessen Wahrscheinlichkeit höher ist als die des Ereignisses  $E' \neq E$ , eben eher und damit auch häufiger eintritt als  $E'$ . Hat man umgekehrt ein zufälliges Ereignis  $E$  beobachtet, so kann man davon ausgehen, dass  $E$  eine höhere Wahrscheinlichkeit hat als die Ereignisse  $E'$ , deren Wahrscheinlichkeit geringer ist. Das kann im Einzelfall durchaus falsch sein, aber *im Allgemeinen* liegt man mit dieser Vermutung richtig. Da nach (3.17) die Wahrscheinlichkeit  $L$  der Daten von den Werten der Parameter  $\theta_a$  und  $\kappa_g$  abhängt, ist es dementsprechend eine plausible Annahme, dass sie  $L$  maximieren.



Man kann  $L$  als Funktion von  $\theta_a$  und  $\kappa_g$  auffassen. Man kann nun  $L$  als Funktion von  $\hat{\theta}_a$  und  $\hat{\kappa}_g$  auffassen. Die Aufgabe ist nun, diejenigen Werte  $\hat{\theta}_a$  und  $\hat{\kappa}_g$  zu bestimmen, für die  $L$  maximal wird. Diese Werte bestimmt man, indem man die *Likelihood-Funktion* einmal nach  $\theta_a$  differenziert, und einmal nach  $\kappa_g$ ; man spricht von *partiellen Ableitungen*. Für diejenigen Werte  $\hat{\theta}_a$  und  $\hat{\kappa}_g$ , für die die beiden partiellen Ableitungen gleich Null werden, nimmt  $L$  einen maximalen Wert an. Man fasst dementsprechend die beiden partiellen Ableitungen als Gleichungen in den Unbekannten  $\hat{\theta}_a$  und  $\kappa_g$  auf, mit  $a = 1, \dots, m$  und  $g = 1, \dots, n$ , d.h. es müssen  $n + m$  Gleichungen gelöst werden.

Die Herleitung dieser Gleichungen soll hier nicht durchgeführt werden; das partielle Differenzieren von  $L$  nach den  $\theta_a$  und  $\kappa_g$  etc. ist ein wenig länglich und trägt zum Verständnis des Folgenden nichts weiter bei. Es genügt, das Resultat dieser Rechnungen anzugeben:

$$m_a = \sum_{g=1}^n \frac{\exp(\hat{\theta}_a - \hat{\kappa}_g)}{1 + \exp(\hat{\theta}_a - \hat{\kappa}_g)}, \quad a = 1, \dots, m \quad (3.18)$$

$$n_g = \sum_{a=1}^m \frac{\exp(\hat{\theta}_a - \hat{\kappa}_g)}{1 + \exp(\hat{\theta}_a - \hat{\kappa}_g)}, \quad g = 1, \dots, n. \quad (3.19)$$

Es ist nicht möglich, diese Gleichungen explizit nach den  $\theta_a$  und  $\kappa_g$  aufzulösen, so dass eine numerische Lösung gefunden werden muß. Programme für die Lösung der Gleichung sind verfügbar. Die Schätzungen haben eine Reihe von wünschenswerten Schätzungen, insbesondere die der *Suffizienz*; dies bedeutet, dass sie alle Informationen über die Parameter, die in den Daten vorhanden sind, enthalten.

Die Lösung dieser Gleichungen ist numerisch aufwändig, allerdings liegen mittlerweile frei verfügbare Programme vor, die jedem Anwender zugänglich sind, s. <https://cran.r-project.org/web/packages/eRm/eRm.pdf>.

### 3.4 Die Bewertung von Schulleistungen

Es ist bereits argumentiert worden, dass eine Benotung auf der Basis einer an die Punkteverteilung angepasste Normalverteilung nicht zu rechtfertigen ist, weil Noten erzeugt werden, die nicht dem objektiven Leistungsstand der Schülerinnen und Schüler entsprechen. Das Rasch-Modell hat aber die Eigenschaft der *spezifischen Objektivität*, d.h. die Personenparameter  $\theta_a$  und die Schwierigkeitsparameter  $\kappa_g$  können unabhängig voneinander geschätzt werden.

Das Rasch-Modell beruht auf der Annahme der logistische Verteilung der Kompetenzen. Wie in der Einleitung bereits angedeutet wurde, ähnelt diese Verteilung der Gauß-Verteilung, hat aber wegen ihrer anderen mathematischen Definition andere Eigenschaften als die Gauß-Verteilung, was sich auf die statistischen Eigenschaften der Parameterschätzungen auswirkt; die Eigenschaft der spezifischen Objektivität kann bei der Gauß-Verteilung nicht erreicht werden. Dies hat

dazu geführt, dass viele Theoretiker das Rasch-Modell gerne als eine Art kanonisches Modell sehen, dass alle anderen Modelle ersetzen sollte.

Es ist aber ebenfalls argumentiert worden, dass die Fähigkeiten z.B. der Gymnasialschüler insgesamt eher einer log-Normalverteilung folgen; die empirisch gefundene Asymmetrie Leistungen kann weder vom Gauß- noch vom logistischen (Rasch-) Modell abgebildet werden. Die Frage, welche der Modelle denn nun wirklich das richtige ist, ist möglicherweise falsch gestellt; denkbar ist ja, dass die Art der Aufgabenstellung implizit die stochastischen Mechanismen und damit die Verteilung der Scores bestimmt, die interpretiert werden muß. Das Rasch-Modell macht es jedenfalls möglich, anhand großer Stichproben (wie sie in PISA-Untersuchungen anfallen) die Schwierigkeitsparameter von Aufgaben unabhängig von den verschiedenen Stichproben zu bestimmen. Da psychologische Betrachtungen das log-Normal-Modell favorisieren, bedeutet die Wahl des Rasch-Modells, dass die Tests so konstruiert werden müssen, dass das Rasch-Modell auch zutrifft, – was keineswegs immer gewährleistet ist: Vorberg und Schwarz (1990) haben gezeigt, dass z.B. Aufgaben, bei denen die (Reaktions-)Zeit bis zum Finden der Lösung der eigentlich interessierende Aspekt der Aufgabenbeantwortung ist, grundsätzlich nicht im Rahmen des Rasch-Modells interpretiert werden können (solche Aufgaben sind u.a. Teil von Intelligenztests). Auch die Gauß- oder die log-Normalverteilung sind dann keine geeigneten Verteilungen mehr. Dieser Befund weist auf den Sachverhalt, dass man das Antwortverhalten nicht durch *ein* Modell beschreiben kann, das für *alle* Aufgabentypen gilt.

Jedenfalls sind die Schwierigkeitsparameter, wenn sie denn im Rahmen des Rasch-Modells geschätzt werden können, Werte auf derselben Skala, auf der auch die Personenparameter  $\theta_a$  liegen. Ist das Rasch-Modell überhaupt anwendbar, so lassen sich über die  $\kappa_g$  nun die Noten definieren, nach Maßgabe einer allgemeinen Diskussion über das, was die Noten 'sehr gut', 'gut', etc bedeuten sollen. Dies bedeutet, dass für jede Altersgruppe eine Schwierigkeitsskala mit  $\kappa$ -Werten aufgestellt wird derart, dass ein geschätzter  $\theta_a$ -Wert zwischen  $\kappa_2$  und  $\kappa_1$  ein 'sehr gut' bedeutet, ein Wert  $\theta_a$  zwischen  $\kappa_2$  und  $\kappa_3$  ein 'gut', etc. Die  $\theta_a$ -Werte werden für alle Schüler  $a = 1, \dots, m$  gemäß (3.18) und (3.19) geschätzt. Da das oben genannte Programm eRm (extended Rasch-model) auf jedem Rechner läuft, ist die Schätzung der  $\theta_a$  kein größeres Problem als das Anpassen einer Normalverteilung.

Die Abbildung der  $\theta_a$ -Werte auf die Notenskala ist natürlich stetig und hat die Form  $y_a = \alpha\theta_a + \beta$ ,  $1 \leq y_a \leq 5$ ; auf diese Weise sind alle Zwischennoten möglich.

Man sieht, dass die Konstruktion einer Notenskala, die vernünftigen und stets auch intuitiv angenommenen Kriterien genügt, keine triviale Aufgabe ist, die an ein Reihe einschränkender Bedingungen geknüpft ist. Das beliebte Modell der Gauß-Verteilung der Leistungen und damit der Noten genügt diesen Bedingungen keineswegs. Da auch das Rasch-Modell mit seinen angenehmen Eigenschaften nicht notwendig auch psychologisch sinnvoll ist, d.h. nicht notwendig der Struktur

der kognitiven Prozesse abbildet, die der Hintergrund der schulischen Leistungen sind, die benotet werden sollen, stellt sich die Frage nach der Sinnhaftigkeit einer Notengebung, von der mehr als nur eine ungefähre Standortbestimmung von SchülerInnen erwartet wird. Auch wenn Schulnoten eine relativ gute regressionsstatistische Vorhersage des Studienerfolgs ermöglichen, so muß doch bedacht werden, dass etwa die Vergabe von Studienplätzen nach Maßgabe der Abiturnoten ihre Eigendynamik erzeugt: die Notenvergabe und die Präsentation der Lehrinhalte muß so eingeregelt werden, dass die "sehr guten" Abiture auch möglich werden, und diese Dynamik wird auf das Studium und die akademischen Prüfungen vererbt. Die Problematik der Benotung wird dabei vergessen.

## 4 Anhang

### 4.1 Zur Gauß-Verteilung

Die Normal- oder Gauß-Verteilung ist durch die Funktion<sup>12</sup>

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty \quad (4.1)$$

definiert.  $\mu = \mathbb{E}(X)$  ist der *Erwartungswert*, d.h. der Mittelwert über *alle möglichen Werte*  $x$ ; normalerweise hat man nur eine Stichprobe von Messwerten, und der Mittelwert

$$\bar{x} = \frac{1}{n}(x_1 + \dots + x_n) = \frac{1}{n} \sum_{i=1}^n x_j \quad (4.2)$$

dient als Schätzung für  $\mu$ .  $\sigma^2$  ist die *Varianz* der  $x$ -Werte, d.h. der Mittelwert der *aller möglichen* quadrierten Abweichungen der  $x_j$  von  $\mu$ . Eine Schätzung für  $\sigma^2$  ist

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_j - \bar{x})^2. \quad (4.3)$$

$\sigma$  bzw.  $s$  ist die *Standardabweichung* der  $x$ -Werte.  $\sigma^2$  ist, wie  $\sigma$ , ein Maß für die Unterschiedlichkeit der  $x$ -Werte.

**Skalentransformation:** Es sei  $Y = aX + b$  eine Skalentransformation einer Gauß-verteilten zufälligen Veränderlichen  $X$ . Die Frage ist, wie  $Y$  verteilt ist. Dazu betrachtet man die *Verteilungsfunktion* von  $X$  bzw.  $Y$ . Diese ist durch  $P(X \leq x)$  als Funktion von  $x$  definiert:

$$P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \quad (4.4)$$

<sup>12</sup>  $\exp$  steht für die  $e$ -Funktion:  $\exp(x) = e^x$ .  $e$  ist die Eulersche Zahl  $e = 2.718281828459045\dots$ , Basis für den natürlichen Logarithmus.

Um die Verteilung von  $Y$  zu erhalten, betrachtet man  $P(Y \leq y)$ :

$$P(Y \leq y) = P(aX + b \leq y) = P\left(X \leq \frac{y-b}{a}\right).$$

Damit ergibt sich die Verteilung von  $Y$  über die Verteilung von  $X$ . Differentiation (Kettenregel) bezüglich  $y$  liefert dann mit  $h = (y-b)/a$

$$g(y) = \frac{dP(Y \leq y)}{dy} = \frac{dP(X \leq h)}{dh} \frac{dh}{dy}$$

$g$  die Dichtefunktion von  $y$ , und

$$\begin{aligned} g(y) &= f_X\left(\frac{y-b}{a}\right) \frac{1}{a} = \frac{1}{a\sigma\sqrt{2\pi}} \exp\left(-\frac{\frac{y-b}{a} - \mu)^2}{2\sigma_x^2}\right) \\ &= \frac{1}{a\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-b-\mu)^2}{2a^2\sigma_x^2}\right) \end{aligned} \quad (4.5)$$

Für den Erwartungswert und die Varianz von  $Y$  gelten die Beziehungen  $\mathbb{E}(Y) = a\mathbb{E}(X) + b = a\mu + b$ ,  $Var(Y) = a^2Var(X) = a^2\sigma_x^2$ . Dementsprechend kann man (4.5) in der Form

$$g(y) = \frac{1}{\sigma_y\sqrt{2\pi}} \exp\left(-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right) \quad (4.6)$$

schreiben.  $Y$  ist wieder Gauss-verteilt mit den Parametern  $\mu_y = a\mu_x + b$  und der Varianz  $\sigma_y^2 = a^2\sigma_x^2$ .

**Geschichte:** Bereits 1733 zeigte Abraham de Moivre<sup>13</sup> dass die Wahrscheinlichkeitsverteilung einer Summe

$$x = x_1 + x_2 + \dots + x_n$$

von zufälligen Veränderlichen  $x_j$  gegen die Verteilung (4.1) strebt, wenn  $n \rightarrow \infty$ , wobei er aber nur den Spezialfall  $x_j = \{0, 1\}$  für alle  $i$  betrachtete, d.h. die  $x_j$  können nur die Werte 0 oder 1 annehmen. Pierre-Simon Laplace (1749 – 1827) zeigte 1782, dass die Aussage unter sehr allgemeinen Randbedingungen gilt, wenn die  $x_j$  nicht auf die Werte 0 oder 1 beschränkt sind (daher die Rede vom *de Moivre-Laplaceschen Grenzwertsatz*. 1809 publizierte Carl Friedrich Gauß (1777 - 1855) sein Werk *Theorie der Bewegung der in Kegelschnitten sich um die Sonne bewegendenden Himmelskörper*, in dem er die Normalverteilung ebenfalls definiert und sie im Zusammenhang mit der Methode der Kleinsten Quadrate diskutiert, die es gestattet, unbekannte Parameter (Konstante, die in die Definition von Funktionen eingehen) aus empirischen Daten zu schätzen. Der belgische Mathematiker Lambert Adolphe Jacques Quetelet (1796 – 1874) fand 1844, dass seine Messungen des Brustumfangs von Tausenden von Soldaten sehr gut mit der Funktion (4.1)

<sup>13</sup>Abraham De Moivre (1667 – 1754), französischer Mathematiker

übereinstimmten (d.h. die Häufigkeiten der verschiedenen Brustumfänge ließen sich durch diese Funktion gut beschreiben), und machte analoge Beobachtungen in anderen biologischen Bereichen; man vermutet, das der Ausdruck *Normalverteilung* auf Quetelet zurückgeht; (4.1) beschreibt demnach "normalerweise" die Häufigkeiten verschiedener Ausprägungen von Merkmalen. Der Ausdruck  $n \rightarrow \infty$  bedeutet für die praktische Anwendung, dass  $n$  zwar endlich, aber "groß" ist, und ein Wert von  $n > 30$  liefert oft schon eine gute Approximation der unbekanntenen Verteilung durch Normalverteilung. Dieser Sachverhalt erklärt die Bedeutung der Normal- oder Gauß-Verteilung.

## 4.2 Simulation log-normalverteilter zufälliger Variablen

Zur Simulation log-normalverteilter zufälliger Veränderlicher (rlnorm in R): Es sei  $X$  log-normalverteilt; dann hat  $X$  den Erwartungswert und die Varianz

$$\mathbb{E}(X) = e^{\mu + \sigma^2/2}, \quad \text{Var}(X) = e^{2\mu + \sigma^2}(e^{\sigma^2} - 1). \quad (4.7)$$

Zur Abkürzung werde  $m = \mathbb{E}(X)$  und  $s = \sqrt{\text{Var}(X)}$  gesetzt. Offenbar gilt  $s^2 = m^2(e^{\sigma^2} - 1)$ .

Der Aufruf von rlnorm erfordert die Eingabe von  $n$ , der Anzahl der Fälle (hier:  $n = 30 =$  Klassengröße, die Übereinstimmung mit dem maximal erreichbaren Score 30 ist Koinzidenz, so wie die Werte meanlog und sdlog. meanlog ist ein Lokations- und sdlog ist ein Formparameter (location, shape). Die Bemerkungen im Text zu rlnorm sind etwas irreführend, deshalb hier eine Erläuterung. Erwünscht ist die Eingabe von  $\mu$  und  $\sigma$ . Bei einer Simulation gibt man aber  $\mathbb{E}(X)$  und  $s$  vor; man muß  $\mu$  und  $\sigma$  also ausrechnen. Aus (4.7) folgt

$$\frac{s^2}{m^2} = e^{\sigma^2} - 1 \Rightarrow e^{\sigma^2} = 1 + \frac{s^2}{m^2},$$

woraus

$$\sigma = \sqrt{\log(1 + s^2/m^2)} \quad (\text{shape}) \quad (4.8)$$

folgt.

Weiter gilt

$$m^2 = e^{2\mu} e^{\sigma^2} = e^{2\mu} (1 + s^2/m^2),$$

woraus sich

$$\frac{m^2}{1 + s^2/m^2} = e^{2\mu} \Rightarrow e^\mu = \frac{m}{\sqrt{1 + s^2/m^2}}$$

so dass

$$\mu = \log \left[ \frac{m}{1 + \frac{s^2}{m^2}} \right] \quad (\text{location}) \quad (4.9)$$

folgt.

## Literatur

- [1] Aitchinson, J. Brown, J.A.C.: The Lognormal distribution. Cambridge 1963
- [2] Billingsley, P.: Probability and Measure. New York 1979
- [3] Bleher, C.: Nicht so viele Einsen, bitte!  
<http://www.sueddeutsche.de/karriere/kritik-an-guten-noten-nicht-zu-viele-einser-bitte-1.592366>
- [4] Dorfman, D.D. (1978) The Cyril Burt Question: New Findings. *Science*, 201 (4362), 1177 – 1186
- [5] Koch, A.L. (1966) The Logarithm in Biology. 1. Mechanisms Generating the Log-Normal Distribution Exactly. *Journal of Theoretical Biology*, 12, 276–219
- [6] Limpert, E., Stahel, E.A., Appt, M.: Log-normal distributions across the sciences. *Biosciences*, 2008, 51(5), 341–352
- [7] Limpert, E., Stahel, W.A. (2011) Problems with using the normal distribution – and ways to improve quality and efficiency of data analysis. *PLoS One*, 6(7), 1 – 8
- [8] McNemar, Q.: The Revision of the Stanford-Binet Scale. An Analysis of the Standardization Data. . Houghton Mifflin, Boston 1942
- [9] Quine, W. v. O. (1951) The two dogmas of empiricism. *Philosophical Review* 60 (1), 20 – 43
- [10] Rasch, G.: Probabilistic Models for some intelligence and attainment tests. The Danish Institute of Educational Research, Copenhagen, 1960
- [11] Rasch, G. (1961) On general laws and the meaning of measurement in psychology. The Danish Institute of Educational Research, Copenhagen.
- [12] Steyr, R. Eid, M.: Messen und Testen. Springer-Verlag Berlin Heidelberg New York 1993
- [13] Sun, K. (2004) Explanation of Log-Normal Distributions and Power-Law Distributions in Biology and Social Science.  
<http://guava.physics.uiuc.edu/nigel/courses/569/Essays2004/files/sun.pdf>
- [14] Thomas, H. (1982) IQ, Interval Scales, and Normal Distributions. *Psychological Bulletin*, 91(1), 198–202
- [15] Tversky, A., Kahneman, D. (1974) Judgment under Uncertainty: Heuristics and Biases. *Science*, 185, 1124–1131

- [16] Vorberg, D., Schwarz, W. (1990) Rasch-representable reaction time distributions. *Psychometrika*, 55, 617 – 632
- [17] Yule, G.U., Kendall, M.G.: An Introduction to the Theory of Statistics. Griffin, London 1937

## Index

Datenstruktur, 32

Itemfunktion, 27

Likelihood-Funktion, 33

log-Normalverteilung, 3

Logit, 30

Maximum-Likelihood, 32

Parametrisierung

    multiplikative, 29

    subtraktive, 29

Rasch-Homogenität, 28

spezifische Objektivität, 30

Verteilungsfunktion, 35

Wettquotient, 30