Verteilte Schulnoten sind normal – aber nicht notwendig normalverteilt

U. Mortensen,

Erste Fassung 01. 07. 2016
neue, korrigierte Fassung 31. 01. 2020
letzte Korrektur 15. 09.2021

Inhaltsverzeichnis

Zusammenfassung Weidemann (2008), Behlert (2010) und andere berichteten über die bayerische Grundschullehrerin Sabine Czerny, die ihre Schülerinnen und Schüler auch für normalerweise ungeliebte Fächer wie Mathematik ("Rechnen") zu begeistern wußte und 91% ihrer SchülerInnen so weit brachte, dass sie zu einem Wechsel in die Realschule oder ins Gymnasium befähigt waren. Sie wurde gebremst, strafversetzt und aufgefordert, "sich an das Niveau ihrer Parallelkollegen" anzupassen, wozu gehöre, dass sie die Schulnoten innerhalb einer Klasse im Durchschnitt durch eine "3" zu charakterisieren sei. Diese Praxis der Notenvergabe scheint trotz aller Proteste bis heute noch üblich zu sein. Im Folgenden wird die Gauß-Normierung von der statistischen Seite beleuchtet.

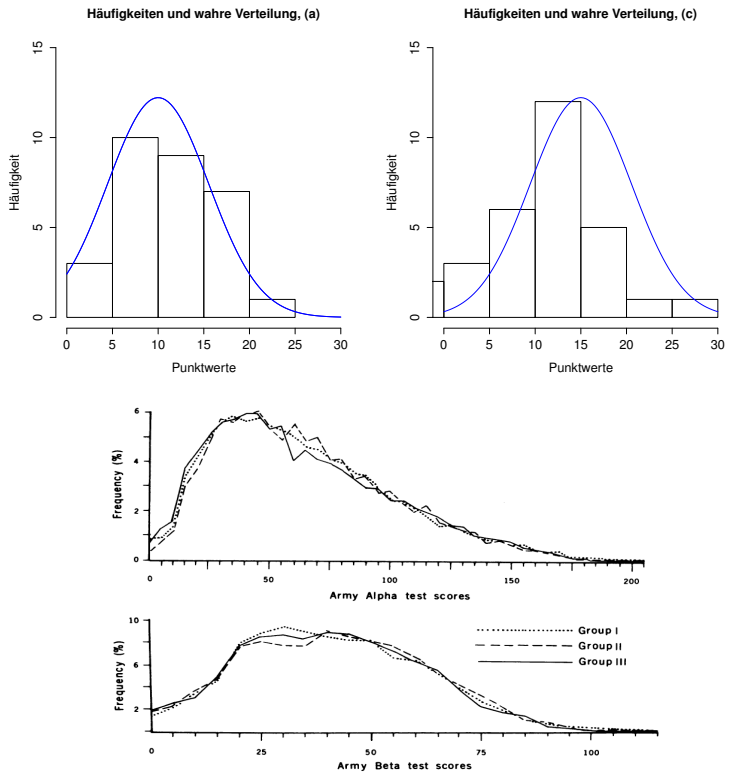
1 Schulnoten und ihre Verteilungen

Schulnoten sollen, für ein gegebenes Fach, die fachliche Kompetenz von SchülernInnen relativ zum Schuljahr abbilden. Die Kompetenz ist eine latente, also nicht

direkt beobachtete Variable, deren Ausprägung anhand von Indikatoren abgeschätzt wird: das sind mündliche oder schriftliche Leistungen. LehrerInnen gehen bei der Vergabe von Noten gelegentlich "ganzheitlich" vor, d.h sie schätzen die Kompetenz eher intuitiv auf der Basis von "Daten" wie mündliche Mitarbeit, Aufsätzen, Lösungen von Mathematik- oder Physikaufgaben, biologische Kenntnisse etc ab und bewerten sie nach Maßgabe ihrer (d.h. des Lehrenden) Erfahrung, wobei sie durchaus bestimmte mehr oder weniger explizit definierte Kriterien verwenden können. Behlert (2010) zitiert aus einer Expertise der Universität Siegen das Experiment des österreichischen Pädagogen Rudolf Weiss, der 153 Lehrer eine Mathematikaufgabe beurteilen ließ. 41 Prozent von ihnen gaben eine Zwei, 42 Prozent eine Drei, die Eins wurde von sieben Prozent vergeben, die Vier von neun Prozent und ein Prozent der Probanden sahen in der Arbeit sogar eine Fünf. Dass ein und derselbe Deutschaufsatz mit allen Noten von "1" bis "5" bewertet werden kann war schon länger bekannt, dass diese Möglichkeit auch für Mathematikaufgaben besteht ist ein eher neuer Befund. Kompetenz ist eine komplexe Eigenschaft, während eine Notenskala eindimensional ist, so dass für die Vergabe einer Note die verschiedenen Aspekte der Kompetenz in geeigneter Weise kombiniert werden müssen. Die Bedeutung des Ausdrucks "geeignete Weise" wird zum einen durch ministerielle Vorgaben und zum anderen durch die individuelle Interpretation eines Lehrers oder einer Lehrerin bestimmt; auf die Details der Spezifikation des Kompetenzbegriffs soll im Folgenden aber nicht weiter eingegangen werden, d.h. der Begriff wird in gewisser Weise naiv verwendet werden. Die intuitive Vergabe von Noten kann nachweislich durch systematische, dem Lehrenden im Allgemeinen nicht bewußte Urteilsfehler verzerrt werden (Kahneman & Tversky (1974), Dawes (1994)), so dass nach objektiveren Methoden gesucht wird. Die Idee ist, Leistungen nach möglichst explizit formulierten Kriterien durch Vergabe von Punkten zu bewerten und dann für jeden Schüler S_i, $i = 1, \ldots, m$, die Gesamtzahl n_i der erreichten Punkte auszuzählen. Die n_i werden dann zu einer *Punkteverteilung* zusammengefasst: dazu wird eine Folge von Intervallen gebildet, etwa $I_1 = [0, 5), I_2 = [5, 10), I_3 = [10, 15), \ldots$ etc[1]. Anschließend wird ausgezählt, wieviele der Punktwerte n_i im Intervall I_1 liegen: es seien h_1 (h von Häufigkeit), h_2 in I_2, etc. Über jedes Intervall I_k wird dann ein Rechteck mit der Höhe h_k bzw. der Höhe h_k/n gezeichnet, wobei n die Maximalzahl der erreichbaren Punkte ist (s. Abbildung 1, obere Reihe). Diese Rechtecke repräsentieren die Punkteverteilung. Die eingezeichneten glockenförmigen Funktionen sind die jeweils bestpassenden Gauß-Verteilungen (Dichtefunktionen, s. Anhang) $f(x)$. Die durch die Aneinanderreihung der Intervalle I_k definierte Skala wird dabei als ein Kontinuum von Werten x einer zufälligen Veränderlichen X mit der Dichtefunktion $f(x)$ interpretiert. Die implizite Annahme ist, dass die Kompetenz eines Schülers durch einen Punkt auf einem Kontinuum repräsentiert werden kann. Die Gesamtzahl (der *Score*) n_i von Punkten, die der Schüler S_i erarbeitet hat, ist eine

[1] Dies sind *halb offene* Intervalle: die Zahl rechts von der eckigen Klammer gehört zum Intervall, die Zahl links von der runden Klammer nicht.

Abbildung 1: Stichproben aus Gauß-verteilten Punkteverteilungen (obere Reihe) und empirische, nicht normalverteilte Army Alpha Scores (aus Dorfman 1978)



Schätzung dieses Punktes.

Im Zentrum der folgenden Betrachtungen steht das Postulat, dass Schulnoten in einer Klasse Gauß-verteilt (normalverteilt) sind. Den Häufigkeitsverteilungen in der oberen Reihe von Abbildung 1 überlagert sind die korrespondierenden Gauß-Dichten; die Noten sind noch nicht eingetragen worden. Darüber hinaus wird es gerne gesehen, wenn die Notenskala eine *Intervallskala* ist, d.h. dass die von den einzelnen Noten abgedeckten Kompetenzintervalle gleich groß sind.

Die erste Frage ist, warum die Häufigkeitsverteilung durch eine Gauß-Verteilung beschrieben bzw. approximiert werden soll. Die Gauß-Verteilung ist auf dem gesamten Intervall $-\infty < X < \infty$ definiert, aber der Begriff der negativen Kompetenz ist kaum jemals definiert worden, so dass die Gauß-Verteilung grundsätzlich nur als Approximation verstanden werden muß. Dabei kann es aber geschehen, dass der ins Negative reichende Teil der bestpassenden Gauß-Dichte abgeschnitten ("trunkiert") wird, so dass eben eine *trunkierte Verteilung* betrachtet werden muß. Die Rechtfertigung für die Wahl der Gauß-Verteilung ist jedenfalls (i), dass

der Wert einer zufälligen Veränderlichen X als eine Summe

$$X = \mu + \varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_n = \mu + \underbrace{\sum_{i=1}^{k} \varepsilon_i}_{\varepsilon} = \mu + \varepsilon, \quad -\infty < X < \infty \tag{1.1}$$

dargestellt werden kann, wobei (ii) die ε_j, $j = 1, \ldots, n$ unabhängige zufällige Veränderliche sind, die alle möglichen, nicht direkt beobachtbbaren Einflüsse auf die gemessene Variable repräsentieren. Dann greift, wenn $k \to \infty$, – das heißt, wenn die Anzahl k der ε_j hinreichend groß ist, der *Zentrale Grenzwertsatz*, demzufolge die Summe ε approximativ Gauß-verteilt mit einem wahren Mittelwert von Null ist. Dann ist X approximativ Gauß-verteilt mit dem "wahren" Mittelwert μ und einer "wahren" Varianz σ^2 (s. Anhang). Für die i-te Person (SchülerIn) nimmt X den Wert X_i an, wobei $X_i \approx n_i$, n_i ist der Punktwert der i-ten Person. X_i ist eine reelle Zahl, während n_i eine natürliche Zahl is, wenn nur ganzzahlige Punktewerte vergeben werden. Die Gleichung (1.1) kann als ein Modell für die Ausprägung der Kompetenz der SchülerInnen gesehen werden.

Man muß an dieser Stelle klären, was eigentlich mit einem X-Wert gemeint ist. Erklärt werden soll, warum die Noten zufällig verteilt sind, – wären sie es nicht, würde man keine Wahrscheinlichkeitsverteilung zu ihrer Beschreibung benötigen. Wir unterliegen täglich zahllosen Einflüssen, die wir nicht alle kontrollieren können und die u.a. auch bewirken können, dass wir bei einer Klausur einmal besser, bei einer anderen Klausur einmmal schlechter vorbereitet sein können, d.h. die individuelle Leistung in einer Klausur ist von zufälligen Elementen überlagert. Dies bedeutet, dass die Summen"scores" der Probanden selbst zufällige Veränderliche sind. Alle diese zufälligen Effekte sollen sich dann so zusammenfügen, dass sich für die Notenskala eine bestimmmte Verteilung ergibt, die der Bedingung (1.1) entspricht. Es wird angenommen, dass jeder Proband durch eine bestimmte Kompetenz μ_i in dem zu prüfenden Fach gekennzeichnet ist und die μ_i insofern zufällig auf dem latenten Kontinuum verteilt sind, als die Klasse im Prinzip eine zufällig zustande gekommene Stichprobe aus der Gesamtpopulation der Schüler aus der entsprechenden Altersgruppe darstellt. Die i-te Person hat zum Zeitpunkt des Tests, also etwa einer Klausur, eine Kompetenz X aus einem kleinen Intervall $[X_i, X_i + dx)$, $X_i = \mu_i + \varepsilon_i$, $\varepsilon_i = \varepsilon_{i1} + \cdots + \varepsilon_{im}$, ε_{ik}, $k = 1, \ldots, m$ sind die individuellen "Störungen" der Leistungsfähigkeit oder Kompetenz der i-ten Person zum Zeitpunkt der Prüfung.

Nun ist allerdings überhaupt nicht klar, warum unkontrollierte Effekte generell *additiv* auf eine Messgröße einwirken sollen, und so wurde schon früh Widerspruch gegen die angeblich allein selig machende Wirkung eines rein additiven Modells erhoben, zumal sich fragen läßt, ob eine zufällige Veränderliche, die nur auf den positiven reellen Zahlen definiert ist, die zufällige Variation von Kompetenzen nicht besser repräsentiert als die Gauß-Verteilung. Die übliche Bezeichnung 'Normalverteilung' mag zum nahezu kritiklosen Gebrauch

der Gauß-Verteilung beigetragen haben. Dieser Ausdruck geht auf den französischen Mathematiker und Astronomen Lambert Adolphe Jacques Quetelet (1796 – 1874) zurück. Quetelet behauptete auf der Basis u.a. von biologischen Messungen (z.B. der Brustumfänge französischer Rekruten), die Gauß-Verteilung sei der Normalfall bei biologischen und sozialen Variablen, weshalb er eben den Ausdruck 'Normalverteilung' prägte. So wurde suggeriert, dass diese Verteilung Ausdruck grundlegender Prozesse in der Natur sei. Dorfman (1978), p. 1180, zitiert die renommierten Statistiker Yule und Kendall (1937) mit der Bemerkung: "The normal curve was, in fact, to the early statisticians what the circle was to the Ptolemaic astronomers"[2], und McNemar (1942) merkte 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[3] for psychological traits" (zitiert nach Dorfman (1978), p. 1180).

So ergibt sich die Frage, welch andere Verteilungen in Frage kommen. Spätere genauere Messungen biologischer Merkmale zeigten, dass diese Merkmale keineswegs normalerweise Gauß-verteilt sind; die Daten legen nahe, dass die *log-Normalverteilung* die Daten besser als die Gauß-Verteilung beschreibt (Aitchinson & Brown (1963), Koch (1966), Limpert, Stahel & Apt (2008), Limpert & Stahel (2011)). Diese ergibt sich, wenn nicht die gemessene Variable, sondern der Logarithmus der Variable Gauß-verteilt ist. So gelangt man zu einer log-Normalverteilung, wenn man statt des additiven Ansatzes (1.1) den multiplikativen Ansatz

$$X = \varepsilon_1 \cdot \varepsilon_2 \cdots \varepsilon_n = \prod_{i=1}^{n} \varepsilon_i, \quad \varepsilon_i > 0,\ i = 1, \ldots, n. \tag{1.2}$$

wählt. Dieser Ansatz ergibt sich, wenn die unbekannten Effekte proportional zueinander wirken. Dann kann man $X = e^{\log X}$ schreiben (es ist stets der natürliche Logarithmus gemeint), und wegen

$$\log X = \log \varepsilon_1 + \log \varepsilon_2 + \cdots + \log \varepsilon_n,$$

gilt wieder der Zentrale Grenzwertsatz, so dass die Verteilung von $\log X$ zumindest approximativ durch eine Gauß-Verteilung gegeben ist. (1.2) impliziert, dass X nicht negativ werden kann, d.h. $0 < X < \infty$, im Gegensatz zu $-\infty < X < \infty$ für X Gauß-verteilte Messwerte. Schulische Leistungen können mangelhaft oder gar nicht vorhanden sein, aber der Begriff einer negativen Leistung ist nicht definiert. Andererseits kann man kaum eine definite obere Grenze für Leistungen angeben, – das Beste ist, man läßt die Skala der möglichen Leistungen nach oben offen. In Abbildung 1 sind die unteren beiden Verteilungen empirische Häufigkeitsverteilungen von Testwerten im *Army-Alpha-Test*, einem vom US-Militär

[2]Der Kreis galt als "perfekte" Figur, woraus man folgerte, dass Gott die Planetenbahnen als Kreisbahnen eingerichtet habe; Kepler hatte Mühe, sich bei der Analye der Daten Tycho Brahes von dieser Vorstellung zu lösen

[3]gemeint ist die Gauß-Verteilung

entwickelten Intelligenztest; die Häufigkeiten der Testwerte ("Scores") entsprechen offenbar nicht einer Gauß-Verteilung, und die obere der beiden Verteilungen könnte mit der Hypothese einer log-Normalerteilung kompatibel sein. Die untere Verteilung entspricht dagegen nicht einer log-Normalverteilung. Trotzdem kann auch hier die "wahre" Verteilung eine log-Normalverteilung sein, die aber wegen einer nicht-repräsentativen Stichprobe oder wegen suboptimal gewählter Aufgaben (Schwierigkeiten) von der log-normalen Verteilung abweicht.

Die Aufgabe, eine theoretische Verteilung, sei sie nun eine Gauß- oder eine log-Normalverteilung, an eine Häufigkeitsverteilung anzupassen, ist im Prinzip nicht schwer; es wird später noch explizit darauf eingegangen. Hier sei angemerkt, dass man den Fit, also das Ausmaß der Anpassung, oft verbessern kann, indem man die Aufgaben anders bepunktet oder sie durch andere nach Maßgabe ihrer Schwierigkeit austauscht, – bei einer normalen Klassenarbeit wird diese Möglichkeit im Allgemeinen nicht gegeben sein. Aber die Frage nach den Schwierigkeiten der Aufgaben wirft eine grundsätzliche Frage auf. Wenn es irgend geht wird die Verteilung so angepasst, dass die Note "3" die Note mit der größten Häufigkeit wird, und die Intervalle für die Noten "2" und "4" in guter Näherung die gleich Breite haben wie die für die "3" (auf die Intervalle für die Noten "1" und "5" wird später noch genauer eingegangen). Was auf den ersten Blick als völlig natürlich erscheinen mag, hat den großen Nachteil, dass dem möglichen Unterschied zwischen den durchschnittlichen Leistungsniveaus verschiedener Klassen nicht nicht Rechnung getragen wird. Es ist ja möglich, dass als Resultat von Stichprobeneffekten das durchschnittliche Begabungsniveau und damit auch das Leistungsniveau in einer Klasse höher ist als in einer anderen Klasse. Darüber hinaus kann der Lehrerin oder der Lehrer durchaus einen spezifischen Einfluß auf das Leistungsniveau der Klasse haben, wie das Beispiel der Frau Czerny zeigt. Wird nun vorgeschrieben, dass in beiden Klassen die "3" die durchschnittliche Note ist, so wird im Vergleich zur schlechteren Klasse die bessere Klasse schlechter benotet. Diese Betrachtung gilt nicht nur für Schulklassen, sondern für ganz Schulen und am Ende für ganze Länder: die Schwierigkeiten für dieselben Aufgaben sind für unterschiedliche Teilpopulationen nicht identisch. Das kann auch der Fall sein, wenn alle "Parallellehrer" gute LehrerInnen im Sinne Sabine Czernys sind. Sollen die SchülerInnen über das Anpassen von Wahrscheinlichkeitsverteilungen benoted werden, so muß eine Verteilung gefunden werden, die eine populationsunabhängige Benotung und damit auch eine populationsunabhängige Abschätzung der Schwierigkeiten der Aufgaben ermöglicht.

Logistische Verteilung und spezifische Objektivtät: Das Problem der Vergleichbarkeit von Noten ergibt sich u.a. aus einer Konfundierung der Abschätzung der Schwierigkeit der einzelnen Aufgaben mit der Abschätzung der Kompetenz der einzelnen SchülerInnen. Also stellt sich die Aufgabe, eine Verteilung zu finden, bei der diese Konfundierung nicht stattfindet. Der dänische Statistiker Georg Rasch (1901 – 1980) fand in der logistischen Verteilung[4] eine Verteilung, die

[4]Der Ausdruck 'logistisch' bezieht sich auf einen stochastischen Wachstumsprozess, den der

der Forderung nach voneinander unabhängiger Schätzung der Aufgabenschwierigkeiten einerseits und Kompetenzparametern der Probanden andererseits zu genügen scheint. Auf die mathematischen Details dieser Verteilung wird weiter unten eingegangen; hier kann angemerkt werden, dass sich die Dichte- und damit auch die Verteilungsfunktionen der Gauß- und der logistischen Verteilung kaum voneinander unterscheiden s. Abbildung 4; der Unterschied zwischen den beiden Verteilungen besteht in der Art und Weise, wie die Lokation der Verteilungen auf der X-Skala und ihre Variation von den Parametern μ und σ abhängen.

Wie sich zeigen läßt gibt es gleichwohl große Unterschiede bezüglich der zu schätzenden Parameter. Die beiden Verteilungen sind ein Beispiel für den Sachverhalt, dass Verteilungen anhand eines gegebenen Datensatzes so gut wie gar nicht voneinander unterschieden werden können, der Effekt der Werte der freien Parameter μ und σ sich aber, wie bereits angemerkt, sich auf die Lokation und Variation sehr unterschiedlich auswirken. Das ursprünglich von Rasch entwickelte Testmodell ist einigermaßen restriktiv und ist in den letzten Jahrzehnten so verallgemeinert worden, dass es allgemeineren praktischen Anforderungen genügt, so dass es von einigen Forschern als eine Art von kanonischem Messmodell propagiert wird. In der Diskussion (Abschnitt 3) muß allerdings darauf hingewiesen werden, dass es auch bei diesen verallgemeinerten Modellen Grenzen der Anwendbarkeit gibt.

2 Zur Konstruktion von Notenskalen

Grundbegriffe der Statistik werden im Anhang, Abschnitt 4.1 kurz dargestellt. Es werde mit einem grundsätzlichen Sachverhalt begonnen:

Satz: Es sei Y eine zufällige Veränderliche mit der Verteilungsfunktion $G(y) = P(Y \leq y)$ und es sei $p_{yj} = G(y_j)$, $j = 1, \ldots, k$. Weiter sei X eine andere zufällige Veränderliche mit der Verteilungsfunktion $F(x) = P(X \leq x)$. Als Verteilungsfunktionen sind G und F monoton mit y bzw. x wachsende Funktionen, so dass die inversen Funktionen G^{-1} und F^{-1} existieren. Dann ist $F^{-1}(p_{yj}) = x_j = \inf\{x|F(x) = p_{yj}\}$, wobei inf der kleinste Wert der Menge $\{x|F(x) = p_{yj}\}$ ist.

Die etwas pompöse Bezeichnung des beschriebenen Sachverhalts als "Satz" dient nur dazu, sich im folgenden Text leichter bzw. unmißverständlich auf den darin ausgedrückten Sachverhalt beziehen zu können. Es wird ja nur von der Tatsache Gebrauch gemacht, dass für eine beliebige Verteilungsfunktion $F(x) = P(X \leq x)$ stets $0 \leq F(x) \leq 1$ gilt. Hat man also ein $p \in (0, 1)$ gegeben, so

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; γ ist eine Normierungskonstante.

existiert stets ein x aus dem Definitionsbereich von X derart, dass $p = F(x)$, so dass $x = F^{-1}(p)$. Eben weil F beliebig gewählt werden kann, kann man zusätzlich annehmen, dass für irgendeine andere Verteilungsfunktion $G(y) = P(Y \leq y)$ einer anderen zufälligen Veränderlichen Y der spezielle Wert $Y = y$ existiert derart, dass $G(y) = p$, und $y = G^{-1}(y)$. So sei $G(y)$ die Verteilung der Punkte pro Klassenarbeit in einer Klasse; $p_y = G(y)$ sei Anteil der SchülerInnen in der Klasse, die $Y \leq y$ Punkte erzielt haben, – also maximal y Punkte. Die Verteilungsfunktion G ist i. A. nicht bekannt, aber es können natürlich Annahmen über G gemacht werden, z.B. dass G durch F gegeben sei. Dann kann man den x-Wert bestimmen, für den $F(x) = p$ ist, d.h. man kann $F^{-1}(p) = x$ bestimmen. So kann man den Anteil p_5 der SchülerInnen *vorgeben*, deren Leistung als "ungenügend" betrachtet werden soll. Unter der Annahme, dass die Punkteverteilung durch F gegeben ist, läßt sich nun der Wert x_4 bestimmen, für den $p_5 = P(X \leq x_4) = F(x_4)$ gilt: x_4 ist durch $x_4 = F^{-1}(p_5)$ gegeben. Ebenso läßt sich p_4 annehmen: dies ist der Anteil der SchülerInnen, die entweder durchgefallen oder bestenfalls die Note "4" erhalten sollen. Es ist dann $x_3 = F^{-1}(p_4)$, und $p_4 = P(X \leq x_4)$, etc. Der Anteil der SchülerInnen mit der Note "4" ist dann durch

$$P(x_4 < X \leq x_3) = F(x_3) - F(x_4) = p_4 - p_5$$

gegeben. Man kann insbesondere annehmen, dass X Gauß-verteilt ist. Z seien die standardisierten Werte von X, und $F_z(z) = P(Z \leq z)$. Dann sei $p_j = F_Z(z_j)$ und man erhält $z_j = F_Z^{-1}(p_j) = z_j$, so dass man speziell

$$P(z_4 < Z \leq z_3) = F_Z(z_4) - F_Z(z_3) = p_4 - p_5 \tag{2.1}$$

erhält. Die tatsächlichen Punktwerte x_4 und x_3 erhält man wegen $z_j = (x_j - \bar{x})/s_x$ gemäß der Beziehung $x_j = s_x z_j + \bar{x}$.

Natürlich werden die Intervallbreiten $|x_j - x_{j-1}|$ nicht für alle j identisch sein, so dass man sagen kann, dass die Skala für die Noten nicht notwendig eine Intervallskala ist.

2.1 Die Gauß-Verteilung und die log-Normalverteilung

Fall 1: ohne Punkteverteilung Eine einfache Anwendung des obigen Satzes hat man, wenn Schülerleistungen intuitiv ("ganzheitlich", ohne Bezug auf eine "objektive" Punkteverteilung) bewertet werden. Ein Lehrer macht bei der Benotung von einer implizit definierten Wahrscheinlichkeitsverteilung G auf einem subjektiven Kontinuum Y Gebrauch; 'implizit' heißt hier, dass er keine explizite Definition von G zugrunde liegt, der Lehrer verteilt die Noten einfach so, wie er es für richtig hält. Am Ende hat er n_1-mal die note "1", n_2-mal die Note "2" etc, und n_5-mal die note "5" vergeben. Wenn der Lehrer b SchüerInnen inn der Klasse hat, so hat er die Noten mit den relativen Häufigkeiten $p_1/n, \ldots, p_5 = n_5/n$ vergeben. Nun *postuliert* der Lehrer, dass die Leistungen, die der Benotung zugrundeliegen, in der Klasse normalverteilt seien, weil er glaubt, Leistungen seien eben

grundsätzlich Normal-, d.h. Gauß-verteilt. Damit postuliert er die Existenz einer "latenten", also nicht explizit von ihm verwendeten kontinuierlichen Skala X, die so in Bereiche eingeteilt werden kann, dass sie den von ihm vergebenen Noten entsprechen. Es gibt also einen Wert x_1 derart, dass $P(X > x_1) = p_1 = 1-F(x_1)$, es gibt weiter einen Wert x_2 derart, dass

$$p_2 = P(x_2 \leq X < x_1) = F(x_1) - F(x_2),$$

etc, bis schließlich

$$p_5 = P(X \leq x_4) = F(x_4).$$

x_1 könnte er nun bestimmen, indem er die Inverse von $F(x_1) = q_1 = 1 - p_1$ bestimmt, – aber dazu müßte er die "wahren" Werte μ und σ der Gauß-Verteilung von X kennen; das Gleiche gilt für die übrigen p_j-Werte. Der Lehrer besinnt sich nun auf den Sachverhalt, dass man eine zufällige Veränderliche stets standardisieren kann, d.h. man kann die Transformation

$$Z = \frac{X - \mu}{\sigma}$$

betrachten, für die

$$P(Z \leq z) = P\left(Z \leq \underbrace{\frac{x - \mu}{\sigma}}_{z}\right) = P\left(Z \leq x\frac{1}{\sigma} - \frac{\mu}{\sigma}\right)$$

gilt. Dies bedeutet, dass er statt $F(x)$ die Verteilungsfunktion $F_Z(z)$ betrachten kann, aus der er die z-Werte $F_Z^{-1}(p) = z$ gewinnen kann, die bis auf eine beliebige lineare (affine) Transformation den x-Werten entsprechen. Es ist, als wolle er die Wärme der Luft messen, wozu er die Celsius-Skala verwendet, die sich von einer beliebigen anderen Skala (z.B. der Fahrenheit-Skala) nur durch eine lineare (affine) Transformation unterscheidet. Da die lineare Transformation umkehrbar eindeutig ist:

$$z = \frac{x - \mu}{\sigma} = x\frac{1}{\sigma} - \frac{\mu}{\sigma} \Longleftrightarrow x = z\sigma + \mu,$$

genügt es, die z-Werte als Skala zu verwenden: die *Relationen* zwischen den x_j-Werten entsprechen denen zwischen den korrespondierenden z_j-Werten.

So sei $q_1 = P(Z \leq z_1) = F_Z(z_1) = 1 - p_1$ die relative Häufigkeit aller Noten schlechter als "1", und z_1 ist durch $F_Z^{-1}(q_1)$ gegeben. p_2 sei die relative Häufigkeit der Note "2", der die Wahrscheinlichkeit

$$p_2 = P(z_2 \leq Z \leq z_1) = F_Z(z_1) - F_Z(z_2)$$

entspricht[5]. Dann hat man $F_Z(z_2) = F_Z(z_1) - p_2$ und man erhält

$$z_2 = F_Z^{-1}(F_Z(z_1) - p_2).$$

[5]exakterweise müßte man von den p_j als den Schätzungen der Wahrscheinlichkeiten sprechen, was hier der Einfachheit halber nicht getan wird.

Für die Note "3" gilt $p_3 = P(z_3 \leq Z < z_2) = F_Z(z_2) - F_Z(z_3)$, so dass man

$$F_Z(z_3) = F_Z(z_2) - p_3 \Rightarrow z_3 = F_Z(^{-1}(p_3 - F_Z(z_2))$$

erhält. Für z_4 erhält man schließlich

$$z_4 = F_Z^{-1}(p_4 - F_Z(z_3))$$

und $p_5 = F_Z(z_4)$. Der Lehrer hat nun Skalenwerte für die Grenzen der Intervalle für die Noten auf einer Skala gefunden, deren Werte Gauß-verteilt sind. Die Intervalle werden im Allgemeinen nicht gleich groß sein, so dass er nicht die beliebte Intervallskala hat, – aber an der Hypothese der Gauß-Verteilung wird nicht gerüttelt.

Der Lehrer hätte aber auch eine andere Verteilung postulieren können, zum Beispiel die log-Normalverteilung. Dann hätte er andere Intervallgrenzen gefunden, aber die Hypothese der log-Normalverteilung wäre nicht in Frage gestellt worden. Er hätte nun, wie im Falle des Postulats der Gauß-Verteilung, den zu erwartenden Befund von ungleich großen Intervallen auf der Skala mit lognormalverteilten Werten interpretieren, – aber da fällt einem immer etwas ein. Tiefer gehende Information läßt sich aus diesen Rechenübungen nicht herleiten, und das Gefühl, mit seiner Annahme über zugrunde liegende Verteilung Recht zu haben, mag schön sein, hat aber keine Bedeutung.

Fall 2: mit Punkteverteilung Nun werde der Fall betrachtet, dass für die Antworten auf Fragen bzw. Aufgaben nach irgendwelchen Kriterien Punkte vergeben werden. Die LehrerInnen müssen nun (ministerieller Erlaß!) das metaphysische Postulat, Noten seien stets Gauß-verteilt, akzeptieren. Das Anpassen einer Gauß-Verteilung an die Daten (Häufigkeiten von Gesamtpunktzahlen) gelingt wieder, und es ist möglich, gleich große Intervalle für die Noten zu berechnen.

Dazu beginnt man zum Beispiel, den Anteil q_1 der Einser-SchülerInnen festzulegen, wobei wieder

$$q_1 = P(Z > z_1)) = 1 - F_Z(z_1) = 1 - p_1$$

gilt. Der Wert von z_1 wird über die Beziehung $z_1 = F_Z^{-1}(q_1)$ bestimmt. Mit $q_1 = 1 - p_1$ hat man wegen der Symmetrie der Gauß-Verteilung auch den Anteil p_5 der Durchgefallenen festgelegt. Es gilt ja

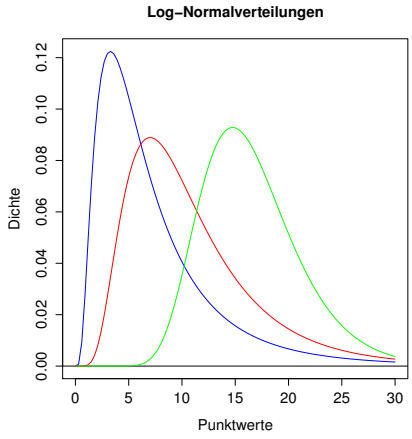
$$P(Z > z_1) = P(Z \leq z_4) = P(Z \leq -z_1).$$

Man dividiert nun das Intervall $[z_4, z_1)$ durch 3 und erhält die Intervallbreite[6]

$$\Delta = \frac{z_1 - z_4}{3},$$

[6] Δ ist das "große"' griechische D und steht gemeinhin für "Unterschied" oder "Differenz".

Abbildung 2: Verschiedene log-Normalverteilungen



und berechnet dann

$$z_3 = z_4 + \Delta, \quad z_2 = z_3 + \Delta,$$

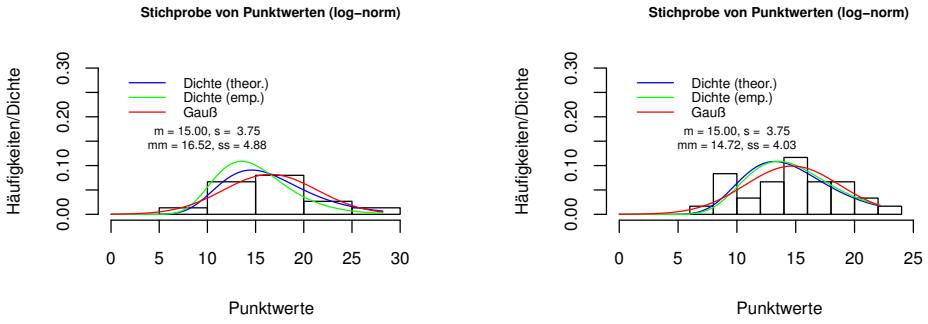
und der Wert von $z_1 = Z_2 + \Delta$ ist ja bereits bekannt. Da $z_j = (x_j - \bar{x})/\hat{s}$ ($\bar{x}$) ist die durchschnittliche erreichte Punktzahl und s^2 ist die Stichprobenvarianz der erreichten Punktzahlen) erhält man daraus die Intervallgrenzen für die Noten in Bezug auf die Häufigkeitkeitsverteilung der Punkte:

$$x_j = \hat{s} z_j + \bar{x}, \quad j = 1, 2, 3, 4$$

SchülerInnen mit einem Punktwert größer als x_1 bekommen eine "1", SchülerInnen mit einem Punktwert zwischen x_1 und x_2 bekommen eine "2", Punktwerte zwischen x_2 und x_3 werden in die Kategorie "3" sortiert, Punktwerte zwischen x_3 und x_4 gehören zur Kategorie "4", und wer weniger als x_4 Punkte hat bekommt eine "5" und ist durchgefallen. Die Noten scheinen nun Gauß-verteilt zu sein mit gleich großen Intervallen zumindest für die Noten "2", "3" und "4". Natürlich kann man den Anteil der Durchgefallenen und damit den Wert von z_4 unabhängig vom Anteil der Einser-SchülerInnen festlegen, aber die Notenskala wird dann nicht mehr symmetrisch zur Note "3" liegen. In derselben Weise läßt sich eine Notenskala für die log-Normalverteilung bestimmen.

Wirklich gewonnen hat man mit einer solchen Übung aber kaum etwas, da die Abbildung der p_j-Werte auf bestimmte z_j-Werte etwa einer Gauß-verteilten (oder log-normalverteilten), standardisierten zufälligen Veränderlichen Z *stets* möglich ist und deswegen über die tatsächliche Verteilung der Leistungen nichts ausgesagt wird. Es sei denn, man ist aus irgendwelchen Gründen überzeugt, dass Leistungen Gauß-verteilt (log-normalverteilt) sein *müssen*, weil die Gauß-Verteilung (log-Normalverteilung) ein Naturgesetz sei. Aber das ist sie nicht, es existiert

Abbildung 3: log-normalverteilte ("theor") Stichproben mit angepassten Gaußverteilungen; die Populationsparameter sind identisch. Die bestpassende Gauß-Verteilung ist gegenüber der Schätzung der wahren log-Normalverteilung nach rechts verschoben



kein hinreichend bestätigtes Modell kognitiver Aktivität, aus dem ein solches Modell folgen würde. Die Bewertung der Antworten der SchülerInnen durch Punkte hat sicher nur die Messqualität einer Rangskala ("Ordinalskala"), denn kodiert man z.B: eine korrekte Antwort mit 1, eine falsche oder gar keine Antwort mit 0, so greift man mit einer 1 ja keine Einheit auf der hypothetischen, latenten Kompetenzskala ab, schon weil korrekte Antworten auf unterschiedlich schwierige Aufgaben alle mit einer 1 kodiert werden. Dieses Argument überträgt sich auf eine Kodierung, bei der eine Aufgabe entweder mit 0 kodiert wird, wenn keine oder eine falsche Antwort gegeben wurde, und bei der ein Punkt vergeben wurde, wenn der Lösungsansatz korrekt gewählt wurde, und bei der darüber hinaus zwei Punkte vergeben wurden, wenn der Ansatz auch noch korrekt durchgeführt wurde. Vielleicht wird für die korrekte Durchführung auch nur ein halber Punkt vergeben, so dass für die Aufgabe maximal 1.5 Punkte vergeben werden konnten. Kann man sagen, dass die Durchführung eines Ansatzes nur eine halbe Einheit auf dem latenten Kompetenzkontinuum bedeutet? Wohl eher nicht, weil das Kontinuum selbst ja gar nicht gegeben ist und eine explizite Beziehung zwischen der Ausprägung der Kompetenz und der Bewertung durch Punkte nicht gegeben ist, und weil ein Punkt oder ein halber Punkt verschiedene Differenzen zwischen den Kompetenzanforderungen verschiedener Aufgaben bedeuten kann. Man könnte argumentieren, dass unterschiedliche Schwierigkeiten, also unterschiedliche Häufigkeiten korrekter Antworten, zweier Aufgaben nur aussagen, dass eine von ihnen schwieriger ist als die andere, aber die unterschiedlichen Häufigkeiten noch keine *metrische* Information bezüglich der beiden Aufgaben liefern: ist N_A die Anzahl SchülerInnen, die A korrekt beantwortet haben, und N_B die Anzahl der SchülerInnen, die B korrekt beantwortet haben, bedeutet dann etwa der Fall $N_B = N_A + 1$, dass B um eine Kompetenzeinheit leichter ist

als A? Kaum, denn wie gerade schon argumentiert wurde bedeutet jede Kodierung mit einer Eins ja nicht einen Zuwachs um eine Kompetenz*einheit*. Da ein Unterschied von einem Probanden bei der Bestimmung der Schwierigkeiten auch zufällig zustande kommen kann, könnte man eine Differenz von $k > 1$ Punkten fordern, die "'signifikant" sein soll, aber das Argument der mangelnden Metrik (Intervallskalenqualität) der Häufigkeiten ist damit nicht ausgehebelt.

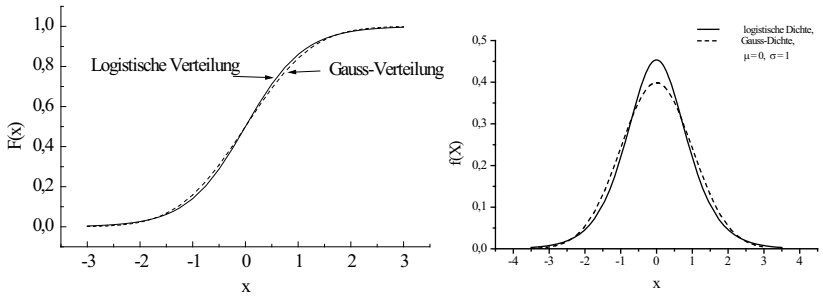
Alle Betrachtungen übertragen sich auf die log-Normalverteilung. Ist F_L die Verteilungsfunktion einer geeignet gewählten log-Normalverteilung, so lassen sich die relativen Häufigkeiten p, mit denen Punkte in bestimmten Intervallen der Punkteskala finden, durch $F_L^{-1}(p) = x_p$ auf Werte einer log-verteilten zufälligen Veränderlichen abbilden. Damit hat man nicht gezeigt, dass die Noten log-verteilt sind, sondern nur, dass auch die log-Normalverteilung verwendet werden kann, um eine Notenskala nach Maßgabe einer Verteilung zu konstruieren, ohne dass etwas über die Validität[7] der Skala aussagen zu können. Abbildung 2 zeigt verschiedenen log-Normalverteilungen; offenbar kann eine log-Normalverteilung einer Gauß-Verteilung ähneln. Abbildung 3 zeigt zwei Stichproben aus ein und derselben log-normal veteilten Population, für die auch jeweils eine Gauß-Verteilung an die Daten angepasst wurde. Die Gauß-Verteilungen sind gegenüber der log-Normalverteilung stets nach rechts verschoben. Man benötigt also tendenziell mehr Punkte als bei der korrekten Wahl der log-Normalverteilung, wenn die Intervallgrenzen für die Notengebung berechnet werden.

2.2 Item-Response-Modelle, insbesondere das Rasch-Modell

Der dänische Statistiker Georg Rasch (1901 – 1980) schlug 1961 ein Testmodell vor, mit dem das Problem der Vergleichbarkeit der Noten im Prinzip gelöst werden kann; das Model ist inzwischen zur Klasse der *Item-Response-Modelle* verallgemeinert worden, mit denen eine Reihe von Einschränkungen des Rasch-Modells aufgehoben werden können. (– Literatur). Hier kann es aber nur um das Prinzip gehen, nämlich die voneinander unabhängige Schätzung von Schwierigkeits - und der Personenparameter (dies sind im gegebenen Zusammenhang die Kometenzen der SchülerInnen. Diese Unabhägigkeit wird als *spezifische Objektivität*) bezeichnet. Das Modell beruht auf der Annahme eine speziellen Verteilung, namlich der logistischen Verteilung. Sie ähnelt der Gauß-Verteilung, ist mathematisch aber so definiert, dass sie die Möglichkeit der spezifischen Objektivität impliziert. Es wird nicht gefragt, ob sie logistische Verteilung die "natürliche" Verteilung in dem Sinne ist, dass sie Eigenschaften kognitiver Prozesse abbildet. Im Zentrum steht vielmehr der rein messtheoretische Aspekt der Repräsentation der Schwierigkeits- und Personenparameter.

[7] Validität ist ein technischer Term aus der Klassischen Theorie psychometrischer Tests (KTT); er bedeutet 'Gültigkeit' des Tests, operationalisiert als Korrelation zwischen Testwert und tatsächlicher Ausprägung des gemessenen Merkmals.

Abbildung 4: Verläufe der standardisierten Gauß- und der logistischen Verteilung; Verteilungsfunktionen links, Dichten rechts.



Zunächst werde die logistische Verteilung eingeführt. Sie ist durch

$$F(x) = P(X \leq x) = \frac{1}{1 - \exp\left(-\frac{(x-\mu))\beta}{\sigma}\right)}, \quad \beta = \pi/\sqrt{3}. \tag{2.2}$$

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

$$P(X \leq x) = \frac{1}{1 + \exp(-(x-\mu)\gamma)} \tag{2.3}$$

μ ist der Erwartungswert von X, d.h. der Mittelwert über alle möglichen Realisierungen von X, und σ ist die korrespondierende Standardabweichung.

Die Verteilung kann für einen individuellen Probanden betrachtet werden. So sei $\theta = \mu$ die "wahre" Kompetenz des Probanden, d.h. eines Schülers, um den die gemessene Kompetenz fluktuiert. Weiter sei κ_g die Schwierigkeit einer, der g-ten Aufgabe. κ_g ist ein Maß für die Kompetenz, die aufgebracht werden muß, um eine Frage zu beantworten bzw. um eine Aufgabe zu lösen. Der Einfachheit halber werde angenommen, dass Fragen entweder korrekt oder inkorrekt (möglicherweise gar nicht) beantwortet werden, dass eine Aufgabe entweder gelöst oder nicht gelöst wird. Dann kann eine *Indikatorvariable* χ_g eingeführt werden:

$$\chi_g = \begin{cases} 0, & g\text{-te Frage nicht beantwortet} \\ 1, & g\text{-te Frage beantwortet} \end{cases} \tag{2.4}$$

wobei natürlich für "Frage beantwortet" oder "nicht beantwortet" auch "Aufgabe gelöst" bzw. "'nicht gelöst" stehen kann. $\{X \leq \kappa_g|\theta\}$ ist dass zufällige Ereignis, dass zum Zeitpunkt der Messung die g-te Frage nicht beantwortet werden kann, wenn die Fähigkeit eines Probanden den Wert θ hat, und das korrespondierende Komplementärereignis ist $\{X > \kappa_g|\theta\}$. Die Wahrscheinlichkeit, dass ein Proband

die g-te Frage korrekt beantwortet ist dann nach (2.3)

$$\begin{aligned} P(\chi_g &= 1|\theta) = 1 - P(X \leq \kappa_g|\theta) \\ &= 1 - \frac{1}{1+\exp(-(\kappa_g - \theta)} = 1 - \frac{1}{1+\exp(\theta - \kappa_g))} \end{aligned} \tag{2.5}$$

wobei der Faktor γ bereits in die Parameter κ_g und θ absorbiert wurde (d.h. man definiert $\kappa_g' = \gamma\kappa_g$ unf $\theta' = \gamma\theta$ und benennt κ_g' und θ' dann wieder in κ_g und θ um). Man sieht durch eine elementare Rechnung, dass (2.5) in der Form

$$P(\chi_g = 1|\theta) = \frac{\exp(\theta - \kappa_g)}{1+\exp(\theta - \kappa_g)}. \tag{2.6}$$

Spezifische Objektivität Dass die Definition von $P(\chi_g = 1|\theta)$ (diese Wahrscheinlichkeit wird ja im"klassischen" Ansatz, bei dem die Gauß- oder die lognormalverteilung angenommmen wird, betrachtet) durch die logistische Verteilung die spezifische Objektivität der Schätzungen für κ_g und θ impliziert, sieht man leicht. Die Gleichung (2.6) impliziert

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

so dass

$$\log \frac{P(\chi_g = 1|\theta)}{P(\chi_g = 0|\theta)} = \theta - \kappa_g \tag{2.7}$$

Für die g-te Frage und zwei Probanden mit den Parametern θ und θ' erhält man dann

$$\log \frac{P(\chi_g = 1|\theta)}{P(\chi_g = 0|\theta)} - \log \frac{P(\chi_g = 1|\theta')}{P(\chi_g = 0|\theta')} = \theta - \theta' + \kappa_g - \kappa_g = \theta - \theta' \tag{2.8}$$

Für zwei Probanden mit identischem θ-Wert erhält man für zwei verschieden schwierige Aufgaben

$$\log \frac{P(\chi_g = 1|\theta)}{P(\chi_g = 0|\theta)} - \log \frac{P(\chi_{g'} = 1|\theta)}{P(\chi_{g'} = 0|\theta)} = \theta - \theta + \kappa_g - \kappa_{g'} = \kappa_g - \kappa_{g'}. \tag{2.9}$$

Diese Gleichungen zeigen, dass es im Prinzip möglich ist, die Personen- und die Itemparameter unabhängig voneinander zu schätzen. Die tatsächliche Schätzung der Parameter für eine gegebene Klausur ist relativ aufwändig, allerdings existieren Programme zur freien Verfügung (Literatur), und die Berechnungnen können auf jedem Laptop durchgeführt werden.Steyr & Eid (1993) liefern eine sorgfältige Einführung in die Theorie der IRT-Modelle.

3 Diskussion

"Kompetenz" ist in den meisten Schulfächern ein komplexer Begriff. Kompetenz ist eine Mischung von Wissen und Umgang mit dem Wissen, gleich ob es sich um Fächer wie Deutsch, Geschichte, Fremdsprachen, oder Mathematik, Physik, Chemie etc handelt. Die Ausprägung von Kompetenz wird bei SchülerInnen auf einer 1-dimensionalen Skala, der Skala der Schulnoten, gemessen. Man kann vermuten, dass LehrerInnen die verschiedenen Subkompetenzen nach Art der multiplen Regression[8] zusammenfassen, bei der die Note aus einer Kombination von Maßen für die Subkompetenzen zusammengefasst werden. Die Noten sollen möglichst objektiv und transparent vergeben werden, und dazu dient die Vergabe von Punkten, deren Verrechnung am Ende eine Note ergibt. Dabei soll die Kompetenzskala zumindest für die Noten "2" bis "4" (oder "2" bis "5", wenn auch noch eine "6" vergeben werden soll) möglichst in gleichgroße Abschhnitte unterteilt werden. Dazu wird eine Häufigkeitsverteilung der Gesamtzahlen n_i der Punkte, die von den SchülerInnen $i = 1, \ldots, n$ erworben wurden erstellt, die dann in Intervalle unterteilt wird derart, dass eine SchülerInn, deren Gesamtzahl in einem bestimmten Intervall liegt, die zu diesem Intervall korrespondierende Note bekommt. Dabei kann es durchaus sein, dass in einem Fach Punkte für einzelne Aspekte der Kompetenz vergeben werden, die dann nach bestimmten Regeln (u.U. wieder analog zur multiplen Regression) zu einer Gesamtnote zusammengefasst werden. Wie auch immer verfahren wird, einige grundsätzliche Probleme bleiben immer.

Diese Probleme ergeben sich aus der Forderung, dass die Benotung eine Messung sein soll, die wenn irgend möglich mehr als nur eine Rangordnung der unterschiedlich ausgeprägten Kompetenzen sein soll. Die erwähnte Gleicheit der Intervalle auf der Notenskala bedeutet in der Sprache der Messtheorie, dass die Notenskala eine Intervallskala ist. Ein bekanntes Beispiel für diesen Skalentyp sind Temperaturskalen. Der Nullpunkt wird willkürlich festgelegt (bei der Celsius-Skala wird, etwas lose formuliert, der Nullpunkt mit dem Gefrierpunkt des Wassers gleichgesetzt). Die Skala erstreckt sich natürlich in Minusgrade ebenso wie Temperaturen über 100 Grad [Celsius]. Die Fahrenheit-Skala ist ebenfalls eine Intervallskala, und die Temperatur auf der Fahrenheit-Skala ist die lineare bzw affine Transformation der Form $y[F^o] = ax[C^o] + b$, a, b Parameter, mit der Celsius-Skala verbunden. Beide Skalen können in die Kelvin-Skala transformiert werden, bei der ein echter Nullpunkt definiert ist: 0^o Kelvin repräsentiert einen Zustand von Null Energie, der de facto nicht erreichbar ist, - auf die Details muß hier nicht eingegangen werden.

Bei Kompetenzmessungen hat aber der Begriff des Nullpunkts keine schar-

[8] Eine Variable Y – hier die Note – ergibt sich diesem Ansatz entsprechend in der Form $Y = b_0 + b_1 X_1 + \cdots + b_n X_n$, wobei die $X_1, \ldots, X_n$ Bewertungen der Subkompetenzen repräsentieren und die $b_0, b_1, \ldots, b_n$ "Gewichte" sind, mit denen die Subkompetenzen in das Gesamturteil eingehen. Die hier geschilderte Problematik überträgt sich auf die "Prädiktoren" X_j und die implizit geschätzten Gewichte b_j.

fe Bedeutung mehr. Wenn von einem Schüler gesagt wird, im Fach Mathematik "könne er nichts" , so ist das eine umgangssprachliche Äußerung, die nur aussagt, dass der Schüler kaum noch etwas im Mathematikunterricht versteht, obwohl er die Regeln der Arithmetik beherrschen kann und ihm auch der Satz des Pythagoras nicht fremd sein muß. Die Bedeutung von "nichts können" ist also relativ zum Leistungsstandard seiner (Schul-)Klasse zu sehen. In einer Klassenarbeit kann eine gegebene Aufgabe vielleicht mit maximal 2 Punkten belohnt werden: einen Punkt für den richtigen Ausatz, einen für den weiteren Punkt für die korrekte Durchführung des Ansatzes. Die Frage ist nun, ob ein Punkt jeweils einen gleichgroßen Abschnitt auf einer latenten Kompetenzskala definiert. Diese Frage ist insofern schwierig zu beantworten, weil nicht wirklich klar ist, was mit der Kompetenz des Ansatzfindens einerseits und der des Durchführens eines Ansatze gemeint ist: auf dem umganspsrachlichen Niveau weiß man intuitiv, was gemeint ist oder gemeint sein könnte, aber man hat keine Vorstellung davon, wie die jeweiligen Subkompetenzen auf eine numerische Skala projiziert werden können. Ebenso versteht man umgangssprachlich, was mit der Summe und dem Mittelwert von Punkten gemeint ist, ohne aber deshalb zu wissen, wie diese Statistiken auf eine Kompetenzskala abgebildet werden können. Dazu müßte auch explizit geklärt werden, wie eine derartige latente Skala zu definieren ist. Vielleicht existiert gar keine derartige Skala. Kompetenz ist eine Mischung von einerseits grundlegenden und andererseits erworbenen kognitiven Fähigkeiten, aber eine explizite Definition 'Kompetenz' fehlt im Allgemeinen. Man wählt also einen adhoc-Ausweg (man könnte auch von einem Deus-ex-machina-Ausweg sprechen), indem man eine auf einem Kontinuum von reellen Zahlen definierte Wahrscheinlichkeitsverteilung, z.B. die Gauß-Verteilung wählt und von dem am Anfang des Abschnitts 2 beschriebenen "Satz" Gebrauch macht. Eine solche Verteilung hat aber ungefähr den Status der Axiome des absoluten Raums und der gleichmäßig und unabhängig davon verfließenden Zeit in der klassischen Physik (Newton): Es ist nicht klar, warum sie gelten sollen, aber in der Physik helfen sie bei der Formulierung von Naturgesetzen bzw. Bewegungsgleichungen. Aber bei der Benotung von Schulleistungen ist bei der Wahl einer Wahrscheinlichkeitsverteilung nicht nur unklar, warum sie die "wahre" Verteilung sein soll, sie hilft auch nicht bei der Verteilung "wahrer" Noten, denn der wirklich kritische Aspekt dieser Wahl ist die Gleichsetzung der durchschnittlchen Leistung innerhalb *einer* Klasse mit der Note "3": je nach Leistungsniveau in einer Klasse erfolgt dadurch eine Unterbewertung von im Durchschnitt besseren SchülerInnen ein einer Klasse bzw. eine Überbewertung von durchschnittlich schlechteren SchülerInnen in einer anderen Klasse. Diesem Nachteil einer verteilungsorientierten Benotung steht bei Wahl der Gauß-Verteilung bestenfalls der Vorteil gleichgroßer Intervalle für die verschiedenen Noten gegenüber. Da viel für die log-Normalverteilung als "wahrer" Verteilung der Leistungsfähigkeit spricht und die log-Normalverteilung eine links-steile Verteilung ist (die Mehrheit der Leistungen befindet sich im unteren Bereich der Skala, s. die Abbildung 1) ergibt sich allerdings die Frage, ob gleichgroße Intervalle tatsächlich zum Ideal einer Notenskala gehören sollen.

Umgangssprachlich ist auch klar, was mit einer "schwierigen" Aufgabe gemeint ist: sie ist "schwierig", wenn nur wenige die Kompetenz haben, sie lösen zu können. Jemand, der eine schwierige Aufgabe lösen kann, ist jemand, der mehr Kompetenzen für das Lösen dieser Art von Aufgaben hat als die meisten anderen Mitbürger. Diese Auffassung führte dann auch zur Definition der Schwierigkeit einer Aufgabe im Rahmen der Klassischen Testtheorie (KTT) als Anteil der Personen, die eine Aufgabe lösen bzw. nicht lösen. Die KTT ist eine statistische Theorie diagnostischer Tests, zu deren Axiomatik die Annahme der Gauß-Verteilung des jeweils gemessenen Merkmals gehört (Lord & Novik, (1971/2008)). Aber bereits die Definition des Begriffs der Schwierigkeit einer Aufgabe impliziert, dass die Testwerte für verschiedene Populationen nicht miteinander vergleichbar sind. So kann der Zufall es wollen, dass in einer Schulklasse mehrere Schüler mit einem IQ größer als 130 sind, so dass in dieser Klasse fast alle Aufgaben gelöst werden, d.h. die Aufgaben erscheinen als "leicht". In der Parallelklasse sind aber, wieder wegen der allgegenwärtigen Stichprobenfluktuationen, eher durchschittliche Schüler, und hier wird ein größerer Anteil von Aufgaben nicht gelöst, so dass die Aufgaben "schwieriger" sind; dabei sollte auch die Rolle der LehrerIn nicht vergessen werden, deren Wirken die Aufgaben "leichter" oder "schwieriger" sein läßt. Daten von Kennedy et al (1963) (zitiert nach Thomas (1982)) legen nahe, dass Intelligenzquotienten log-normal und nicht Gauß-verteilt sind. Wie am Anfang des Abschnitts 2.1 gezeigt wurde ("Satz"), kann man praktisch jede Verteilungsfunktion an eine Menge relativer Häufigkeiten anpassen, wobei allerdings nicht notwendig eine Intervallskala entsteht, die den verschiedenen Noten zwischen "2" und "4" gleichgroße Abschnitte auf der latenten Kompetenzdimension entsprechen; gelegentlich wird ja behauptet, dass die Wahl einer Gauß-Verteilung stets auch eine Intervallskala erzeugt (z.B. Jensen (1974)), aber das ist sicherlich nicht der Fall. Die Wahl der log-Normalverteilung hat jedenfalls den Vorteil, auf der Halbachse $(0, \infty)$ statt auf dem Intervall $(-\infty, \infty)$ definiert zu sein. Die Wahl der Gauß-Verteilung kann u.a. die Betrachtung trunkierter Verteilungen[9] bedeuten, weil schon die Ausprägungen der Begabungen in einer Klasse keineswegs Gaußverteilt sein müssen: Begabungen unterhalb eines bestimmten Niveaus werden im Gymnasium mit großer Wahrscheinlichkeit aussortiert.

Die *Item-Response-Theorien*, von denen das Rasch-Modell die einfachste Version ist, legen die logistische Verteilung nahe, weil dann die Schwierigkeits- und die Kompetenzparameter in einem gewissenen Sinne unabhängig voneinander und dem Anspruch nach populationsunabhängig geschätzt werden können. Wieder stellt sich zunächst die Frage nach der "wahren" Verteilung. Nun hatte ja schon McNemar daraufhin gewiesen, dass man Abweichungen von der Gauß-Verteilung leicht durch Hinzufügen oder Wegnehmen bestimmter Aufgaben oder durch ge-

[9]Es wird eine Stichpobe von Personen ausgewählt, die einen IQ X oberhalb eines kritischen Wertes X_0 haben. Wenn die Gauß-Verteilung die Verteilung für die Gesamtpopulation ist, so ist die Verteilung für IQs in Stichproben, die die Personen mit einem IQ $< X_0$ gar nicht aufgenommen werden die *trunkierte* Verteilung für Werte größer als X_0 (trunkieren = abschneiden).

eignete Punktvergabe erzeugen kann. So gesehen, macht es womöglich gar keinen Sinn, von einer 'wahren' Verteilung zu sprechen, was nicht bedeutet, dass Kompetenzen in einer Population nicht zufällig im Sinne der Statistik verteilt sind, dass aber die Verteilung von Kompetenzen eine Mischung von Verteilungen ist, deren Komponenten kaum eindeutig identifiziert werden können. Da es auf die spezifisch objektive Schätzung von Personen- und Itemparametern (= Schwierigkeitsparametern) κ_g und θ_j ankomme, müsse man eben von vornherein die Aufgaben so aussuchen, dass sie die spezifisch objektive Schätzung der κ_g und θ_j ermögliche. Darüber hinaus läßt sich zeigen, dass die Schätzungen $\hat{\kappa}_g$ und $\hat{\theta}_j$ der Parameter sogenannte *suffiziente Schätzungen*[10] sind, d.h. es geht alle Information über die Parameter, die in den Daten vorhanden ist, in die Schätzungen dieser Parameter ein. Daraus ist gefolgert worden, dass IRT-Modelle die einzig sinnvollen Messmodelle überhaupt seien (Fischer (1974)), andere Messmodelle, die diese Eigenschaft nicht haben, seien suboptimal. Vorberg und Schwarz (1990) haben allerdings 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), eine interessante Diskussion von Raktionszeiten im Rahmen von IRT-Modellen findet man bei Ranger (2009). Die Idee, dass man SchülerInnen nach Maßgabe von bis auf eine angeblich bis auf eine Dezimalstelle genaue Schulnote für weiterführende Schulen selegieren kann sollte man wohl aufgeben.

Bisher ist in dieser Diskussion die Rolle der Lehrerin oder des Lehrers fast unberücksichtigt geblieben; oben wurde nur angemerkt, dass die Schwierigkeit der Aufgaben auch von der LehrerIn abhängt. Jeder, der einmal zur Schule gegangen ist, wird die Wichtigkeit dieser Rolle bestätigen, und Sabine Czerny hat gezeigt, dass jenseits aller Betrachtungen über Skalen und Wahrscheinlichkeitsverteilungen LehrerInnen einen großen Anteil am Erfolg der Mehrzahl der SchülerInnen haben können. Wollte man die Wirkungen von Faktoren wie Lehrmethode, Begabung, familiärer Hintergrund und LehrerIn auf die Schulnoten überprüfen, so sollte man die *Wechselwirkung* LehrerIn $\times$ SchülerIn[11] als gesonderte Einflußgröße mit berücksichtigen, statt sich mit einem Hinweis auf "Parallelkollegen" als Standard der Beurteilung von SchülerInnen zu begnügen.

4 Anhang

4.1 Allgemeine Betrachtungen

Mit $\mathbb{R}$ wird im Folgenden die Menge der reellen Zahlen bezeichnet. Für X wird gewöhnlich ein Intervall aus $\mathbb{R}$ angegeben, das den potentiellen Variationsbereich

[10] Ein Begriff aus der mathematischen Statistik

[11] So die Bezeichnungsweise in der Theorie der Versuchspläne.

von X angibt. So kann ohne weitere Einschränkung $X \in \mathbb{R}$ gelten, oder X ist auf den Bereich der positiven reellen Zahlen beschränkt: $X \in R_+ = \{0 \leq 0 \leq \infty\}$. $F(x) = P(X \leq x)$ heißt Verteilungsfunktion, und $f(x) = \frac{dF}{dx}$ ist die zugehörige Dichtefunktion, so dass

$$F(x) = \int_{-\infty}^{x} f(u)du. \tag{4.1}$$

Gilt $X \in \mathbb{R}_+$, so ist $f(x) = 0$ für alle $x < 0$ und $F(x) = \int_0^\infty f(u)du$. Verteilungen sind durch bestimmte Parameter bestimmt. Dabei sind zwei Parameter von besonderer Bedeutung für das Folgende: (i) $\mu = \mathbb{E}(X)$, und $\sigma^2 = \mathbb{V}(X)$. μ ist der Erwartungswert, d.h. der Mittelwert über alle möglichen Werte ("Realisierungen") von X, und σ^2 ist die Varianz von X. σ^2 ist der Mittelwert über alle möglichen Quadrate der Differenzen $(X-\mu)^2$. Da stets nur eine Stichprobe von n X-Werten beobachtet bzw. gemessen werden kann, werden die Parameter durch die entsprechenen arithmetischen Mittel $\bar{x}$ und s^2 geschätzt[12]

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2. \tag{4.2}$$

Will man Noten über eine Wahrscheinlichkeitsverteilung definieren, so bedeutet dies, dass man auf irgendeine Weise die zu beurteilende Kompetenz derart auf Zahlen abbildet, dass die Relation zwischen den Zahlen die Relation zwischen den korrespondierenden Kompetenzen abbildet. Die Kompetenz einer Person kann aufgrund physischer und psychischer Einflüsse fluktuieren, und da diese Einflüsse kaum jemals alle explizit erfasst werden können betrachtet man diese Fluktuationen als zufällig; "zufällig" bedeutet also nicht, dass das Prinzip von Ursache und Wirkung aufgehoben ist, sondern dass *nicht alle* Einflüsse auf die Kompetenz erfasst werden können. X ist dementsprechend eine zufällige Veränderliche. Die Kompetenz wird durch eine Leistung angezeigt; die Leistung wird oft durch die Vergabe von Punkten repäsentiert. Die Vergabe einer Schulnote bedeutet dann, dass X nach Maßgabe der gezeigten Leistung (Punktwert) einen Wert aus einem bestimmten, für diese Note charakteristischen Intervall aus dem Definitionsbereich von X annimmt. Hat man sich für den Notenbereich von "1" ("sehr gut") bis "5" ("ungenügend") entschieden, so wird man den Definitionsbereich von X in Teilintervalle aufteilen: eine "1" wird vergeben, wenn $X > x_1$, eine "2", wenn $x_2 < X \leq x_1$, etc, und eine Leistung wird mit "5" bewertet, wenn $X \leq x_4$. Die Wahrscheinlichkeit für die Vergabe der j-ten Note N_j ist dann durch

$$P(N_j) = P(x_{j-1} \leq X \leq x_j) = F(x_{j-1}) - F(x_j), \quad j = 1, \ldots, 5 \tag{4.3}$$

gegeben. Die konkrete Bestimmung der x_i-Werte hängt von der Wahl der Wahrscheinlichkeitsverteilung ab, die Details werden in den folgenden Abschnitten gezeigt.

[12] Als Schätzung für σ^2 wird oft $\hat{s}^2 = \frac{1}{n-1}\sum_i (X_i - \bar{x})^2$ gewählt, weil die Schätzung s^2 mit einer systematischen Tendenz, kleiner als σ^2 zu sein, ("'Bias") behaftet ist. Diese Tendenz wird vermieden, wenn durch $n-1$ dividiert wird.

X läßt sich *standardisieren*: man definiert die neue zufällige Veränderliche

$$Z = \frac{X - \mu}{\sigma} \tag{4.4}$$

Die Wert von μ und σ sind nicht bekannt und werden für praktische Berechnungen durch den Stichprobenmittelwert $\bar{x}$ und die Sticbprobenstreuung $\hat{s}$ ersetzt, man erhält die Schätzung $\hat{Z} = (X - \bar{x})/\hat{s}$. Es läßt sich zeigen, dass der Mittelwert standardisierter Werte Z_i oder $\hat{Z}_i$ stets gleich Null ist, und die Standardabweichung ist gleich 1:

$$\bar{\hat{z}} = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i - \bar{x}}{\hat{s}} = \frac{1}{\hat{s}}\frac{1}{n}\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\bar{x}\right) = \frac{1}{\hat{s}}\frac{1}{n}(n\bar{x} - n\bar{x}) = 0, \tag{4.5}$$

$$\begin{aligned} \hat{s}_z^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(\hat{Z}_i - \bar{\hat{z}})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{Z}_i - \bar{\hat{z}})^2 \\ &= \frac{1}{\hat{s}^2}\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2 = \frac{\hat{s}^2}{\hat{s}^2} = 1. \end{aligned} \tag{4.6}$$

Die Verteilungsfunktionen für X und Z lassen sich jeweils durcheinander ausdrücken:

$$P(Z \le z) = P\left(\frac{X - \bar{x}}{\hat{s}} \le z\right) = P(X \le z\hat{s} + \bar{x}) \tag{4.7}$$

$$P(X \le x) = P(\hat{s}Z + \bar{x} \le x = P\left(Z \le \frac{x - \bar{x}}{\hat{s}}\right). \tag{4.8}$$

Zur Abkürzung werde $p_i = P(X \le x_i) = P(Z \le z_i)$ gesetzt. Für den Spezialfall, dass X Gauß-verteilt ist, wird $\Phi(z_i) = P(Z \le z_i)$ geschrieben. Die Beziehung zwischen x und $P(X \le x)$ bzw. z und $PZ \le z)$ ist i.A. umkehrbar: ist $p_x = P(X \le x)$, so ist $P^{-1}(p_x) = x$ und analog dazu gilt für $p_z = P(Z \le z)$ die inverse Relation $P^{-1}(p_z) = z$.

Die Dichtefunktion f und die Verteilungsfunktion $F(x) = P(X \le x)$ der Gauß-Verteilung sind durch

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad F(x) = \int_{-\infty}^{x} f(u)du. \quad -\infty < X < \infty \tag{4.9}$$

gegeben. Dabei ist $\exp(x)$ eine Schreibweise für e^x. μ und σ sind die Parameter der Verteilung; μ ist der Mittelwert *über alle* möglichen Werte von X. μ heißt auch der Erwartungswert von X, man schreibt $\mathbb{E}(X)$ dafür. μ ist nur eine andere Bezeichnung für $\mathbb{E}(X)$.

Für die log-Normalverteilung gilt die Definition: Es sei Y eine mit dem Erwartungswert μ und der Varianz σ^2 Gauß-verteilte zufällige Veränderliche. Weiter sei

$X = e^Y$ eine auf $0 \leq X < \infty$ definierte zufällige Veränderliche. Dann heiüt $\log X$ *log-normalverteilt*. Die Dichtefunktion von $X = x$ ist durch

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(\log x - \mu)^2}{\sigma^2}\right]. \tag{4.10}$$

Zum Nachweis betrachte man

$$P(X \leq x) = P(E^Y \leq x) = P(Y \leq \log x);$$

dann ist

$$f_X(x) = \frac{d(P(X \leq x)}{dx} = \frac{dP(Y`\log x}{dx} = f_Y(\log x)\frac{1}{x},$$

(Kettenregel: $d(\log x)/dx = 1/x$), und damit hat man (4.10). Für den Erwartungswert und die Varianz von X erhält man

$$\mathbb{E}(X) = e^{\mu+\sigma^2/2} \tag{4.11}$$

$$\mathbb{V})X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1) \tag{4.12}$$

4.2 Methode der Kleinsten Quadrate

μ ist eine Konstante, die gelegentlich auch "wahrer Wert" genannt wird; μ ist der Mittelwert *über alle möglichen* Werte von X, also nicht nur der tatsächlich gemessenen, in einer Stichprobe zusammengefassten X-Werte. Man schreibt $\mathbb{E}(X)$ für derartige Mittelwerte und nennt sie *Erwartungswert*. Man hat dann $\mathbb{E}(X) = \mathbb{E}(\mu + \varepsilon) = \mu + \mathbb{E}(\varepsilon)$, woraus $\mathbb{E}(\varepsilon) = 0$ folgt. σ^2 ist der Erwartungswert der quadrierten Anweichungen $(X - \mu)^2$, d.h. $\sigma^2 = \mathbb{E}(X - \mu)^2$.

μ ist wie ε nicht direkt beobachtbar und muß aus den Messungen X_i geschätzt werden. Gauß verwendete dazu die Methode der Kleinsten Quadrate: μ wird als derjenige Wert bestimmt, für den die Summe

$$Q(\mu) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \tag{4.13}$$

minimal wird. Es zeigt sich, dass das arithmetische Mittel $\bar{x}$ der X_i-Werte die Funktion $Q(\mu)$ minimalisiert, d.h. $\min_\mu Q(\mu) = Q(\bar{x})$. Der Ausdruck

$$Q(\bar{x}) = s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2 \tag{4.14}$$

ist die Stichprobenvarianz und ist eine Schätzung für den Parameter σ^2; als tatsächliche Schätzung wird allerdings die Größe

$$\hat{s}^2 = \frac{1}{n-1}\sum_{i}(X_i - \bar{x})^2 \tag{4.15}$$

verwendet, weil sie anders als s^2 keinen systematischen Fehler enthält.

Literatur

[1] Aitchinson, J. Brown, J.A.C.: The Lognormal distribution. Cambridge 1963

[2] Billingsley, P.: Probability and Measure. New York 1979

[3] Binet, A., Simon, T. (1904) Methodes nouvelles pour le diagnostiqc du niveau intellectuel des anormaux. *L'Année Psychologique,* Vol. 11, 191–244

[4] Bleher, C.: Nicht so viele Einser, bitte! http://www.sueddeutsche.de/karriere/kritik-an-guten-noten-nicht-zu-viele-einser-bitte-1.592366

[5] Dawes, R.M.: House of Cards – Psychology and Psychotherapy built on Myths- New York 1994

[6] Dorfman, D.D. (1978) The Cyril Burt Question: New Findings. *Science*, 201 (4362), 1177 – 1186

[7] Fischer, G.H.: Einführung in die Theorie psychologischer Tests. Bern 1974

[8] Jensen, A.R. (1974) Cumulative deficit: A testable hypothesis? *Developmental Psychilogy,* 39, 1 – 123

[9] Koch, A.L. (1966) The Logarithm in Biology. 1. Mechanisms Generating the Log-Normal Distribution Exactly. *Journal of Theoretical Biology*, 12, 276–219

[10] Limpert,E., Stahel, E.A., Appt, M.: Log-normal distributions across the sciences. *Biosciences*, 2008, 51(5), 341–352

[11] 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

[12] McNemar, Q.: The Revision of the Stanford-Binet Scale. An Anlysis of the Standardization Data. . Houghton Miffin, Boston 1942

[13] Quine, W. v. O. (1951) The two dogmas of empiricism. *Philosophical Review* 60 (1), 20 – 43

[14] Ranger, J.M.: Der Nutzen von Reaktionszeiten bei psychologischen Tests im Rahmen von Item-Response Modellen. Gießen 2009 https://d-nb.info/997129034/34

[15] Rasch, G.: Probabilistic Models for some intelligence and attainment tests. The Danish Institute of Educational Research, Copenhagen, 1960

[16] Rasch, G. (1961) On general laws and the meaning of measurement in psychology. The Danish Institute of Educational Research, Copenhagen.

[17] Steyr, R. Eid, M.: Messen und Testen. Springer-Verlag Berlin Heidelberg New York 1993

[18] Sun, K. (2004) Explanation of Log-Normal Distributions and Power-Law Distributions in Biology and Social Science.
$http://guava.physics.uiuc.edu/\ nigel/courses/569/Essays_2004/files/sun.pdf$

[19] Thomas, H. (1982) IQ, Interval Scales, and Normal Distributions. *Psychological Bulletin*, 91(1), 198–202

[20] Tversky, A., Kahneman, D. (1974) Judgment under Uncertainty: Heuristics and Biases. *Science*, 185, 1124–1131

[21] Vorberg, D., Schwarz, W. (1990) Rasch-representable reaction time distributions. *Psychometrika*, 55, 617 – 632

[22] Thomas, H. (1982) IQ, Interval scales, and normal distributions. *Psychological BBulletin*, 91(1), (1998 - 2020)

[23] Weidemann,B. (2008) Die Angst vor guten Noten. Frankfurter Rundschau, https://www.fr.de/wissen/angst-guten-noten-11602814.html

[24] Yule, G.U., Kendall, M.G.: An Introduction to the Theory of Statistics. Griffin, London 1937