Eine elementarmathematische Begründung des BENFORD-Gesetzes

von Hans Humenberger

1 Einleitung

1881 entdeckte der Astronom und Mathematiker Simon NEWCOMB bei der Arbeit mit Logarithmenbüchern, dass diese auf den Anfangsseiten viel abgegriffener und abgenutzter waren als auf den hinteren. Dies wäre bei anderen Büchern als Logarithmentafeln in Bibliotheken durchaus erklärbar, denn viele Leute beginnen ein Buch zu lesen (Roman, Gedichte, Theaterstück, Kurzgeschichten, Sachbücher, Fachbücher etc.), hören aber vorzeitig damit wieder auf, weil sie keine Zeit mehr haben, weil es ihnen zu langweilig wird, weil es ihnen zu kompliziert wird (Fachbücher) u. ä. Wenn viele die Lektüre unfertig unterbrechen, ist es klar, dass der Anfang von Büchern abgenutzter sein kann als der Schluss. Aber warum soll dies bei Logarithmentafeln der Fall sein – diese werden ja nach anderen Gesichtspunkten benutzt. Die einzige Erklärung, die es dafür gibt, ist, dass der Logarithmus von Zahlen mit niedrigen Anfangsziffern (1, 2, ...) häufiger gesucht wurde als von Zahlen mit hohen Anfangsziffern (9, 8, ...). Aber warum? Kommen Zahlen mit niedrigen Anfangsziffern „in der Welt" häufiger vor? Warum sollte die Natur eine Präferenz für die 1 als Anfangsziffer haben?

NEWCOMB gab auch schon eine mathematische Formel an, die seine Beobachtungen gut beschreiben konnte: Die relative Häufigkeit, mit der die Ziffer d als Anfangsziffer einer Zahl auftritt, ist ca. $\log_{10}\left(\frac{d+1}{d}\right)$. Er gab aber keine Erklärungen dafür, sondern empfand diese Tatsache einfach als interessante Kuriosität, die bald danach auch wieder vergessen wurde.

Es dauerte über 50 Jahre, bis der Physiker Frank BENFORD (1938) dieselbe Entdeckung an Logarithmenbüchern machte. Er war von dieser Tatsache viel mehr fasziniert und sammelte mit Akribie eine Unmenge von Daten aus den verschiedensten Bereichen, um immer wieder festzustellen, dass 1 als führende Ziffer mit einer relativen Häufigkeit von ca. 30 % auftrat, 2 mit ca. 18 % usw. bis 9 mit ca. 5 %. Wenn die Anfangsziffern von Werten tatsächlich[1] mit diesen beobachteten relativen Häufigkeiten bzw. Wahrscheinlichkeiten vorkommen, ist es einleuchtend, dass bei einer Logarithmentafel die Seiten mit führender Ziffer 1 (das sind eben die vorderen) abgenützter sind als die mit führender Ziffer 9 (ca. sechsmal so stark!).

Intuitiv würden die meisten sicher Gleichverteilung erwarten: Warum soll eine bestimmte Ziffer als führende Ziffer bevorzugt sein? Dann müsste die Wahrscheinlichkeit für alle möglichen Anfangsziffern (1, 2, ..., 8, 9) bei ca. $\frac{1}{9} \approx 0{,}1111$ liegen[2].

BENFORD hat z. B. untersucht: Oberflächen von Seen, Halbwertszeiten radioaktiver Substanzen, Energieverbrauchszahlen von Haushalten, Entfernungen zwischen Orten, Baseball-Statistiken etc. Aber auch er hat keine Erklärung dafür geben können; die erste mathematische Erklärung stammt von Roger S. PINKHAM (1961).

Man kann sich heutzutage z. B. mit Google sehr schnell selbst einen Überblick über große Datenmengen verschaffen: Man wählt z. B. eine beliebige 3-stellige Zahl (473) und gibt in Google diese Zahl der Reihe nach mit einer führenden 1, ..., 9 als Suchbegriff ein: 1473, ..., 9473. Z. B. bei 1473 erhält man ca. 30,5 Millionen „Treffer", für 9473 nur mehr ca. 3,5 Millionen Treffer. In relativen Häufigkeiten ergibt sich das Bild von **Abbildung 1**, wobei auch die theoretisch nach dem Benford-Gesetz zu erwartenden Werte zum Vergleich zu sehen sind.

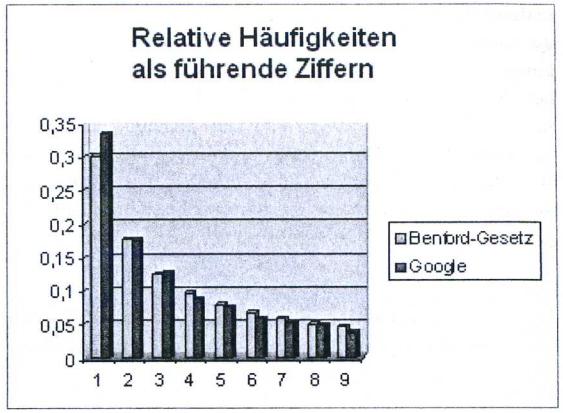

Abb. 1: Ein Versuch mit Google

Es hat natürlich keinen Sinn, Daten zu betrachten, die von vornherein auf einen Bereich eingeschränkt sind, der die Möglichkeiten für die erste Ziffer ziemlich einengt – z. B. Lottozahlen, die 1000-m-Zeiten in Sekunden bei Leichtathletikereignissen, die Anzahl der Buchstaben in den Familiennamen der Bewohner eines Landes, die Gebäudehöhen in einer Stadt, das Alter von Studierenden an einer Universität (das Alter generell!), die Anzahl der Schulbildungsjahre, die Anzahl der Sitze in Fahrzeugen, die Wurzeln der ersten 1000 natürlichen Zahlen usw. Für eine Realisierung des BENFORD-Gesetzes sollen sich die Daten schon über einige Zehnerpotenzen verteilen.

Wir nehmen $\mathbb{R}^+$ als das potentielle Universum der physikalischen Maßzahlen, aus denen die Daten stammen sollen[3] und wollen im Folgenden dem „BENFORD-Gesetz" auf die Spur kommen[4]:

$$P(Z \in Z_d) := P(1.\text{ Ziffer von } Z = d) = \lg(d+1) - \lg d \qquad d = 1, \ldots, 9,$$

wobei Z_d die Menge aller positiven reellen Zahlen bezeichnet, die in der Dezimaldarstellung mit Ziffer d beginnen. Da es sich bei der Zahl Z eigentlich um eine Zufallsvariable handelt, bezeichnen wir sie mit einem Großbuchstaben.

Danach hätten die einzelnen Ziffern die in **Tab. 1** angegebenen Wahrscheinlichkeiten (für das Auftreten als 1. Ziffer), die mit den in vielen Datensätzen beobachteten gut übereinstimmen, so auch bei unserem obigen Versuch mit Google (diese Zahlen sind die numerischen Werte der **Abb. 1** in der Rubrik „BENFORD-Gesetz").

1	2	3	4	5	6	7	8	9
0,301	0,176	0,125	0,097	0,079	0,067	0,058	0,051	0,046

Tab. 1: Wahrscheinlichkeiten für die einzelnen Ziffern nach BENFORD

An anderen Stellen (vgl. HUMENBERGER 1996, 1997, 2000) haben wir deutlich gemacht, dass und wie dieses Phänomen im Schulunterricht anschaulich thematisiert werden könn-

te durch *Einschränkung auf natürliche Zahlen*: $P_n(1)$ bzw. $P_n(9)$ seien die *relativen Anteile* der natürlichen Zahlen $\leq n$, die mit 1 bzw. 9 beginnen. Anschauliche und einfache Überlegungen liefern schnell: Nur wenn n von der Form 9 … 9 ist, sind diese beiden Anteile gleich, bei allen anderen n ist immer $P_n(1) > P_n(9)$. Dies liefert schon die erste Einsicht in die Tatsache, dass die 1 bei der 1. Ziffer von Zahlen gegenüber der 9 doch bevorzugt ist. Es finden sich dort (1996, 2000) auch schon Überlegungen für den Fall *positiver reeller Messwerte*, aber diese enthalten einiges an *Maßtheorie*, was wir hier aus Gründen der Elementarität vermeiden wollen. Sie würde einer möglichen Umsetzung im Schulunterricht deutlich entgegenstehen.

2 Wahrscheinlichkeiten bei unbeschränkten Mengen und eine zunächst vordergründige Argumentation

Beispiel:
Die Lüftung eines Tunnels wird automatisch in Betrieb gesetzt und wieder ausgeschaltet; und zwar ist sie von jeder vollen Stunde an 20 Minuten lang in Betrieb (d.h. sie wird z.B. um 10.00 Uhr eingeschaltet und um 10.20 Uhr wieder ausgeschaltet). Wie groß ist die Wahrscheinlichkeit, dass ein zu einem *zufälligen* Zeitpunkt[5] in den Tunnel einfahrendes Auto die Lüftung in Betrieb vorfindet?

Die „Lösung" scheint hier ziemlich klar zu sein: Innerhalb jeder vollen Stunde spielt sich dasselbe Szenario ab (20 von 60 Minuten Betrieb); wegen dieser offensichtlichen Periodizität wird man auch intuitiv den möglichen großen Stichprobenraum $\mathbb{R}$ einschränken auf $[0; 1)$ (in Stunden) und dort die Wahrscheinlichkeit mit $\frac{20}{60} = \frac{1}{3}$ ausrechnen. Dabei ist ebenfalls intuitiv klar, dass es keine Rolle spielt, ob die Lüftung jeweils zu den vollen Stunden oder jeweils zu irgendeinem anderen fixen Zeitpunkt 20 Minuten lang pro Stunde eingeschaltet wird (z.B. jeweils um *.10 Uhr).

Dies ist ein Beispiel, bei dem einer Teilmenge (Vereinigung unendlich vieler Intervalle) einer *unbeschränkten* Menge (nämlich $\mathbb{R}^+$) ein relatives Maß („Wahrscheinlichkeit") zugeordnet wurde.

Welchen relativen Anteil $P(A)$ hat eine gewisse Teilmenge A an einer Gesamtmenge Ω? Bei beschränkten Mengen ist die Antwort darauf kein Problem, z.B. macht das Intervall $A = [1; 2)$ den Bruchteil $\frac{1}{9}$ von $\Omega = [1; 10)$ aus.

Bei unbeschränktem Ω ist dies aber i.A. gar nicht mehr so leicht, es bedarf oft relativ komplizierter und langwieriger Überlegungen („Maßtheorie"). Wie groß sind „Anteile" bei unbeschränkten Mengen? Klar ist aber auch hier, dass das Wahrscheinlichkeitsmaß jeder *beschränkten* Menge den Wert 0 zuordnen muss, denn „$\frac{\text{endlich}}{\infty} = 0$".

Im Folgenden wird zunächst obiges Beispiel leicht verallgemeinert und formal aufgeschrieben: Die unbeschränkte Teilmenge von $\mathbb{R}$

$$K_{a,b} := \bigcup_{n=-\infty}^{\infty} [n + a, n + b) \quad \text{mit} \quad 0 \leq a \leq b \leq 1$$

ist eine Vereinigung unendlich vieler halboffener Intervalle und ist in **Abb. 2** dargestellt.

Hier ist auch rein intuitiv klar (vgl. obiges Beispiel): Wenn dieser Menge $K_{a,b}$ ein relatives Maß $P(K_{a,b})$ zukommen soll, dann

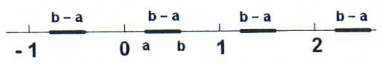

Abb. 2: Die Menge $K_{a,b}$ mit $P(K_{a,b}) = b - a$

kann dies nur $b - a$ sein, denn in jedem Intervall $[n; n + 1)$ spiegelt sich das Verhältnis von $K_{a,b}$ zu $\mathbb{R}$ wider und dort ist es mit $(b - a) : 1$ abzulesen. Die Wahrscheinlichkeit, dass eine „zufällig" gewählte reelle Zahl Z in $K_{a,b}$ liegt, ist daher mit $b - a$ zu quantifizieren:

$$P(Z \in K_{a,b}) = b - a,$$

wobei hier aber wieder – wie beim Tunnel-Beispiel – *Gleichverteilung* vorausgesetzt werden muss.

Wir erinnern uns nun an die Mengen Z_d: die Menge aller *positiven* reellen Zahlen, die Anfangsziffer d haben:

$$Z_d = \bigcup_{n=-\infty}^{\infty} [d\,10^n,\ (d+1)\,10^n) \quad d = 1, \ldots, 9.$$

Unser Ziel ist $P(Z \in Z_d) = \lg(d + 1) - \lg d$ plausibel zu machen („BENFORD-Gesetz").

Für $Z \in \mathbb{R}^+$ ist $\lg Z \in \mathbb{R}$ und für $\lg Z_d$ (d.h. die Menge aller $\lg Z$ mit $Z \in Z_d$) erhalten wir eine Menge, deren Elemente sich über ganz $\mathbb{R}$ erstrecken:

$$\lg Z_d = \bigcup_{n=-\infty}^{\infty} [n + \lg d,\ n + \lg(d + 1)).$$

D.h. die Menge $\lg Z_d$ ist nichts anderes als $K_{\lg d,\ \lg(d+1)}$ in obiger Notation ($a = \lg d$, $b = \lg(d + 1)$). Für die Wahrscheinlichkeit, dass eine zufällig gezogene Zahl aus $\lg Z_d$ stammt, ergibt sich daher $\lg(d + 1) - \lg d$, wenn man Gleichverteilung von $\lg Z$ voraussetzt.

Man könnte nun etwas vordergründig bzw. voreilig argumentieren:
Wegen $\lg Z \in \lg Z_d \Leftrightarrow Z \in Z_d$ erhalten wir „klarer Weise"

$$P(Z \in Z_d) = P(\lg Z \in \lg Z_d) \stackrel{?}{=} \lg(d + 1) - \lg d$$

und dadurch das gewünschte Resultat

$$\boxed{P(Z \in Z_d) = \lg(d + 1) - \lg d}.$$

Über dem letzten Gleichheitszeichen steht ein Fragezeichen. Die zugehörige Frage kann man so ausdrücken:

? Warum ist $\lg Z$ *gleichverteilt*?
Bei dem obigen einfachen Argument für $P(Z \in K_{a,b}) = b - a$ war ja Gleichverteilung die Voraussetzung – „relativer Anteil von Mengen". Anders formuliert: Warum darf (bzw. muss man sogar) bei der Frage nach der Wahrscheinlichkeit in *Logarithmen denken?*

Um die Klärung genau dieser Frage soll es im Folgenden noch gehen – siehe insbesondere den nächsten Abschnitt.

Man darf ja nicht automatisch davon ausgehen, dass Zufallsgrößen gleichverteilt sind. Insbesondere das Anwenden einer Funktion f auf eine Größe Z verändert i.A. das zugehörige Verteilungsgesetz.

Dazu ein einfaches Beispiel mit der Quadratfunktion: Es soll eine Zahl Z zufällig aus $[0; 1)$ gezogen werden. Das interessierende Intervall sei dabei $I := \left[0, \frac{1}{2}\right)$. Mit günstigen und möglichen Intervalllängen argumentiert („geometrische Wahrscheinlichkeit") ergibt sich

$$P\left(Z \in I := \left[0, \tfrac{1}{2}\right)\right) = \tfrac{1}{2}.$$

Nun betrachten wir die Quadratfunktion f: $Z \mapsto Z^2$. Wegen $f([0, 1)) = [0, 1)$ und $f(I) = \left[0, \frac{1}{4}\right)$ ergibt sich „analog" $P(f(Z) \in f(I)) = \frac{1}{4} \neq \frac{1}{2}$. Wenn man also in beiden Fällen – vor und nach dem Quadrieren – ohne weiter darüber nachzudenken Gleichverteilung voraussetzt, so kommt man dabei in Schwierigkeiten, denn wegen $Z \in I \Leftrightarrow f(Z) \in f(I)$ muss natürlich $P(\in I) = P(f(Z) \in f(I))$ sein.

Oft werden LAPLACE- oder „geometrische" Wahrscheinlichkeiten in sehr naiver Weise benutzt: P = |günstig|/|möglich|, wobei $|\cdot|$ für eine *Anzahl* im diskreten Fall bzw. für *Längen*, *Flächen*, *Volumina* im „geometrischen" Fall steht. „Sehr naiv" soll dabei bedeuten, dass man sich oft zu wenig Gedanken macht, ob wirklich kein Ausgang des Zufallsexperiments bevorzugt ist, d.h. ob wirklich Gleichverteilung vorliegt (widrigenfalls wäre ja P = |günstig|/|möglich|falsch). Das Anwenden einer Funktion f (oben: Quadratfunktion) ist eine Transformation einer Zufallsvariable, und dabei darf man *nicht* davon ausgehen, dass das zugehörige Verteilungsgesetz gleich bleibt[6].

Wir brauchen uns zwar über das Verteilungsgesetz von Z (d.h. vor dem Logarithmieren) gar keine Gedanken zu machen, aber wir müssen die Frage beantworten: *Warum* ist $\lg Z$ *gleichverteilt*? Denn das obige einfache Argument des relativen Anteils von Mengen setzte ja *Gleichverteilung* voraus.

3 Skaleninvarianz und die Gleichverteilung der logarithmierten Werte

Wenn es überhaupt ein Verteilungsgesetz für die erste Ziffer von Zahlen gibt[7], so muss dieses doch ein *universelles* sein, d.h. es kann doch nichts ausmachen, in welchen Einheiten man die entsprechenden Größen angibt, da Einheiten ja nicht vom Universum oder einer höheren Macht vorgegeben, sondern willkürliches Menschenwerk sind.

Die Einheiten für eine feste physikalische Größe unterscheiden sich i.A. nur um einen konstanten Faktor $s \in \mathbb{R}^+$, z.B. unterscheiden sich km und Meilen ungefähr um den Faktor $s = 1,609344$. Wenn Entfernungen in km statt Meilen angegeben werden sollen, so müssen die entsprechenden Werte mit $s = 1,609344$ multipliziert werden. Wenn Preise von Dollar in € umgerechnet werden, so muss man die Zahlen durch ca. 1,33 dividieren[8].

D.h. ein Verteilungsgesetz für die erste Ziffer von Zahlen soll sich natürlich nicht ändern, wenn jede Zahl mit einem konstanten Faktor multipliziert wird. Mit anderen Worten: Wenn es ein vernünftiges Verteilungsgesetz für die erste Ziffer von Zahlen gibt, so muss dieses **skaleninvariant** sein, d.h. sich nicht ändern, wenn alle Werte mit einer positiven Konstante multipliziert werden.

Welche Verteilungsgesetze für die erste Ziffer kommen nun dafür in Frage?
Zunächst ein Test, ob die Gleichverteilung als skaleninvariantes Verteilungsgesetz in Frage kommt:
Dazu nehmen wir mal an, dass alle Ziffern 1, …, 9 gleichwahrscheinlich als führende Ziffer wären[9], und betrachten als Beispiel eine Vielzahl von Geldwerten in Euro. Bei einer Währungsänderung, z.B. wenn man statt der €-Werte in die alte DM-Welt zurückfallen möchte, muss jeder Geldwert mit (dem gerundeten Wert) 2 multipliziert werden. Kann dabei die ursprünglich in der €-Welt angenommene Gleichverteilung erhalten bleiben?

Mit Sicherheit nicht:
Alle Euro-Werte mit führender Ziffer 5, 6, 7, 8, 9 haben als DM-Wert die führende Ziffer 1, d.h. nach der Multiplikation mit 2 wäre $P(1) = \frac{5}{9}$.

Alleine damit ist schon klar, dass hier 1, …, 9 als führende Ziffern nicht mehr gleich wahrscheinlich sein können, die Anfangsziffer 1 wäre deutlich bevorzugt! D.h. die **Gleichverteilung** als Verteilungsgesetz zwischen der Ziffern 1, …, 9 ist sicher **nicht skaleninvariant**!

Begründung, warum die Skaleninvarianz der Messwerte zu gleichverteilten Logarithmen führt

Wir interessieren uns für die erste Ziffer von positiven reellen Messwerten Z (wobei wir führende Nullen nicht zählen). Es bietet sich also die so genannte „wissenschaftliche" Schreibweise von Zahlen an („Gleitkommazahl"): $Z = M \cdot 10^n$, wobei $1 \leq M < 10$ ist[10] („Mantisse"). So kann man alle positiven Zahlen darstellen. Diese Schreibweise hat den Vorteil, dass die interessierende Ziffer einfach die 1. Ziffer von M ist, denn M hat sicher keine führenden Nullen. Indem wir statt Z nur mehr die zugehörige Mantisse M betrachten, befreien wir uns sozusagen von den – hier nur lästigen – Zehnerpotenzen, die für das Problem der Anfangsziffer ja irrelevant sind.

Multiplikationen mit $s \in \mathbb{R}^+$ bewirken in der Welt der Zahlen $Z \in \mathbb{R}^+$ (bis auf 10er-Potenzen) genau dasselbe wie in der Welt der Mantissen $M \in [1, 10)$. Damit ist sehr einleuchtend[11]: Wenn die Verteilungsgesetze von Z und $Z \cdot s$ gleich sind („Skaleninvarianz"), dann sind es auch jene von M und $M \cdot s$[12].

Wenn das Verteilungsgesetz für M skaleninvariant ist (d.h. die Verteilungsgesetze für M und $M \cdot s$ sind gleich), dann müssen auch die Verteilungsgesetze von $\lg M$ und $\lg(M \cdot s)$ gleich sein. Wegen $\lg(M \cdot s) = \underbrace{\lg M}_{=:Y} + \underbrace{\lg s}_{=:c}$ bedeutet dies, dass das Verteilungsgesetz

unverändert bleiben muss, wenn man eine beliebige Konstante c *addiert*: die Größen Y und $Y + c$ haben für alle $c \in \mathbb{R}$ dasselbe Verteilungsgesetz[13].

Es ist aber bedeutend leichter die Frage zu beantworten, welche Verteilungsgesetze durch beliebige *Additionen* unverändert bleiben, als die ursprüngliche Frage, welche Verteilungsgesetze durch beliebige Multiplikationen unverändert bleiben. Verteilungsgesetze können durch Dichtefunktionen beschrieben werden, wobei sich zugehörige Wahrscheinlichkeiten als Flächeninhalte unter diesen Dichtefunktionen berechnen lassen.

Nun ist sehr *plausibel*, dass nur die konstante Funktion als Dichtefunktion eines Verteilungsgesetzes in Frage kommt, das durch beliebige Additionen (d.h. horizontale Verschiebungen) nicht verändert wird – siehe **Abb. 3a**. Wie sonst sollte man jemals erreichen, dass sich die Dichtefunktion des zugrunde liegenden Verteilungsgesetzes durch beliebiges horizontales Verschieben nicht ändert[14]? Der konstante Funktionswert der Dichte muss 1 sein, da der mögliche Bereich für $Y = \lg M$ die Länge 1 hat und der Gesamtflächeninhalt unter der Dichtefunktion den Wert 1 haben muss (= „gesamte Wahrscheinlichkeitsmasse").

Die Sache ist jetzt schon deutlich einfacher, denn mit dieser konstanten 1-Funktion als Dichte („Gleichverteilung") lassen sich Wahrscheinlichkeiten der Art $P(a \leq Y < b)$ besonders leicht ausrechnen (Flächeninhalt unter der Dichtefunktion zwischen a und b):

$$P(a \leq Y < b) = (b - a) \cdot 1 = b - a; \text{ siehe } \textbf{Abb. 3b}.$$

Damit können wir die gewünschten Wahrscheinlichkeiten bestimmen, für alle Ziffern $d = 1, \ldots, 9$ ergibt sich unmittelbar:

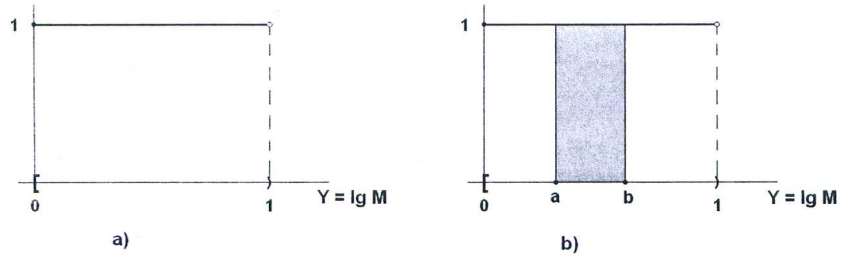

Abb. 3: Dichte der Gleichverteilung von $Y = \lg M$ auf [0; 1)

$$\boxed{\text{P(1. Ziffer von } Z = d)} = \text{P(1. Ziffer von } M = d) = \text{P}(d \le M < d+1)$$
$$= \text{P}(\lg d \le \underbrace{\lg M}_{Y} < \lg(d+1))$$
$$= \boxed{\lg(d+1) - \lg d} \text{ – das BENFORD-Gesetz!}$$

Die Gleichverteilung der logarithmierten Werte ist hiermit also bewiesen mittels der Skaleninvarianz der Messwerte. Die obige, zunächst nur vordergründige Argumentation ist also exaktifiziert und gerechtfertigt! Das oben noch ausstehende „Warum sind die logarithmierten Werte gleichverteilt?" kann also beantwortet werden mit: Wegen der Skaleninvarianz der Messwerte! Diese ist einfach eine vernünftig scheinende Forderung und braucht nicht weiter bewiesen zu werden.

In **Abb. 4** wird beides gleichzeitig dargestellt: die Gleichverteilung von $Y := \lg M$ auf [0; 1) und die daraus (von unten nach oben!) resultierende „logarithmische" Verteilung von M auf [1; 10). Darin sind auch schön die immer kleiner werdenden Anteile der Ziffern $d = 1, \ldots, 9$ zu sehen: $\text{P(1. Ziffer von } Z = d) = \lg(d+1) - \lg d$.

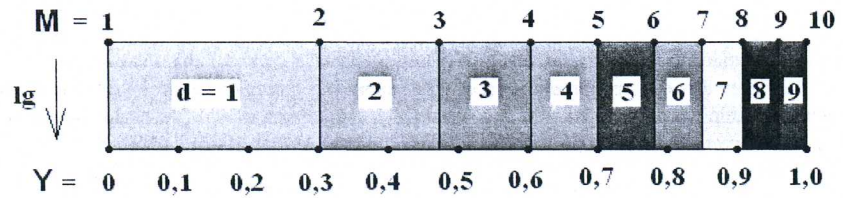

Abb. 4: Gleichverteilung von $Y := \lg M$ auf [0; 1) bzw. die Verteilung von M auf [1; 10) – „logarithmisch"

Arithmetisches versus geometrisches Zählen

Wir Menschen zählen bekanntlich in arithmetischer Folge $0 \overset{+1}{\to} 1 \overset{+1}{\to} 2 \overset{+1}{\to} 3 \ldots$ mit konstanten Differenzen (konstantes absolutes Wachstum, additives Zählprinzip). Wir drücken Verschiedenheiten aber oft auch durch Quotienten (Verhältnisse) aus, wobei hier nicht das additive (arithmetische), sondern das „multiplikative (geometrische) Zählprinzip" im Vordergrund steht.

Es gibt viele Phänomene (insbesondere Wachstumsphänomene, auch beim Tasten, Hören und Sehen, d. h. generell beim Empfinden[15]), bei denen auch die Natur quasi geometrisch zählt, d. h. von Schritt zu Schritt immer mit einer Konstanten multipliziert:

$$a \cdot q^0 \overset{\cdot q}{\to} a \cdot q^1 \overset{\cdot q}{\to} a \cdot q^2 \overset{\cdot q}{\to} a \cdot q^3 \ldots$$

Dies entspricht einem konstanten relativen Wachstum. Solche Werte haben dann die Eigenschaft, dass sich die logarithmierten Werte um eine additive Konstante unterscheiden:

$$\log a \overset{+\log q}{\to} \log a + \log q \overset{+\log q}{\to} \log a + 2\log q \overset{+\log q}{\to} \log a + 3\log q \ldots$$

Wenn diese Werte (nach dem Logarithmieren) gleichverteilt sind, was man aus der Forderung nach Skaleninvarianz bei den ursprünglichen Werten folgern kann, schlägt in diesen Situationen „naturgemäß" das BENFORD-Gesetz voll zu – siehe oben.

STEWART (1994, S. 20) schreibt dazu: „Wir Menschen zählen in arithmetischer Folge 1, 2, 3, . . . und wundern uns, ungleiche Wahrscheinlichkeiten für die Anfangsziffern zu finden. Aber das lässt sich dadurch erklären, dass die Natur mit gleichen Wahrscheinlichkeiten unter den Termen einer geometrischen Folge wählt $x, x^2, x^3, \ldots$". Diese Aussage erklärt aber noch nicht, warum daraus das Benford-Gesetz folgt. Die vernünftig scheinende *Forderung nach Skaleninvarianz* ist eine mögliche Erklärung dafür.

So wie oben die Werte erst nach dem Logarithmieren *gleichverteilt* waren (man könnte von einer „Log-Gleichverteilung" sprechen), so trifft man statt der gewöhnlichen Normalverteilung oft auch auf die so genannte Log-Normalverteilung (d. h. die entsprechenden Werte sind erst nach Logarithmieren *normalverteilt*): z. B. Durchmesser von Bäumen, Durchmesser von Bakterien, Partikelgrößen in Eis oder Wasserwolken, etc.

4 Anwendungen

Das BENFORD-Gesetz ist sicher ein interessantes und überraschendes Resultat. Aber hat es auch reale Anwendungen? Kann man dieses Wissen irgendwo mit Nutzen einsetzen? Wenn jemand allzustark an das BENFORD-Gesetz glaubt, könnte er ja meinen, dass auch beim Lottospielen Zahlen mit Anfangsziffer 1 bevorzugt seien. Aber das ist sicher nicht der Fall, dort herrscht einfach jedes Mal aufs Neue der neutrale Zufall, d. h. das BENFORD-Gesetz hilft nicht, um bessere Tipps beim Lotto zu erhalten, leider!

Der amerikanische Mathematiker MARK NIGRINI hat dieses Gesetz erst Anfang der 90-er Jahre wirklich berühmt gemacht, indem er Anwendungen in die Tat umgesetzt hat. Wenn z. B. Steuerpflichtige (große Betriebe mit wirklich vielen Daten) ihre Steuererklärung beim Finanzamt einreichen, so sind die Daten in manchen Fällen ja etwas manipuliert: Gewisse

Daten wurden vielleicht verändert, gewisse erfunden, gewisse gestrichen etc. In vielen Fällen tendieren die Manipulateure dazu, bei ihren erfundenen Zahlen nicht zu kleine aber auch nicht zu große Anfangsziffern zu wählen, also z. B. sehr viele mit 4, 5, 6 beginnen zu lassen, oder aus falscher Intuition heraus die Anfangsziffern 1, ..., 9 relativ gleichmäßig zu benutzen. Dies führt dazu, dass die 1 (oder auch die 2) als Anfangsziffer nicht mehr genügend häufig vertreten ist.

MARK NIGRINI hat eine Software entwickelt, die überprüft, ob irgendwelche übermittelten Daten dem BENFORD-Gesetz gehorchen. Diese Software wird schon vielfach eingesetzt in Amerika, in Deutschland und in der Schweiz. Wenn ein Datensatz das Benford-Gesetz zu schlecht erfüllt, so ist dies natürlich kein Beweis, dass die Daten gefälscht sind, aber es können die Alarmglocken läuten, und eine genauere Untersuchung (Steuerprüfung) kann veranlasst werden. Auch die Steuererklärungen von Bill Clinton und Bill Gates wurden angeblich mit NIGRINIS Programm überprüft, es ergaben sich dabei aber keine Anzeichen von Steuerbetrug (siehe WALTHOE).

Bei der Entdeckung so mancher berühmter gefälschter Bilanzen, z. B. bei den Riesenskandalen (2002) um die Bilanzfälschungen von *Enron* und *Worldcom*, bei denen unzählige Anleger um ihr Kapital betrogen wurden, konnte angeblich eine BENFORD-Überprüfung helfen (Wikipedia-Artikel zum BENFORD-Gesetz).

Es ist gar nicht so leicht, Daten „passend" zu manipulieren, denn es gibt nicht nur ein Verteilungsgesetz für die 1. Ziffer von Zahlen, sondern auch welche für die nachfolgenden Ziffern, aber da sind die Unterschiede zwischen den einzelnen Ziffern 1, ..., 9 nicht mehr ganz so groß wie bei der 1. Ziffer. Die Ziffern folgen umso näher einer Gleichverteilung, je kleiner ihr Stellenwert ist. Wir geben die Werte hier nur ohne Begründung an (die zugehörige Formel zur Berechnung ist z. B. im Wikipedia-Artikel über das BENFORD-Gesetz zu lesen; vgl. **Tab. 2**).

Ziffer	0	1	2	3	4	5	6	7	8	9
1. Ziffer		0,301	0,176	0,125	0,097	0,079	0,067	0,058	0,051	0,046
2. Ziffer	0,120	0,114	0,109	0,104	0,100	0,097	0,093	0,090	0,088	0,085
3. Ziffer	0,1018	0,1014	0,1010	0,1006	0,1002	0,0998	0,0994	0,0990	0,0986	0,0983

Tab. 2: Wahrscheinlichkeiten für die Ziffern

Und außerdem muss ein professioneller Fälscher noch einer Reihe anderer stochastischer Gesetzmäßigkeiten Rechnung tragen. Trimmt der Datenfälscher die Daten allzu genau auf die theoretische Erwartung hin, besteht Gefahr, dass die Manipulationen eben daran erkannt werden!

Manche Daten passen aber auch ungefälscht nicht zum BENFORD-Gesetz, eine Verletzung des BENFORD-Gesetzes ist eben nie ein Beweis, dass Daten gefälscht sind, sondern nur ein Hinweis darauf, dass sie gefälscht sein könnten. Man denke z. B. auch an Preise (von denen ja auch viele in Bilanzen und Steuererklärungen vorkommen); hier sind oft aus psychologischen Gründen Werte knapp unterhalb von Zehnerpotenzen deutlich häufiger anzutreffen (9,90 oder 99,90 etc.), so dass die 9 als führende Ziffer auch in ungefälschten Verkaufsbilanzen sicher häufiger vorkommen wird, als ihr laut BENFORD-Gesetz zusteht.

Es gibt noch so manche andere Anwendung des BENFORD-Gesetzes. In Belgien arbeitet man daran, Ungereimtheiten bei den Krankenhausabrechnungen auf die Spur zu kommen, in der Wissenschaft kann das BENFORD-Gesetz helfen zu erkennen, ob ein „übereifriger" Wissenschaftler (z. B. aus dem Drang heraus, durch ein *signifikantes* Ergebnis mehr Aufmerksamkeit zu erregen) seinen Daten zur Signifikanz etwas nachgeholfen hat. In Freiburg wird angeblich (vgl. WALTHOE u. a.) auch daran gearbeitet, die Verwaltung von Speicherplatz auf Festplatten mit Hilfe des BENFORD-Gesetzes zu optimieren[16]. Die Anwendungen des BENFORD-Gesetzes scheinen also immer weitere Kreise zu ziehen.

Bemerkungen:

- Am Ergebnis und an der prinzipiellen Vorgangsweise ändert sich nichts, wenn man realistischererweise nicht $\mathbb{R}^+$, sondern $\mathbb{Q}^+$ (oder gar nur die *endlichen* Dezimalzahlen) als das mögliche Universum aller physikalischen Konstanten ansieht (bei allen in Dezimalschreibweise angegebenen Werten aller möglichen Tabellen können ja nur endlich viele Stellen berücksichtigt werden).
- Das BENFORD-Gesetz hat mit der Darstellung im *Dezimalsystem* **nichts** zu tun. Auch bei Darstellungen mit jeder anderen natürlichen Zahl $a > 2$ als Basis ergäbe sich ein analoges Gesetz für die Wahrscheinlichkeit, dass eine Zahl mit einer Ziffer d beginnt:
  $P(Z \in Z_d) = \log_a(d+1) - \log_a(d)$ für $d = 1, 2, 3, \ldots, a-1$.

Zusammenfassung

Das Anliegen dieses Aufsatzes ist zu zeigen, wie eine elementarmathematische Begründung des in letzter Zeit sehr populären BENFORD-Gesetzes (siehe die populärwissenschaftlichen Literaturzitate von DWORSCHAK 1998, MATTHEWS 1999, ALBRECHT 2000) auch auf Schulniveau möglich ist.

Eine „außermathematische Anwendung" des prima vista vielleicht höchst theoretisch scheinenden Gesetzes wurde durch MARK NIGRINI realisiert (NIGRINI 2000), der mittels dieses Gesetzes Steuersündern auf die Spur gekommen ist. Internationale Konzerne und Finanzbehörden interessieren sich mittlerweile für die Software von M. NIGRINI.

Anmerkungen

1 D. h. nicht nur in den beobachteten Stichproben, sondern allgemein.
2 Mit *Anfangsziffer* sei im Folgenden stets *die erste Ziffer ungleich 0 bzw. erste „signifikante" Ziffer* gemeint; also z. B. 3 in 0,0367.
3 *Messwerte* sind zwar naturgemäß rational, aber mit reellen Zahlen rechnet es sich leichter. Alles Folgende würde auch mit $\mathbb{Q}^+$ statt $\mathbb{R}^+$ funktionieren. Wenn physikalische Werte in Wirklichkeit negativ wären, kann man z. B. deren Betrag nehmen.
4 In weiterer Folge schreiben wir $\lg := \log_{10}$ („Logarithmus generalis"). Diese Formel gilt auch für $d = 9$.
5 „Zufällig" soll hier heißen, dass die Ankunftszeiten der Autos *gleichverteilt* angenommen werden: kein Zeitpunkt soll bevorzugt werden.
6 Wenn man eine Zufallsvariable einer Transformation f unterzieht, so muss man immer überlegen, wie sich das auf das zugehörige Verteilungsgesetz (Dichte- bzw. Verteilungsfunktion) auswirkt, um danach wieder Überlegungen zu Wahrscheinlichkeiten anstellen zu können. Es geht uns hier nur um das zu Grunde liegende Phänomen, und nicht darum, wie man technisch zu der jeweils neuen Dichtefunktion bei der Transformation f kommt; deswegen ist dies hier auch nicht näher für die log- bzw. für die Quadratfunktion ausgeführt.
7 Empirische Beobachtungen unterstützen diese These!
8 Dieser Wert variiert natürlich von Tag zu Tag.
9 So würde es die ursprüngliche Intuition eigentlich nahe legen: $P(1) = \ldots = P(9) = \frac{1}{9}$
10 So wie Z kann auch M als Zufallsvariable aufgefasst werden, deswegen wieder ein Großbuchstabe.
11 Für genauere Ausführungen dazu bräuchte man Maßtheorie, die hier aber vermieden werden soll.

12 Wenn der Wert von $M \cdot s$ dabei nicht in [1; 10) liegen sollte, so muss man dabei erneut die Mantisse bilden, z. B. hat man also für $M = 9$ und $s = 2$ bei $M \cdot s$ an 1,8 zu denken!

13 Bei $Y + c$ muss man dabei eigentlich „modulo 1" denken, damit $Y + c$ wieder dieselbe Wertemenge [0; 1) wie Y hat.

14 Natürlich wieder modulo 1 gedacht, d. h. jener Teil des Graphen der Dichtefunktion, der beim Verschieben den Bereich [0; 1) auf *einer* Seite verlässt, wandert auf der *anderen* Seite wieder herein – siehe Abb. 3. Man könnte auch einen *formalen Beweis* für diese Tatsache führen, dass hier nur die konstante Funktion in Frage kommt.

15 Vgl. die Lautstärkeeinheiten „Bel" bzw. „Dezibel", bei denen in Logarithmen gedacht werden muss. Auch bei der Tonhöhe werden die Unterschiede als gleich wahrgenommen, wenn die Töne dasselbe Frequenz*verhältnis* haben. Dies alles wird subsumiert unter „Weber-Fechner'sches Grundgesetz".

16 Ich weiß nicht genau, wie.

Literatur

Albrecht, J. (2000): Die Eins von Planet Zob. Die Zeit (40, 28. 09. 2000), 35.

Benford, F. (1938): The law of anomalous numbers. In: Proceedings of the American Philosophical Society 78, 551–572.

Dworschak, M. (1998): Weiter Weg zur Zwei – ein kurioses Gesetz der Wahrscheinlichkeitstheorie kann Finanzbeamten helfen, Steuersünder aufzuspüren. In: Der Spiegel 47/1998, 228–229.

Humenberger, H. (1996): Das Benford-Gesetz über die Verteilung der ersten Ziffer von Zahlen. In: Stochastik in der Schule 16, 3, 2–17. Kurzfassung: Beiträge zum Mathematikunterricht 1997, 251–254.

Humenberger, H. (1997): Eine Ergänzung zum Benford-Gesetz – weitere mögliche schulrelevante Aspekte. In: Stochastik in der Schule 17, 3, 42–48.

Humenberger, H. (2000): Das „BENFORD-Gesetz" – warum ist die Eins als führende Ziffer von Zahlen bevorzugt? (überarbeitete und kombinierte Version von Humenberger 1996 und 1997.) In: Henn, H. W., F. Förster u. J. Meyer (Hrsg., 2000): Materialien für einen realitätsbezogenen Mathematikunterricht, Band 6, 138–150. Schriftenreihe der ISTRON-Gruppe, Franzbecker, Hildesheim.

Matthews R. (1999): The power of one. In: New Scientist 2194 (July 10), 27–30.

Nigrini, M. (2000): Digital Analysis Using Benford´s Law: Tests & Statistics for Auditors. Global Audit Publications, Vancouver.

Pinkham, R. S. (1961): On the distribution of first significant digits. In: Annals of Mathematical Statistics 32, 1223–1230.

Stewart (1994): Mathematische Unterhaltungen. In: Spektrum der Wissenschaft (April 1994), 16–20.

Walthoe, J., R. Hunt u. M. Pearson: Looking out for number one. Internet: http://plus.maths.org/issue9/features/benford/