Das „Benford-Gesetz“ – warum ist die Eins als führende Ziffer von Zahlen bevorzugt?

Hans Humenberger

Zusammenfassung

Der „Original-Aufsatz“ (gleichen Titels) in Band 6 aus dem Jahr 2000 enthält bei den mathematischen Erklärungen viele maßtheoretische Aspekte. Diese sind zwar elementarer Natur, aber eben doch Maßtheorie. Dies soll in diesem Aufsatz vermieden werden, weil Maßtheorie kaum mit Schulunterricht kompatibel ist. Das Anliegen dieses Aufsatzes ist zu zeigen, wie eine elementarmathematische Begründung des populären Benford-Gesetzes 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.

1 Einleitung und historische Aspekte

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 ist 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,

Dieser Beitrag ist ursprünglich erschienen in „Materialien für einen realitätsbezogenen Mathematikunterricht“ (Schriftenreihe der ISTRON-Gruppe), Band 6, Franzbecker 2000, S. 138–150, und wurde für diesen ISTRON-Jubiläumsband aktualisiert.

H. Humenberger ✉
Fakultät für Mathematik, Universität Wien, Österreich

H.-S. Siller, G. Greefrath, W. Blum (Hrsg.), *Neue Materialien für einen realitätsbezogenen Mathematikunterricht 4*,
Realitätsbezüge im Mathematikunterricht, https://doi.org/10.1007/978-3-658-17599-3_12

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 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 diesem Phänomen 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 *Anfangsziffer von Werten* tatsächlich eine Wahrscheinlichkeitsverteilung hat, die ca. diesen relativen Häufigkeiten entspricht, 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. $1/9 \approx 0{,}1111$ liegen[1].

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 angegeben, 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 eine beliebige 3-stellige Zahl (z. B. 473) und gibt in Google diese Zahl der Reihe nach mit einer führenden 1, … , 9 als Suchbegriff ein: 1473, … , 9473. Zum Beispiel bei 1473 erhält man ca. 79,6 Mio „Treffer“, für 9473 nur mehr ca. 13,7 Mio Treffer. In relativen Häufigkeiten ergibt sich Abb. 1, wobei auch die theoretisch nach dem Benford-Gesetz zu erwartenden Werte zum Vergleich eingezeichnet sind.

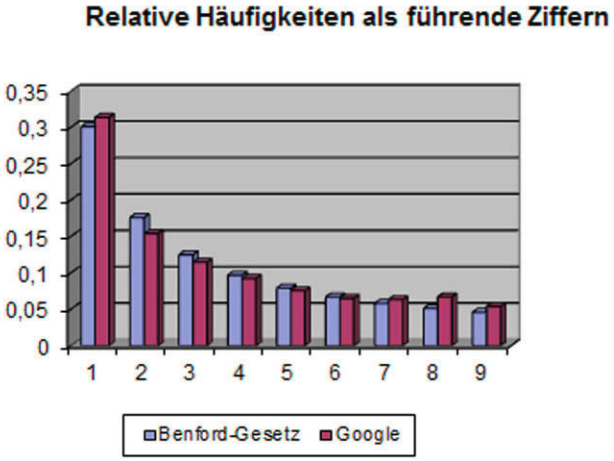

Abb. 1 Ein Versuch mit Google (Februar 2016)

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 Laufzeiten in Sekunden bei 1000 m-Laufbewerben, 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 von Personen, die Anzahl der Sitze in Fahrzeugen, die Wurzeln der ersten 1000 natürlichen Zahlen usw. Eine statistische Analyse vielfältiger Daten zeigt, dass die Verteilung der führenden Ziffer gut mit dem Benford-Gesetz übereinstimmt, wenn sich die Daten wenigstens über einige Zehnerpotenzen verteilen.

[1] Mit *Anfangsziffer* sei im Folgenden stets die *erste Ziffer ungleich 0* bzw. erste „signifikante“ Ziffer gemeint; also z. B. 3 in 0,0367.

Wir nehmen $\mathbb{R}^+$ als das potentielle Universum der physikalischen Maßzahlen, aus denen die Daten stammen sollen[2] und wollen im Folgenden dem „Benford-Gesetz“ auf die Spur kommen[3]:

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

Dabei bezeichnet Z_d die Menge aller positiven reellen Zahlen, die in Dezimaldarstellung mit Ziffer d beginnen. Ziel der Überlegungen ist es, ein stochastisches Modell für das Auftreten der 1. Ziffer herzuleiten. Dabei muss man natürlich die Zahl Z als Realisierung einer *Zufallsvariable* auffassen. Konsequenterweise bezeichnen wir die Zahl Z mit einem Großbuchstaben.

Nach diesem Gesetz hätten die einzelnen Ziffern die in Tab. 1. angegebenen Wahrscheinlichkeiten, die mit den in vielen Datensätzen beobachteten relativen Häufigkeiten gut übereinstimmen, so auch bei unserem obigen Versuch mit Google (diese Zahlen sind die numerischen Werte der Graphik in obiger Abbildung in der Rubrik „Benford-Gesetz“).

Tab. 1 Wahrscheinlichkeiten für die einzelnen Ziffern nach Benford

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

[2] Messwerte sind zwar naturgemäß rational, aber mit reellen Zahlen rechnet es sich hier leichter. Alles Folgende würde aber auch mit $\mathbb{Q}^+$ statt $\mathbb{R}^+$ funktionieren.

[3] In weiterer Folge schreiben wir meist $\lg(x) := \log_{10}(x)$ („Logarithmus generalis“). Diese Formel gilt auch für $d = 9$. Eine Erklärung für Daten, die Schwankungen unterworfen sind, gibt z. B. Dworschak (1998, S. 229): *„Die 1 ist auf der Zahlenskala von der 2 nicht weiter entfernt als die 5 von der 6. Für die wirklichen Dinge allerdings, die gezählt, gemessen oder gewogen werden, kann der Weg von der 1 zur 2 sehr lang sein: Um ihn zurückzulegen, müssen sie auf das Doppelte wachsen. Einer 5 fehlt dagegen nur ein Fünftel, um zur 6 zu werden. [. . .] Angenommen der Deutsche Aktienindex stünde gerade bei 1000 Punkten, dann müssten sich die Aktienkurse im Schnitt verdoppeln, ehe der DAX die 2000 erreicht. Solange bliebe die führende 1 erhalten, solange erschiene sie auf allen Listen. Stünde der DAX aber bei 5000 Punkten, so müssten die Werte nur noch um 20 Prozent steigen, ehe mit 6000 die 5 als erste Ziffer abgelöst wird. Noch kleiner ist im Verhältnis der Schritt von 9000 auf 10 000. [...] Was wächst oder schrumpft, verharrt deshalb relativ lang im Bereich der führenden 1.“*

Wir wollen im Folgenden noch etwas tiefere mathematische, aber trotzdem elementare Erklärungen geben[4], davor aber seien noch einige Anwendungen angegeben.

2 Anwendungen

Das Benford-Gesetz über die Häufigkeit der 1. Ziffer von Zahlen ist ein interessantes und überraschendes Resultat. Aber hat es auch reale Anwendungen? Kann man dieses Wissen irgendwo mit Nutzen einsetzen? Wenn jemand allzu stark an das Benford-Gesetz glaubt, könnte er ja meinen, dass auch beim Lottospielen Zahlen mit Anfangsziffer 1 bevorzugt seien. Aber das ist nicht der Fall: Jede Zahl aus $\{1, \ldots, 49\}$ hat bei den Ziehungen dieselbe Chance, es „herrscht“ einfach jedes Mal aufs Neue der „neutrale Zufall“. Das Benford-Gesetz hilft nicht, um bessere Tipps beim Lotto zu erhalten!

Der amerikanische Mathematiker Mark Nigrini hat dieses Gesetz erst Anfang der

[4] Für tiefere fachmathematische Analysen verweisen wir insbesondere auf Hungerbühler.

neunziger Jahre des 20. Jahrhunderts der Öffentlichkeit bekannt gemacht, indem er Anwendungen dieses Gesetzes 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, einige wurden erfunden, andere gestrichen etc. In vielen Fällen tendieren Manipulateure dazu, bei ihren erfundenen Zahlen die Anfangsziffern 1, …, 9 relativ gleichmäßig zu benutzen, nicht zu kleine aber auch nicht zu große Anfangsziffern zu wählen, also z.B. sehr viele mit 3, 4, 5, 6, 7 beginnen zu lassen. Dies führt dazu, dass die 1 (oder auch die 2) als Anfangsziffer im Vergleich zum Benford-Gesetz zu selten auftritt.

Mark Nigrini hat eine Software entwickelt, die überprüft, ob irgendwelche übermittelten Daten dem Benford-Gesetz „gehorchen“. Diese Software wird schon vielfach eingesetzt in Amerika, 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 Steuererklärung von Bill Clinton wurde 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, war angeblich eine Benford-Überprüfung (Wikipedia-Artikel zum Benford-Gesetz) mit im Spiel.

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 so groß wie bei der 1. Ziffer. Die Ziffern folgen umso besser einer Gleichverteilung, je kleiner ihr Stellenwert ist. Wir geben in Tab. 2 nur die Werte ohne Begründung an (die zugehörige Formel zur Berechnung ist z.B. im Wikipedia-Artikel über das Benford-Gesetz zu lesen):

Tab. 2 Wahrscheinlichkeiten für die Ziffern

Ziffer	1. Ziffer	2. Ziffer	3. Ziffer
0		0,12	0,1018
1	0,301	0,114	0,1014
2	0,176	0,109	0,101
3	0,125	0,104	0,1006
4	0,097	0,1	0,1002
5	0,079	0,097	0,0998
6	0,067	0,093	0,0994
7	0,058	0,09	0,099
8	0,051	0,088	0,0986
9	0,046	0,085	0,0983

Außerdem muss ein professioneller Fälscher noch eine Reihe anderer stochastischer Gesetzmäßigkeiten berücksichtigen (z.B. Häufigkeit von Ziffernpaaren). Trimmt der Datenfälscher die Daten allzu genau auf die theoretische Erwartung aus dem Benford-Gesetz 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, sondern nur ein Hinweis darauf, dass die Daten 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 häufiger vorkommen wird als ihr laut Benford-Gesetz zusteht.

Es gibt noch so manche andere Anwendung des Benford-Gesetzes. 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. Oder wenn in einer Firma die Preisgrenze für selbständige Anschaffungen bei

300 € liegt[5], dann kann es doch sein, dass umtriebige Mitarbeiter/innen dieser Firma guten Kontakt zu anderen Firmen haben, und statt einer Rechnung über 810 € mehrere Rechnungen (z. B. dreimal 270 €) bekommen – dafür kann eben die Bestellung ganz rasch und unbürokratisch erfolgen . . . In solchen Fällen ist bei der 1. Ziffer von Rechnungen sicher besonders häufig die 2 vertreten. Sollte sich das bewahrheiten, können ja genauere Nachforschungen angestellt werden …

Laut B. F. Roukema (Universität Torun, Polen) gibt es starke Hinweise, dass die Präsidentschaftswahlen im Iran 2009 manipuliert wurden (für Mahmud Ahmadinedschad), auch dafür wurde das Benford-Gesetz verwendet. Die Stimmenzahlen der einzelnen Wahlkreise dürften manipuliert worden sein (Spiegel-Artikel: http://www.spiegel.de/wissenschaft/mensch/0,1518,632541,00.html bzw. Aufsatz von Roukema: http://arxiv.org/PS_cache/arxiv/pdf/0906/0906.2789v1.pdf).

Eine weitere Anwendung aus der neueren Zeit ist der griechische Finanzskandal: Vermutlich wurden dort Bilanzen in großem Stil gefälscht, so dass einem Beitritt zur EU und zum Euroraum nichts im Wege stand. Auch hier hätte ein entsprechender Test mit dem Benford-Gesetz schon frühzeitig die Alarmglocken läuten lassen können (vgl. Rauch u. a. 2011). Dieser Test wurde aber erst im Nachhinein gemacht, schade, denn man hätte aus frei zugänglichen Daten den griechischen Finanzbetrug evtl. schon lange erkennen können. Noch einmal: Solche Tests sind natürlich noch lange keine Beweise, dass da etwas manipuliert wurde, aber man hätte Hinweise, dass eine genauere Prüfung der Daten stattfinden sollte (anhand weiterer Belege etc.).

Nun kommen wir zu elementaren mathematischen Erklärungen, die auch im Schulunterricht möglich sind. Eine etwas abgewandelte Zugangsweise ist in Schuppar/Humenberger 2015 zu lesen (Kapitel 1.5, S. 78–92). In Abschn. 3 beschränken wir uns bewusst zunächst auf natürliche Zahlen als „mögliche Zufallszahlen", das ist jener „Topf" aus dem Zahlen zufällig gezogen werden sollen, wobei man sich für die Wahrscheinlichkeit interessiert, dass die gezogene Zahl mit Ziffer d ($d = 1, \ldots, 9$) beginnt. In Abschn. 4 beginnen dann die Überlegungen, wenn $\mathbb{R}^+$ als der Topf der möglichen Zufallszahlen betrachtet wird, die meisten Maßzahlen gehen ja über $\mathbb{N}$ hinaus.

3 Natürliche Zahlen als mögliche Zufallszahlen

Zunächst ist klar, dass die Wahrscheinlichkeiten für die einzelnen Ziffern, als erste Ziffer einer Zufallszahl zu stehen, von der Grundgesamtheit des „Topfes" abhängen, aus dem die Zahl gezogen wird. Wenn z. B. aus den ersten 20 natürlichen Zahlen zufällig gezogen wird, so ist offenbar die 1 als erste Ziffer ziemlich übermächtig (11 von 20 möglichen), 2 steht in zwei von 20 Fällen an erster Stelle und jede andere Ziffer d ($\neq 0$) genau einmal.

Wir erkennen auch sofort, dass bei natürlichen Zahlen die Wahrscheinlichkeit $\frac{1}{9}$ nur dann bei jeder Ziffer $d = 1, 2 \ldots, 9$ auftritt, wenn die Grundgesamtheit aus den ersten 9, 99, 999, 9 999 usw. Zahlen besteht. Verfolgen wir (bei wachsendem n) z. B. einmal die Wahrscheinlichkeit von 1 als erste Ziffer, wenn der „Topf" aus den ersten n natürlichen Zahlen besteht, d. h. wir betrachten die Folge $(P_n(1))_{n\in\mathbb{N}}$: Bei $n = 1$ ist $P_1(1) = 1$, bei $n = 2$ ist $P_1(1) = \frac{1}{2}$, usw., diese Wahrscheinlichkeit sinkt dann bis $P_9(1) = \frac{1}{9}$ bei. Dann steigt die Wahrscheinlichkeit wieder bis $n = 19$ (auf $P_{19}(1) = \frac{11}{19}$), um dann wieder bis $n = 99$ abzufallen (wieder $P_{99}(1) = \frac{1}{9}$), dann kommt natürlich wieder ein Anstieg bis $n = 199$ (auf $P_{199}(1) = \frac{111}{199}$), dem wieder ein Abfall bis $n = 999$ folgt (s. Tab. 3 und zur Veranschaulichung Abb. 2).

So geht dies natürlich „ewig" (von 10er-Potenz zu 10er-Potenz) weiter, die relative Häufigkeit wird sich nie (mit wachsendem n) bei einem stabilen Wert einpendeln, es werden nur die Phasen des Anstiegs bzw. des Abfalles klarer Weise länger,

[5] Ohne explizite Genehmigung aus der Chefetage können also Mitarbeiter/innen bei Beträgen bis zu 300 € einfach so bestellen; d. h. ein Computer um 1000 € kann nicht „einfach so" angeschafft werden, sehr wohl aber eine Tonercassette für einen Drucker um 140 €.

Tab. 3 Wahrscheinlichkeiten $P_n(1)$: Obergrenzen $O_m(1) = \frac{11\ldots1}{19\ldots9}$ und Untergrenze $1/9$

n	$P_n(1)$
1	$1/1$
9	$1/9$
19	$11/19$
99	$11/99 = 1/9$
199	$111/199$
999	$111/199 = 1/9$
1999	$1111/1999$
9999	$1111/1999 = 1/9$
19999	$11111/19999$

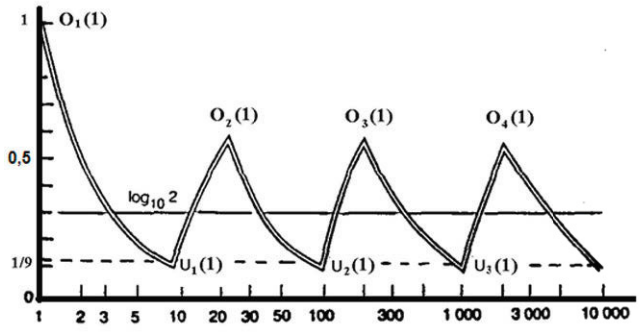

Abb. 2 Verlauf von $P_n(1)$ als kontinuierliche Funktion von n

die Folge $(P_n(1))_{n\in\mathbb{N}}$ ist sicher divergent. Einfach durch Grenzwertbildung ($\lim_{n\to\infty}$) kann man also nicht zur gesuchten Wahrscheinlichkeit „P(1)" in ganz $\mathbb{N}$ kommen (falls diese überhaupt existiert), denn $P_n(1)$ schwankt immer zwischen $1/9$ (Untergrenze!) und einer nur fast gleichbleibenden Obergrenze $O_m(1)$ ($^{1}1/19, {}^{111}1/199, {}^{111}1/1999, \ldots$) ; bei der m-ten Obergrenze O_m bezeichne m die Anzahl der Ziffern in Zähler bzw. Nenner. Schon an dieser Stelle ist plausibel, dass $P(1) > 1/9$ ist, da für „fast alle" $n \in \mathbb{N}$ der entsprechende Wert $P_n(1) > 1/9$ ist (nur bei $n = 10^1 - 1$ ist $P_n(1) > 1/9$). Salopp formuliert: nach jeder neu dazukommenden Stelle wird die Ziffer 1 lange Zeit „bevorzugt" (zwischen 10 ... 0 und 19 ... 9), bevor die anderen Ziffern der Reihe nach diesen Rückstand wieder aufholen.

In Abb. 2 ist der Verlauf von $P_n(1)$ als kontinuierliche Funktion in Abhängigkeit von n graphisch dargestellt, so dass die oben schon angesprochenen Phasen des Anstieges bzw. Abfalles von $P_n(1)$ auch *sichtbar* werden. (Die n-Achse wurde logarithmisch skaliert, damit die Abstände zwischen zwei aufeinander folgenden Zehnerpotenzen nicht immer größer werden, sondern gleich bleiben; die Werte $P_n(1)$ „schwanken" in einem gewissen Sinn um den Wert $\log_{10}(2)$.)

Die Obergrenzen $O_m(1)$ könnten auch im Schulunterricht näher studiert werden. Was passiert mit diesen Obergrenzen? Existiert ein Grenzwert für $m \to \infty$? Eine Frage, die durchaus von Schülern/innen selbständig bearbeitet werden kann (Vernetzung: Grenzwert von Folgen), indem sie die Zahlen $O_m(1) = \frac{11\ldots1}{19\ldots9}$ näher betrachten. Sie sollten z.B. in der Lage sein zu beweisen, dass die Folge dieser Obergrenzen $O_m(1)$ monoton fallend und nach unten beschränkt ist (z.B. durch 0 oder sogar durch ½), wodurch die Konvergenz gesichert wäre. Die Schüler/innen werden auch schnell (z.B. durch Probieren, d.h. Berechnen einiger Werte mit dem Computer) auf die Vermutung kommen, dass es der Wert $0,\overline{5} = 5/9$ ist, dem sich diese Zahlen von oben nähern. Diese Tatsache kann nun durch direktes Einsetzen in die Grenzwertdefinition auch bewiesen werden. Dafür ist zu zeigen, dass es für alle $\varepsilon > 0$ ein $M(\varepsilon)$ gibt, so dass für alle $m > M(\varepsilon)$ die Beziehung

$$-\varepsilon < \frac{11\ldots1}{19\ldots9} - \frac{5}{9} < \varepsilon \text{ gilt;}$$

man sieht rasch, dass immer $\frac{11\ldots1}{19\ldots9} - \frac{5}{9} > 0$ ist,

so dass man sich auf die rechte Ungleichung beschränken kann; diese ist für $19 \ldots 9 > \frac{4}{9\cdot\varepsilon}$ erfüllt (so ist der Grenzwert sogar *ohne* explizite „Termdarstellung" der Zahlen

$O_m(1) = \frac{11\ldots1}{19\ldots9}$ bestimmbar).

Einfacher ist die Grenzwertbestimmung, wenn man einen geschlossenen Term für die Zahlen $O_m(1)$ findet. Auch dies sollte für die Schüler/innen kein

Problem darstellen: die Zähler 11…1 (m Stellen) können z.B. als $(10^m–1)/9$ und die Nenner 19…9 z.B. durch

$10^m/5–1$ oder durch $2 \cdot 10^{m–1} –1$

dargestellt werden, wodurch man für $O_m(1)$ die geschlossene Darstellung

$$O_m(1) = \frac{5}{9} \cdot \frac{10^{m-1}}{10^{m-5}}$$

und damit den Grenzwert $^5/_9$ leicht finden kann.

Die Ober- und Untergrenzen der anderen Ziffern 2, 3, … , 9

Betrachten wir als nächstes z.B. die Ziffer 9. Verfolgen wir (bei wachsendem n) analog die Werte $(P_n(9))_{n\in\mathbb{N}}$, d.h. die Wahrscheinlichkeiten, dass die erste Ziffer einer Zahl 9 ist, wenn aus den ersten n natürlichen Zahlen zufällig gezogen wird:

Bei $n = 1, \ldots, 8$ ist $P_n(9) = 0$, bei $n = 9$ ist $P_9(9) = 1/9$ usw., diese Wahrscheinlichkeit sinkt dann bis $P_{89}(9) = 1/89$ bei $n = 89$. Dann steigt die Wahrscheinlichkeit wieder bis $n = 99$ (auf $P_{99}(9) = {}^{11}/_{99} = 1/9$), um dann wieder bis $n = 899$ abzufallen auf $P_{899}(9) = {}^{11}/_{899}$), dann kommt wieder eine hundert Zahlen lange Aufholphase bis $n = 999$ auf $P_{999}(9) = {}^{111}/_{999} = 1/9$), der wieder ein Abfall bis $n = 8999$ folgt usw. (siehe Tab. 4 und Abb. 3).

$P_n(9)$ schwankt immer zwischen $1/9$ (Obergrenze) und der Untergrenze $U_m(9) = 0, 1/89, {}^{11}/_{899}, {}^{111}/_{8999}, {}^{1111}/_{89999} \ldots$. Es ist plausibel, dass für $P_n(9)$ (falls existent) $P_n(9) < 1/9$ gilt, da für „fast alle" $n \in \mathbb{N}$ der entsprechende Wert $P_n(9) < 1/9$ ist (nur bei $n = 10^1 –1$ ist $P_n(9) = 1/9$). Salopp formuliert: nach jeder neu dazukommenden Stelle (also nach 10…0) hat die Ziffer 9 lange Zeit „Rückstand", bis sie unmittelbar vor der nächsten dazukommenden Stelle den Rückstand wieder aufholt (zwischen 90…0 und 99…9). Die Ziffer 9 hat also nie einen „Vorteil" gegenüber einer anderen, sie kann ihren Rückstand nur manchmal wettmachen!

Tab. 4 Wahrscheinlichkeiten $P_n(9)$: Obergrenze $1/9$ und Untergrenzen $U_m(9) = \frac{1\ldots1}{89\ldots9}$

n	$P_n(9)$
8	0
9	$1/9$
89	$1/89$
99	${}^{11}/_{99} = 1/9$
899	${}^{11}/_{899}$
999	${}^{111}/_{999} = 1/9$
8999	${}^{111}/_{8999}$
9999	${}^{1111}/_{9999} = 1/9$
89999	${}^{1111}/_{89999}$

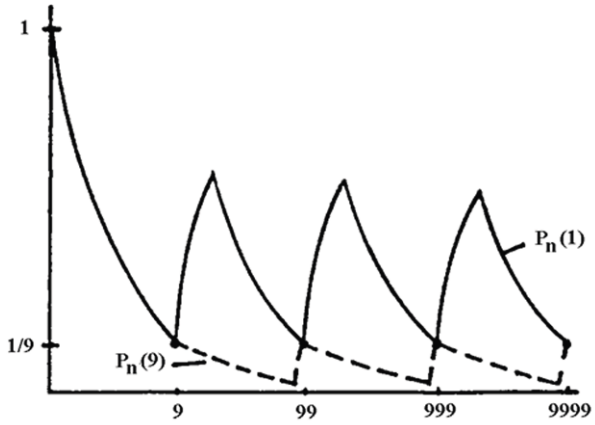

Abb. 3 Gleichzeitige graphische Darstellung der Verläufe von $P_n(1)$ und $P_n(9)$ als kontinuierliche Funktionen von n

Die Folge der Untergrenzen $U_m(9) = \frac{1\ldots1}{89\ldots9}$ (m sei die Anzahl der Einsen im Zähler) ist monoton wachsend und nach oben beschränkt, sie konvergiert gegen $1/81$, wie man durch analoge Überlegungen wie oben feststellen kann

(explizite Darstellung: $U_m(9) = \frac{1}{81} \cdot \frac{10^m - 1}{10^m - \frac{1}{9}}$).

In Abb. 3 sieht man nochmals deutlich: $1/9$ ist der *kleinste* Wert im Verlauf von $P_n(1)$ („das Schlimmste, was der Ziffer 1 passiert"), während $1/9$ der *größte* Wert im Verlauf von $P_n(9)$ ist („das Beste, was der

Tab. 5 Wahrscheinlichkeiten $P_n(3)$: Obergrenze $O_m(3) = \frac{11\ldots1}{39\ldots9}$ und Untergrenzen $U_m(3) = \frac{1\ldots1}{29\ldots9}$

n	$P_n(3)$
2	0
3	1/3
29	1/29
39	11/39
299	11/299
399	111/399
2999	111/2999
3999	1111/3999
29999	1111/29999

Ziffer 9 passiert“), so dass in Abb. 3 die Relationen $P_n(1) > 1/9$, $P_n(9) < 1/9$ und $P_n(1) \gg P_n(9)$ auch „sichtbar“ sind.

Während bei der Ziffer 1 die Untergrenze konstant $U_m(1) = 1/9$ war und bei der Ziffer 9 dies die konstante Obergrenze $O_m(9)$ war, sind bei den anderen Ziffern d = 2, ..., 8 die Werte der jeweiligen Ober- bzw. Untergrenzen $O_m(d)$ bzw. $U_m(d)$ nicht mehr konstant. Tab. 5 gibt einen Überblick über die entsprechenden Werte bei der Ziffer 3.

Die Schüler/innen können ganz selbständig die Form der Unter- bzw. Obergrenzen für jede beliebige Ziffer d erarbeiten: ganz allgemein haben diese für die Ziffer d die Gestalt (m sei dabei die Anzahl der Einsen im Zähler):

$$U_m(d) = \frac{1\ldots1}{(d-1)9\ldots9} = \frac{1}{9d} \cdot \frac{10^m - 1}{10^m - 1/d} \text{ und}$$

$$O_m(d) = \frac{11\ldots1}{d9\ldots9} = \frac{10}{9(d+1)} \cdot \frac{10^m - 1}{10^m - 10/(d+1)}$$

Die Schüler/innen sollten erkennen (und begründen) können, dass die Obergrenzen $O_m(d)$ monoton fallend und die Untergrenzen $U_m(d)$ monoton wachsend sind, wobei auch deren Grenzwerte für $m \to \infty$ nicht schwierig zu finden sind (durch die explizite Darstellung der Zähler und Nenner und somit der Werte $U_m(d)$ und $O_m(d)$ selbst):

$$U(d) := \lim_{m\to\infty} U_m(d) = \lim_{m\to\infty}\left(\frac{1}{9d} \cdot \frac{10^m - 1}{10^m - 1/d}\right) = \frac{1}{9d}$$

und

$$O(d) := \lim_{m\to\infty} O_m(d) =$$

$$= \lim_{m\to\infty}\left(\frac{10}{9(d+1)} \cdot \frac{10^m - 1}{10^m - 10/(d+1)}\right) = \frac{10}{9(d+1)}$$

Die so definierten (Grenz-)Werte $U(d)$ bzw. $O(d)$ sind natürlich Schranken für die Wahrscheinlichkeiten $P(d)$, falls diese überhaupt existieren. Man sieht (vgl. auch Tab. 6), dass die Intervalle $[U(d); O(d)]$ für alle $d = 1,\ldots,9$ den Wert 1/9 enthalten, mit den beiden Extremfällen: bei $d = 1$ als konstante Untergrenze und bei $d = 9$ als konstante Obergrenze – bei $n = 9, 99, 999\ldots$ ist ja $P_n(d) = 1/9$ für alle $d = 1,\ldots,9$ (die Wahrscheinlichkeiten $P_n(d)$ „kehren mit wachsendem n immer wieder zum Wert 1/9 zurück“).

Als Schätzwert für $P(d)$ böte sich in erster Näherung z. B. der jeweilige Mittelwert

$$P(d) \approx \frac{U(d) + O(d)}{2} \text{ an,}$$

eine Schätzung, die – wie wir sehen werden – gar nicht so schlecht ist. Tab. 6 gibt einen Überblick über die jeweiligen Werte von $U(d)$ und $O(d)$ (als Intervallgrenzen) in exakter Form und in Dezimaldarstellung (auf drei Dezimalen gerundet); des Weiteren ist der jeweilige Intervallmittelpunkt

$$\frac{U(d) + O(d)}{2}$$

angegeben und der wirkliche Wert $P(d)$ (auf drei Dezimalen gerundet; siehe Tab. 6. Ein Vergleich dieser beiden Werte zeigt, dass der jeweilige Schätzwert durch den Intervallmittelpunkt eine Differenz zum jeweils exakten Wert von nur 0,016 (bei $d = 9$) bis 0,037 (bei $d = 2$) aufweist.

Ein möglicher Schritt zur Verbesserung der Genauigkeit der durch die Intervallmittelpunkte gegebenen Schätzwerte könnte folgender sein: Da die Summe dieser Schätzwerte (vierte Spalte in Tab. 6

Tab. 6 Überblick über die Werte von $U(d)$ bzw. $O(d)$

d	$[U(d)\,;O(d)]$ Brüche	$[U(d)\,;O(d)]$ 3 Dezimalen	$(U(d)+O(d))/2$ 3 Dezimalen	$\mathrm{P}(d)=\log(d+1)-\log(d)$ 3 Dezimalen
1	[1/9 ; 5/9]	[0,111 ; 0,555]	0,333 (0,271)	0,301
2	[1/18 ; 10/27]	[0,056 ; 0,370]	0,213 (0,173)	0,176
3	[1/27 ; 5/18]	[0,037 ; 0,278]	0,157 (0,128)	0,125
4	[1/36 ; 2/9]	[0,028 ; 0,222]	0,125 (0,102)	0,097
5	[1/45; 5/27]	[0,022 ; 0,185]	0,104 (0,085)	0,079
6	[1/54 ; 10/63]	[0,019 ; 0,159]	0,089 (0,072)	0,067
7	[1/63 ; 5/36]	[0,016 ; 0,139]	0,077 (0,063)	0,058
8	[1/72 ; 10/81]	[0,014 ; 0,123]	0,069 (0,056)	0,051
9	[1/81 ; 1/9]	[0,012 ; 0,111]	0,062 (0,050)	0,046

nicht 1, sondern 1,229 ergibt, ist es naheliegend, alle diese Werte durch 1,229 zu dividieren, was bei $d=2,\ldots,9$ die Genauigkeit beträchtlich und bei $d=1$ auch ein wenig erhöhte (Werte in Klammern in der vierten Spalte von Tab. 6). Die Differenz zum exakten Wert beträgt dann nur 0,003 bis 0,006 bei $d=2,\ldots,9$ und 0,030 bei $d=1$. So könnten plausible und passable Schätzwerte für die Wahrscheinlichkeiten $\mathrm{P}(d)$ gewonnen werden – und zwar *a priori*, ohne folgende (oder andere) Überlegungen, die den *mathematischen* Hintergrund näher beleuchten.

Nun verlassen wir den Bereich der natürlichen Zahlen und wenden uns den (positiven) reellen Zahlen zu (als „Topf, aus dem eine Zufallszahl gezogen wird“). Für Größen, die *zeitlichen Veränderungen* unterworfen sind, haben wir oben schon ein passendes Zitat von Dworschak 1998 gegeben über den DAX. Auch für Größen, die sich im Lauf der Zeit nicht wesentlich ändern, kann man intuitive Überlegungen anstellen:

„Es gibt einfach mehr Pfützen als Tümpel, mehr Tümpel als Ozeane. Folglich gibt es wahrscheinlich auch mehr Gewässer zwischen 10 und 20 Hektar als zwischen 20 und 30, mehr zwischen 100 und 200 als zwischen 200 und 300 – und so fort. Damit ist Benfords Gesetz vollends auf dem Weg zur Weltenformel. Denn es gibt auch mehr Kieselsteine als Felsbrocken und überhaupt mehr kleine Dinge als große. Warum sich dies so verhält, ist wieder eine andere Frage.“ (Dworschak 1998, S. 229)

4 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[6] 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 20‰ = ⅓ 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 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 ⅑ von $\Omega=[1;10]$ aus.

Bei unbeschränktem Ω ist dies aber i.A. gar nicht mehr so leicht, es bedarf oft langwieriger Überlegungen aus der so genannten *Maßtheorie*. Wie groß sind „relative Anteile" bei unbeschränkten Mengen? Klar ist aber auch hier, dass das Wahrscheinlichkeitsmaß jeder *beschränkten* Menge den Wert 0 zuordnen muss, denn

$$\text{„}\frac{|\text{beschränkt}|}{|\text{unbeschränkt}|}=\frac{\text{endlich}}{\infty}=0\text{“}.$$

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

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

ist eine Vereinigung unendlich vieler halboffener Intervalle und ist in Abb. 4 dargestellt:

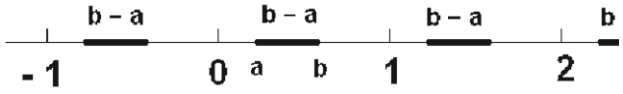

Abb. 4 Die Menge $K_{a,b}$ mit $P(K_{a,b}) = b - a$

Kein Intervall $[n; n+1[$ sei bevorzugt, innerhalb dieser Intervalle sind die Ankunftszeiten gleichverteilt, d.h. das gesuchte Maß entspricht dem relativen Anteil von $K_{a,b}$ an $\mathbb{R}$. Hier ist auch intuitiv klar (vgl. obiges Beispiel): Dieser relative Anteil ist $b - a$ (in jedem Teilintervall mit $(b - a) : 1$ abzulesen). Die Wahrscheinlichkeit, dass eine „zufällig" gewählte Ankunftszeit (bzw. reelle Zahl) in $K_{a,b}$ liegt[7], ist daher mit $b - a$ zu quantifizieren:

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

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

$$Z_d=\bigcup_{n=-\infty}^{\infty}\left[d\cdot 10^n\ ;\ (d+1)\cdot 10^n\right[\qquad d=1,\ldots,9$$

Unser Ziel ist

$$P\left(Z\in Z_d\right)=\lg(d+1)-\lg(d)$$

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:

$$\lg\left(Z_d\right)=\bigcup_{n=-\infty}^{\infty}\left[n+\lg(d)\ ;\ n+\lg(d+1)\right[$$

[6] „Zufällig" soll hier heißen, dass die Ankunftszeiten der Autos gleichverteilt angenommen werden: kein Intervall und kein Zeitpunkt soll bevorzugt werden.

[7] Gleichverteilung von Z vorausgesetzt.

D.h. die Menge $\lg\left(Z_d\right)$ ist nichts anderes als

$$K_{\lg(d),\,\lg(d+1)}$$

in obiger Notation ($a=\lg(d)$, $b=\lg(d+1)$). Wie können wir dieser Menge ein Maß bzw. eine Wahrscheinlichkeit zuordnen? Bei dem obigen einfachen Argument für $\mathrm{P}(Z\in K_{a,b})=b-a$ war ja Gleichverteilung von Z die Voraussetzung („relativer Anteil von Mengen“). Verwenden wir dieses Argument hier analog, so müssen wir voraussetzen, dass $\lg(Z)$ gleichverteilt ist. Damit erhalten wir:

$$\mathrm{P}\left(\lg\left(Z_d\right)\right)=\lg(d+1)-\lg(d)$$

Man könnte nun etwas vordergründig bzw. voreilig argumentieren:
Wegen

$$\lg(Z)\in\lg\left(Z_d\right)\Leftrightarrow\ Z\in Z_d$$

erhalten wir „klarer Weise“

$$\mathrm{P}\left(Z\in Z_d\right)=\mathrm{P}\left(\lg(Z)\in\lg\left(Z_d\right)\right)=\lg(d+1)-\lg(d)$$

und dadurch das gewünschte Resultat

$$\mathrm{P}\left(Z\in Z_d\right)=\lg(d+1)-\lg(d)\,.$$

Diese Folgerung gilt aber nur unter der Voraussetzung, dass $\lg(Z)$ (und nicht Z) gleichverteilt ist. Die Frage muss also lauten: Warum ist $\lg(Z)$ gleichverteilt?

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

Einschub: 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 (im Sinne der „geometrischen Wahrscheinlichkeit“: kein Bereich bevorzugt) aus $[0;1[$ gezogen werden.

Das interessierende Intervall sei dabei

$$I=\left[0;\frac{1}{2}\right[\,.$$

Mit günstigen und möglichen Intervalllängen argumentiert ergibt sich

$$\mathrm{P}(Z\in I)=1/2\,.$$

Nun betrachten wir die Quadratfunktion

$$f:Z\mapsto Z^2\ .$$

Wegen $f\left([0;1[\right)=[0;1[$ und $f(I)=\left[0;\frac{1}{4}\right[$ ergibt sich „analog“ $\mathrm{P}\left(f(Z)\in f(I)\right)=1/4\neq 1/2\,.$

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

$$Z\in I\Leftrightarrow f(z)\in f(I)\ \text{muss natürlich}$$

$$\mathrm{P}\left(Z\in I\right)=\mathrm{P}\left(f(Z)\in f(I)\right)\text{sein.}$$

Oft werden Laplace- oder „geometrische“ Wahrscheinlichkeiten in sehr naiver Weise benutzt

$$\mathrm{P}=\frac{|\text{günstig}|}{|\text{möglich}|}\,,$$

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 zu wenig Gedanken macht, ob wirklich kein Ausgang des Zufallsexperiments bevorzugt ist, d.h. ob wirklich Gleichverteilung vorliegt (widrigenfalls wäre ja

$$\mathrm{P}=\frac{|\text{günstig}|}{|\text{möglich}|}\ \text{falsch}).$$

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.

5 Skaleninvarianz und die Gleichverteilung der logarithmierten Werte

Wenn es überhaupt ein Verteilungsgesetz für die erste Ziffer von Zahlen gibt[8], 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. Es wäre ja wirklich höchst merkwürdig, wenn das Verteilungsgesetz von den gewählten Maßeinheiten abhinge, durch Wechsel vom anglo-amerikanischen ins metrische System würde sich dieses Gesetz ändern.

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, so muss man die entsprechenden Werte mit $s = 1{,}609344$ multiplizieren, wenn Preise von Dollar in Euro umgerechnet werden, so muss man die Zahlen durch ca. 1,33 dividieren[9].

D.h. ein Verteilungsgesetz für die erste Ziffer von Zahlen soll sich 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. es darf sich nicht ändern, wenn alle Werte mit einer positiven konstanten Zahl multipliziert werden.

Welche Verteilungsgesetze für die erste Ziffer kommen dafür in Frage?

Zunächst ein Test, ob Gleichverteilung der 1. Ziffer die Bedingung der Skaleninvarianz erfüllt: Dazu nehmen wir einmal an, dass alle Ziffern 1, ..., 9 gleichwahrscheinlich als führende Ziffer wären[10] und betrachten als Beispiel eine Vielzahl von Geldwerten in Euro. Bei einer Währungsänderung, z.B. wenn man statt der €-Werte in Deutschland in die alte DM-Welt zurückfallen möchte, muss jeder Geldwert mit (dem gerundeten Wert) 2 multipliziert werden. Bleibt dabei die ursprünglich in der €-Welt angenommene Gleichverteilung erhalten?

Nein, denn: Alle €-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 ist deutlich bevorzugt! D.h. die *Gleichverteilung* als Verteilungsgesetz für die Ziffern 1, . . ., 9 als erste Ziffern ist *nicht skaleninvariant!*

Begründung, warum die Skaleninvarianz der Messwerte zwingend zu gleichverteilten Logarithmen führt

Wir interessieren uns für die erste Ziffer positiver reeller Messwerte 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 $1 \le M < 10$ ist[11] („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 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.

[8] Empirische Beobachtungen unterstützen die These, dass es ein solches gibt.

[9] Dieser Wert variiert natürlich von Tag zu Tag.

[10] So würde es die ursprüngliche Intuition eigentlich nahe legen: $P(1) = \ldots = P(9) = 1/9$.

[11] So wie Z kann auch M als Zufallsvariable aufgefasst werden, deswegen wieder ein Großbuchstabe.

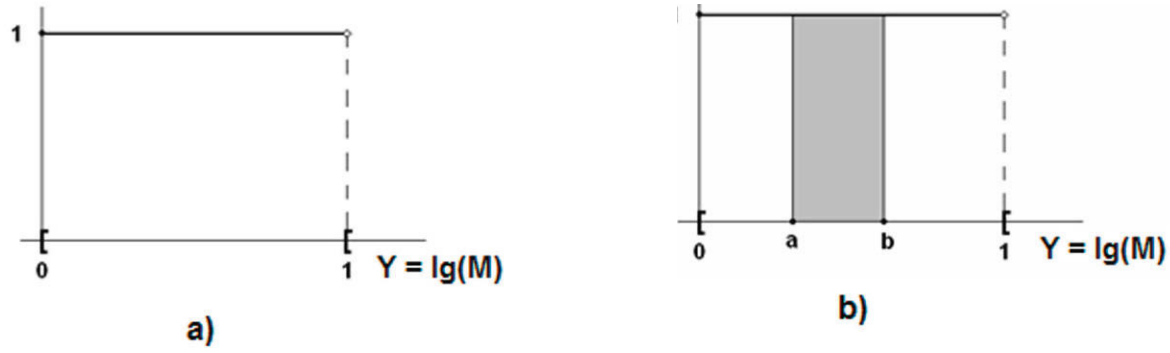

Abb. 5 Dichte der Gleichverteilung von $Y = \lg(M)$ auf [0;1[

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[12]: Wenn die Verteilungsgesetze von Z und $Z \cdot s$ gleich sind („Skaleninvarianz“), dann sind es auch jene von M und $M \cdot s$.

Wenn das Verteilungsgesetz für M skaleninvariant ist (d.h. die Verteilungsgesetze für M und $M \cdot s$ [13] gleich sind), 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[14].

Es ist aber bedeutend leichter die Frage zu beantworten, welche Verteilungsgesetze durch beliebige *Additionen* unverändert bleiben, als die ursprüngliche Frage, welche Verteilungsgesetze durch Multiplikationen mit beliebigen positiven Zahlen unverändert bleiben.

Verteilungsgesetze können durch Dichtefunktionen[15] beschrieben werden, wobei sich zugehörige Wahrscheinlichkeiten als Flächeninhalte unter den Graphen dieser Dichtefunktionen berechnen lassen.

Nun ist sehr *plausibel*, dass nur die konstante Funktion als Dichtefunktion so eines Verteilungsgesetzes in Frage kommt, das durch beliebige Additionen (d.h. horizontale Verschiebungen) nicht verändert wird – siehe Abb. 5a

Wie sonst sollte man jemals erreichen, dass sich die Dichtefunktion des zugrunde liegenden Verteilungsgesetzes durch beliebiges horizontales Verschieben nicht ändert[16]? 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 Dichte („Gleichverteilung“) lassen sich Wahrscheinlichkeiten der Art $P(a \le Y < b)$ besonders leicht ausrechnen (Flächeninhalt unter dem Graphen der Dichtefunktion zwischen a und b – siehe Abb. 5b:

[12] Für genauere Ausführungen dazu bräuchte man Maßtheorie, die hier aber vermieden werden soll.

[13] 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!

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

[15] Diskrete Verteilungen kommen für $Z \in \mathbb{R}^+$ nicht in Frage, jedes Intervall soll ja positive Wahrscheinlichkeit haben.

[16] 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. 5. Ein formaler Beweis für diese sehr plausible Tatsache ist hier wohl nicht nötig.

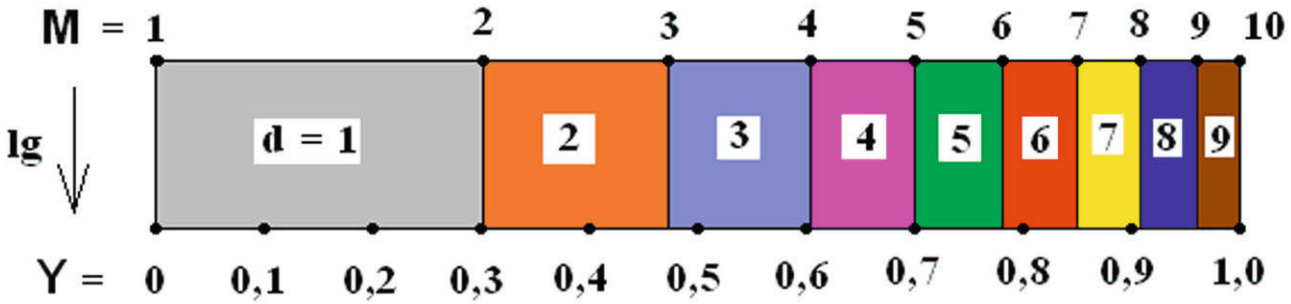

Abb. 6 Gleichverteilung von $Y := \lg(M)$ auf [0;1[bzw. die Verteilung von M auf [1;10[– „logarithmisch“

$$P(a \le Y < b) = (b-a)\cdot 1 = b-a\ .$$

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

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

Wir sind beim Benford-Gesetz angelangt!

Die Gleichverteilung der logarithmierten Werte ist hiermit also aus der Annahme der Skaleninvarianz der Messwerte abgeleitet. Die Lücke in der obigen, anschaulichen, zunächst nur vordergründigen Argumentation ist hiermit geschlossen! Das oben noch ausstehende „Warum sind die logarithmierten Werte gleichverteilt?“ kann beantwortet werden mit: Wegen der Skaleninvarianz der Messwerte! Diese Forderung erscheint vernünftig und kann nicht weiter bewiesen werden.

In Abb. 6 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 relativen Anteile der Ziffern d = 1, ..., 9 zu sehen: P (1. Ziffer von $Z = d$) = lg (d + 1) – lg(d).

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[17]), bei denen auch die Natur quasi geometrisch zählt, d.h. von Schritt zu Schritt immer mit einer Konstanten multipliziert:

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

Dies entspricht einem konstanten relativen Wachstum. Solche Werte haben dann die Eigenschaft,

[17] Vgl. die Lautstärkeeinheiten „Bel“ bzw. „Dezibel“, bei denen in Logarithmen gedacht werden muss. Auch bei der Tonhöhe werden Unterschiede als gleich wahrgenommen, wenn die Töne dasselbe Frequenzverhältnis haben. Dies alles wird subsumiert unter „Weber-Fechner'sches Grundgesetz“.

dass sich die logarithmierten Werte um eine additive Konstante unterscheiden:

$$\log(a) \overset{+\log(q)}{\rightarrow} \log(a)+\log(q) \overset{+\log(q)}{\rightarrow} \log(a)+2\log(q) \overset{+\log(q)}{\rightarrow} \log(a)+3\log(q)...$$

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.

6 Zusammenfassung und Ausblick

Empirische Daten legen es nahe, dass es ein stabiles Verteilungsgesetz für die Häufigkeit des Auftretens der 1. Ziffer von Zahlen gibt, und diese Daten sprechen auch dafür, dass diese Verteilung sich in der Nähe des Benford-Gesetzes befindet. Wenn es ein stochastisches Gesetz gibt, dann muss es wohl unabhängig von den zugrunde gelegten Skalen sein. Es würde doch sehr befremdlich anmuten zu sagen: „Dass die Daten so gut der Benford-Verteilung folgen, liegt nur an der (zufällig genau passenden?) Wahl der Einheiten, bei anderen Einheiten gäbe es das Benford-Gesetz gar nicht.“ Durch Umwandlung der Datenmengen in andere Einheiten könnte man dies auch widerlegen und empirisch bestätigen, dass das Benford-Gesetz auch dann weiterhin Gültigkeit hat.

Wenn das Verteilungsgesetz der Zahl Z selbst und damit der Mantisse M (in wissenschaftlicher Schreibweise) skaleninvariant ist (und das ist eine sehr vernünftige bzw. plausible Annahme), dann sind die logarithmierten (Mantissen-) Werte zwingend gleichverteilt, woraus das Benford-Gesetz unmittelbar folgt.

Es gibt also bei Skaleninvarianz keine Alternative zum Benford-Gesetz, außer *kein* Gesetz (dagegen sprechen aber empirische Daten).

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 Ziffer d beginnt:

$$P(Z \in Z_d) = \log_a(d+1) - \log_a(d)$$

für $d = 1,2, ..., a-1$

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.

Diekmann, A. (2012): Making Use of Benford's Law for the Randomized Response Technique. In: Sociological Methods and Research 41, 2, 325–334.

Diekmann, A. u. B. Jann (2010): Benford's Law and Fraud Detection – Facts and Legends. In: German Economic Review 11, 397–401.

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.

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? 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.

Humenberger, H. (2008): Eine elementarmathematische Begründung des Benford-Gesetzes. In: Der Mathematikunterricht 54, 1, 24–34.

Hungerbühler, N.: Benfords Gesetz über führende Ziffern: Wie die Mathematik Steuersündern das Fürchten lehrt. http://www.educ.ethz.ch/content/dam/ethz/special-interest/dual/educeth-dam/documents/Unterrichtsmaterialien/mathematik/Benfords%20Gesetz%20%C3%BCber%20f%C3%BChrende%20Ziffern%20%28Artikel%29/benford.pdf
(Letzter Zugriff 14. 04. 2016)

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.

Rauch, B. u.a. (2011): Fact and Fiction in EU-Governmental Economic Data. In: German Economic Review 12, 3, 243–255.

Schuppar, B. u. H. Humenberger (2015): Elementare Numerik für die Sekundarstufe. Springer Spektrum, Berlin-Heidelberg.

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

Tödter, K.-H. (2009): Benford's Law as an Indicator of Fraud in Economics. In: German Economic Review 10, 3, 339–351.

Walthoe, J.: Looking out for number one. https://plus.maths.org/content/os/issue9/features/benford/index
(Letzter Zugriff 14. 04. 2016)