Benfords Gesetz über führende Ziffern: Wie die Mathematik Steuersündern das Fürchten lehrt

Norbert Hungerbühler, Fribourg

1 Vorspann

Das Benfordsche Gesetz bietet eine ganze Reihe von Anknüpfungspunkten für den gymnasialen Mathematikunterricht. Es weist insbesondere Bezüge auf

- zum Stellenwertsystem
- zu Logarithmen
- zur Statistik (Histogramme)
- zur Wahrscheinlichkeitsrechnung
- zu Folgen (und Reihen)

Ausserdem lässt sich am Benfordschen Gesetz exemplarisch darlegen, wie man ein mathematisches **Modell** aufstellt, und wie man es diskutieren, anwenden und hinterfragen kann. Insbesondere macht es den Unterschied zwischen einem phänomenologischen **Gesetz** und einem mathematischen **Satz** deutlich. Aber mehr noch: Im Zusammenhang mit dem Benfordschen Gesetz lassen sich **Simulationen** und spannende **Experimente** mit überraschenden Ergebnissen durchführen. Und zu guter Letzt: Obwohl das Benfordsche Gesetz auf einer vergleichsweise einfachen Beobachtung aufbaut, hat es doch in jüngster Zeit aktuelle und pfiffige **Anwendungen** gefunden, unter anderem im Bereich der forensischen Mathematik.

Dieser Artikel ist, grob gesagt, folgendermassen aufgebaut: Wir erklären im Abschnitt 2, was das Benfordsche Gesetz besagt und wie es gefunden wurde. Im Abschnitt 3 werden ausgewählte Beispiele dargestellt. Der Abschnitt 4 beleuchtet einige mathematische Aspekte des Benfordschen Gesetzes. Schliesslich werden im Abschnitt 5 einige real world applications vorgestellt.

2 Das Mantissengesetz von Newcomb

2.1 Das Geheimnis der abgegriffenen Seiten

1881 stellte der amerikanische Mathematiker und Astronom Simon Newcomb[1] (Abbildung 1) beim Betrachten seiner Logarithmentafeln fest, dass die vorderen Seiten deutlich stärker abgegriffen waren, als die hinteren (siehe Abbildung 2). Als genau beobachtender Mathematiker, fragte er sich sofort nach dem Grund. Um Newcombs Staunen zu verstehen, rufen wir uns kurz in Erinnerung, wie eine Logarithmentafel aufgebaut ist. In Abbildung 3 ist ein Ausschnitt einer Seite aus [25] wiedergegeben. In der linken Spalte unter "N." sind die ersten drei Stellen der Numeri aufgelistet, rechts davon die Mantissen der Logarithmen. Will man etwa den gemeinen oder Zehnerlogarithmus zum Numerus 30.58 finden, so überlegt man sich zunächst, dass dieser eine 1 vor dem Komma hat, und findet dann in der Kolonne 305 in der Spalte 8 den Wert 1.48544 (die ersten beiden Nachkommastellen stehen ganz vorn unter "L."). Ganz rechts finden sich unter "P.P." (partes proportionales) noch die Tabellen für die lineare Interpolation. Natürlich kann man die Tafeln auch in umgekehrter Richtung benutzen, um bei gegebenem Logarithmus den Numerus zu finden.

[1] Biographie siehe z. B. www-history.mcs.st-andrews.ac.uk/history

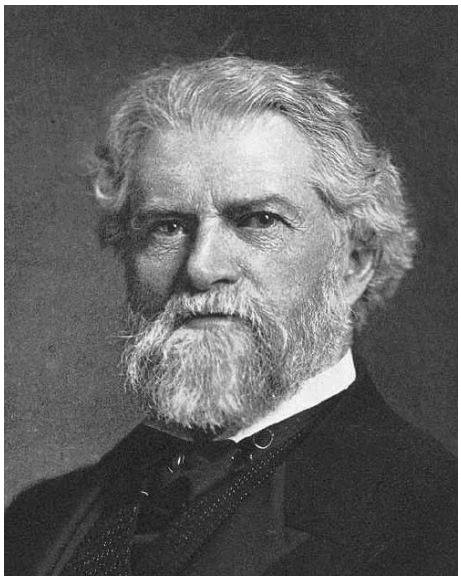


Abbildung 1: Simon Newcomb (1835–1909)

Abbildung 2: Die vorderen Seiten einer Logarithmentabelle sind stärker abgegriffen, als die hinteren.

N.	L. 0	1	2	3	4	5	6	7	8	9	P.P.	
300	47 712	727	741	756	770	784	799	813	828	842		
301	857	871	885	900	914	929	943	958	972	986		
302	48 001	015	029	044	058	073	087	101	116	130		
303	144	159	173	187	202	216	230	244	259	273		**15**
304	287	302	316	330	344	359	373	387	401	416	1	1,5
305	430	444	458	473	487	501	515	530	544	558	2	3,0
306	572	586	601	615	629	643	657	671	686	700	3	4,5
307	714	728	742	756	770	785	799	813	827	841	4	6,0
308	855	869	883	897	911	926	940	954	968	982	5	7,5
309	996	*010	*024	*038	*052	*066	*080	*094	*108	*122	6	9,0
310	49 136	150	164	178	192	206	220	234	248	262	7	10,5
311	276	290	304	318	332	346	360	374	388	402	8	12,0
											9	13,5

Abbildung 3: Ausschnitt aus einer Logarithmentafel

Jedenfalls ist nun klar, dass auf den vorderen Seiten der Loarithmentafeln zunächst die Numeri mit führender Ziffer 1, dann 2 usw. verzeichnet sind und auf den hinteren Seiten schliesslich die Numeri mit führender Ziffer 9 [2].

Was Newcomb also bemerkte war, dass in seiner Logarithmentafel häufiger Numeri mit führender Ziffer 1 nachgeschlagen wurden, als etwa mit mit führender Ziffer 8 oder 9. Die moderne Version dieser Beobachtung beschreibt Thomas Jech in seinem Artikel [16]: "When the 1 key on my old computer gave out I was not surprised".

Newcomb wäre nicht Mathematiker gewesen, wenn er nicht sofort versucht hätte, seine Beobachtung zu quantifizieren. In seinem Artikel [19] schrieb er als Quintessenz:

Mantissengesetz von Newcomb

Die Häufigkeit von Zahlen ist so, dass die Mantissen ihrer Logarithmen gleichverteilt sind.

Newcomb gibt zwar eine heuristische Begründung, spezifiziert jedoch nicht wirklich, für was für Zahlmengen dieses Mantissengesetz gelten sollte. Um es zu verstehen, bedarf zunächst der Begriff der Mantisse einer Erklärung, denn er wird nicht einheitlich verwendet. Newcomb versteht unter der Mantisse einer positiven Zahl ihren fraktionalen Teil: Für $x \in \mathbb{R}_+$ ist

$$\text{Mantisse von } x := \langle x \rangle := x - \lfloor x \rfloor \equiv x \mod 1$$

Beispiel $\langle \pi \rangle = 0.1415926\ldots$

Wie kam nun Newcomb zu seinem Mantissengesetz? Nehmen wir dazu eine Menge von positiven Zahlen, die wir uns auf der reellen Achse als Perlenkette aufgereiht denken (siehe Abbildung 4 oben). Im Beispiel in Abbildung 4 handelt es sich um Weibull-verteilte Zufallszahlen (siehe Abschnitt 4.2.2). Diese Zahlen unterwerfen wir nun der Logarithmusfunktion[3] (siehe Abbildung 4 Mitte). Anschliessend wird die Mantisse dieser Zahlen genommen, d.h. die Zahlen werden modulo 1 betrachtet. Anschaulich kann man sich das so vorstellen, dass die Perlenkette auf einem Kreis mit Umfang 1 aufgewickelt wird (siehe Abbildung 4 unten). Wenn sich die Logarithmen der Zahlen über einen genügend grossen Bereich erstrecken, mischen sich so die Mantissenwerte zu einer Gleichverteilung auf dem Kreis (respektive auf dem Intervall $[0, 1[$). Donald Knuth vergleicht die Situation mit einem Roulette-Tisch: Es gibt sozusagen keinen offensichtlichen Grund, warum sich die Logarithmen-Werte zum Beispiel in der Nähe der ganzen Zahlen häufen sollten (siehe [18, vol. 2, §4.2.4]). Wir werden im Abschnitt 4.2.2 sehen, wann dies alles bei Weibull-verteilten Zufallszahlen in der Tat in guter Näherung zutrifft.

2.2 Folgerungen aus dem Mantissengesetz

Wie konnte nun Newcomb mit Hilfe seines Mantissengesetzes das Phänomen der abgenutzten Seiten seiner Logarithmentafel erklären? Dazu nehmen wir an, eine Menge von zufälligen Zahlen sei so verteilt, dass sie dem Mantissengesetz gehorcht. Dann definieren wir für die Ziffern $i \in \{1, 2, \ldots, 9\}$ die Mengen

$$\begin{aligned} E_i &:= \bigcup_{k \in \mathbb{Z}} [\, i10^k, (i+1)10^k [\\ &= \{x \in \mathbb{R}_+ \ : \ \text{führende Ziffer von } x \text{ ist } i\} \end{aligned}$$

[2] Die zitierte Logarithmentafel führt hinten allerdings nochmals Numeri zwischen 10'000 und 11'009 und deren siebenstellige Logarithmen auf, weil das Rechnen mit Zinsfaktoren oft diese höhere Genauigkeit erforderlich macht.

[3] Wo nicht anders spezifiziert meinen wir immer den Zehnerlogarithmus.

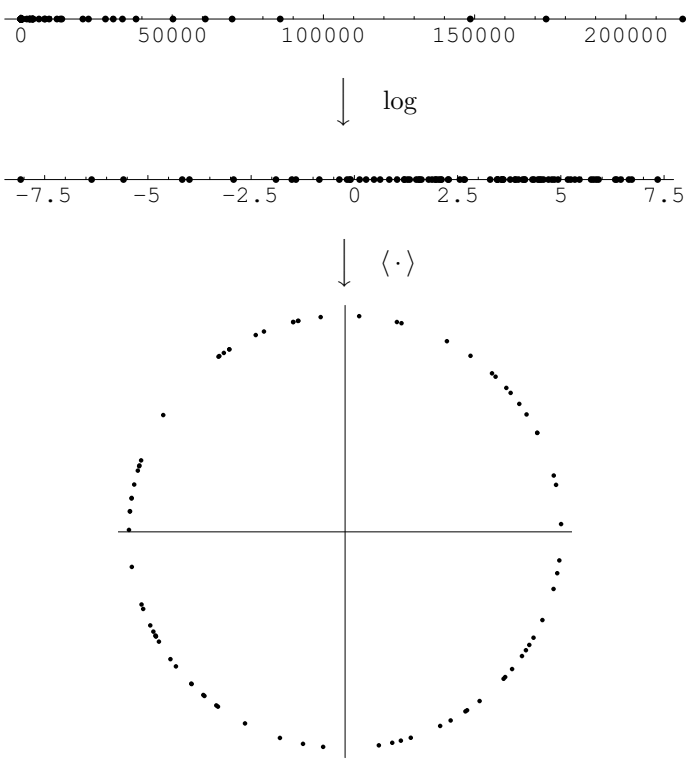


Abbildung 4: Heuristische Herleitung des Mantissengesetzes

Insbesondere ist dann $\mathbb{R}_+ = \bigcup_{i=1}^{9} E_i$. In Abbildung 5, oben, ist die Menge E_1 angedeutet (sie erstreckt sich natürlich noch weiter in Richtung Null und Unendlich). Darunter ist zu sehen, wohin die Menge E_1 unter der Funktion $\langle \log(\cdot) \rangle$ abgebildet wird. Man beachte dabei, dass die Funktion $\langle \log(\cdot) \rangle$ nichts davon merkt, wenn ihr Argument mit 10 multipliziert wird: $\langle \log(x) \rangle = \langle \log(10x) \rangle$. Wir möchten nun die Wahrscheinlichkeit berechnen, dass eine unserer Zufallszahlen in die Menge E_i zu liegen kommt, also mit der Ziffer i beginnt:

$$\begin{aligned} P(X \in E_i) &= P\big(\langle \log(X) \rangle \in [\log i, \log(i+1)[\big) \\ &= \log(i+1) - \log i \\ &= \log\big(1 + \frac{1}{i}\big) \end{aligned}$$

Dabei haben wir für die zweite Zeile eben das Mantissengesetz verwendet, wonach $\langle \log(X) \rangle$ gleichverteilt auf $[0, 1[$ ist. Die Formel

$$P(X \in E_i) = \log\big(1 + \frac{1}{i}\big)$$

für $i \in \{1, 2, \ldots, 9\}$ heisst **Benford's first significant digit law**. Die Abbildung 6 zeigt die entsprechende Wahrscheinlichkeitsverteilung der führenden Ziffern. Zunächst einmal ist dies ein kontraintuitives Ergebnis, denn weshalb sollte die 1 häufiger als führende Ziffer vorkommen als die 8 oder die 9? Andererseits sollte man das Benfordsche Gesetz weniger als eine Eigenschaft der Zahlen selber auffassen, sondern vielmehr als eine Eigenschaft unseres Stellenwertsystems, d.h. der Art und Weise, wie wir eben Zahlen darstellen.

Newcomb konnte aus dem Mantissengesetz natürlich in analoger Weise auf das Verhalten zum Beispiel der ersten *zwei* Ziffern schliessen. Wenn man etwa wissen will, wie gross die Wahrscheinlichkeit

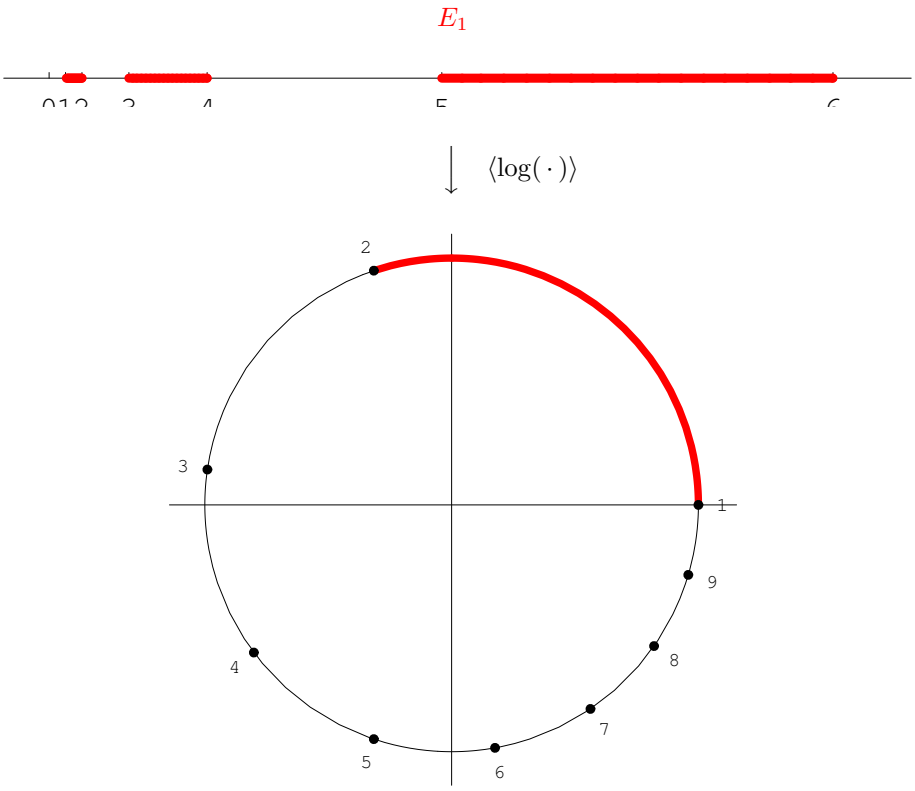


Abbildung 5: Drei Intervalle der Menge E_1 und deren Bild unter der Funktion $\langle \log(\,\cdot\,) \rangle$

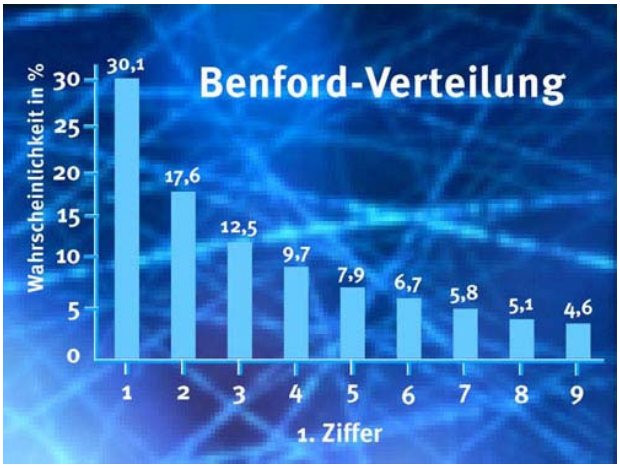


Abbildung 6: Benford-Verteilung

ist, dass unsere Zufallszahl mit der Ziffernfolge 31 beginnt, betrachtet man halt die Menge

$$\begin{aligned} E_{31} &:= \bigcup_{k\in\mathbb{Z}} [\,31\cdot 10^k, 32\cdot 10^k[\\ &= \{x \in \mathbb{R}_+ \;:\; x \text{ beginnt mit der signifikanten Ziffernfolge } 31\}. \end{aligned}$$

Man bekommt mit der gleichen Rechnung wie oben

$$P(X \in E_{31}) = \log\big(1 + \frac{1}{31}\big).$$

Natürlich lassen sich jede Menge schöne Aufgaben damit konstruieren, etwa diese beiden:

Aufgaben

- Wie gross ist die Wahrscheinlichkeit, dass die dritte Ziffer i ist?
- Wie gross ist die bedingte Wahrscheinlichkeit, dass die zweite Ziffer i ist, unter der Bedingung, dass die erste Ziffer j ist?

Auch die Antwort auf die zweite Frage ist wieder zunächst kontraintuitiv. Die Verteilung der zweiten Ziffer hängt von der ersten ab. Aber warum sollte die zweite Ziffer etwas von der ersten wissen?!

Das allgemeine (oder starke) Benfordsche Gesetz ergibt sich wie oben als **Folgerung aus dem Mantissengesetz**. Es lautet:

Allgemeines (oder starkes) Benfordsches Gesetz

Für $Z \in \mathbb{N}$ ist die Wahrscheinlichkeit, dass X mit der signifikanten Ziffernfolge $Z_{\overline{10}|}$ beginnt gegeben durch

$$P(X \in E_Z) = \log_{10}\big(1 + \frac{1}{Z}\big).$$

Dabei meint $Z_{\overline{10}|}$ die Ziffernfolge im Zehnersystem, und, wie zuvor,

$$\begin{aligned} E_Z &:= \bigcup_{k\in\mathbb{Z}} [\, Z \cdot 10^k, (Z+1) \cdot 10^k [\\ &= \{x \in \mathbb{R}_+ \ : \ x \text{ beginnt mit der signifikanten Ziffernfolge } Z\} \end{aligned} \tag{1}$$

Diaconis hat bemerkt, dass auch umgekehrt das Mantissengesetz aus dem allgemeinen Benfordschen Gesetz folgt (siehe [5]):

Mantissengesetz $\iff$ allgemeines Benfordsches Gesetz

Es schliessen sich an dieser Stelle zwei Bemerkungen an:

1. Zunächst einmal zeichnet sich die Basis 10 in keiner relevanten Weise vor anderen Basen aus. Was wir also bisher für das Zehnersystem festgehalten haben, lässt sich sofort auf jede andere Basis übertragen[4]. Immerhin schränken wir dies auf natürliche Basen ein[5]. Auf die Basis-Unabhängigkeit werden wir im Abschnitt 4.3 nochmals kurz zu sprechen kommen.

[4] Wir merken an, dass das "first digit law" für die Basis 2 trivial ist, da jede von Null verschiedene Zahl im Binärsystem 1 als führende signifikante Ziffer hat.

[5] Bei der Darstellung in gebrochenen Basen kommt es zu unerwünschten Effekten. Beispielsweise stimmt die lexikographische Ordnung nicht mehr mit der natürlichen Ordnung der Zahlen überein. Zahlensysteme zu gebrochenen Basen hängen eng mit dem Josephus-Problem zusammen (siehe [4]).

2. Die zweite Bemerkung bezieht sich auf die Skaleninvarianz des Mantissengesetzes. Diese besagt folgendes:

   Lemma *Ist X eine positive Zufallsvariable für welche $\langle \log X \rangle$ gleichverteilt ist, so gilt dies auch für λX, $\lambda > 0$.*

   Der Beweis erschöpft sich in der Bemerkung, dass $\langle \log(\lambda X) \rangle = \langle \log \lambda + \log X \rangle$.

   Anschaulich kann man sich die Skaleninvarianz etwa so vorstellen: Denken wir uns eine zufällige Menge von Quantitäten in der realen Welt, die durch Messungen in gewissen Einheiten zustande gekommen sind. Ändert man die Einheiten, so ändern sich die Zahlen. Werden auf einen Schlag alle Zahlen im Universum mit einer Konstanten multipliziert, so würden auch die geänderten Zahlen dem Benfordschen Gesetz folgen, wenn sie es vorher taten. Oder kurz und knapp: Wenn ein Datensatz in Metern Benford-verteilt ist, so ist er dies auch noch nach Umrechnung in Meilen. Ob auch eine andere als die Benford-Verteilung diese Eigenschaft der Skaleninvarianz besitzt, werden wir uns in Abschnitt 4.3 noch überlegen.

3 Empirische Belege für Benfords Gesetz

3.1 Benfords Daten

Kehren wir zurück zur Geschichte: Newcombs Artikel geriet leider bald nach seinem Erscheinen in Vergessenheit. Aber wie alle guten Gedanken, wurde auch dieser nochmals gefunden. 1938 gelangte der amerikanische General Electric Physiker Frank Benford (siehe Abbildung 7) zur selben

Abbildung 7: Frank Benford (1883–1948)

Schlussfolgerung, wie Newcomb ein halbes Jahrhundert vor ihm. Anders als Newcomb unterlegte Benford seine Beobachtung, das heute nach ihm benannte Gesetz, mit insgesamt 20'229 gesammelten Daten. Die Abbildung 8 zeigt Benfords Auswertung. Die erste Zeile erfasst die Verteilung der ersten Ziffern von Entwässerungsgebieten von 335 Flüssen. Die zweite Zeile gibt Einwohnerzahlen amerikanischer Ortschaften wieder. In der dritten Zeile wertete Benford physikalische Konstanten in einem Tabellenwerk aus. Die Zeile D handelt von Auflagen von Zeitschriften. Das geht dann munter weiter bis hin zu Zahlen aus Tabellenwerken mit Inversen und Wurzeln in Zeile K, Zahlen aus Artikeln des *Reader's Digest* in Zeile M oder Resultaten der American Football League in Zeile P. Viele Zeilen weisen überraschend gute Annäherung an die theoretischen Benford-Werte auf. Die vielleicht überraschendste Beobachtung ist aber, dass die Vereinigung aller Datensätze im letzten Teil der Tabelle die beste Approximation an Benford darstellt. Diese Beobachtung wollen wir nun an einem weiteren Beispiel wiederholen.

TABLE I

PERCENTAGE OF TIMES THE NATURAL NUMBERS 1 TO 9 ARE USED AS FIRST DIGITS IN NUMBERS, AS DETERMINED BY 20,229 OBSERVATIONS

Group	Title	First Digit									Count
		1	2	3	4	5	6	7	8	9	
A	Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
B	Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
C	Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
D	Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
E	Spec. Heat	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
F	Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
G	H.P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
H	Mol. Wgt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
I	Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
J	Atomic Wgt.	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
K	$n^{-1}, \sqrt{n}, \cdots$	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5000
L	Design	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
M	*Digest*	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
N	Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
O	X-Ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
P	Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1458
Q	Black Body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
R	Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
S	$n^1, n^2 \cdots n!$	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
T	Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
Average		30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1011
Probable Error		±0.8	±0.4	±0.4	±0.3	±0.2	±0.2	±0.2	±0.2	±0.3	—

Abbildung 8: Benfords Daten aus [3]

3.2 Gilt Benfords Gesetz für Autokennzeichen?

Gibt man seinen Schülern den Auftrag, zwei Stunden lang an einer viel befahrenen Strasse die erste Ziffer jedes Autokennzeichens zu notieren und anschliessend ein Histogramm zu erstellen, so lassen sich daran interessante Beobachtungen machen. Je nachdem, wo die Erhebung durchgeführt wurde, kann das Ergebnis recht unterschiedlich ausfallen. Betrachten wir zuerst Daten in Abbildung 9, die an einer Autobahnraststätte bei 388 Autos erhoben wurden. Die Verteilung folgt verblüffend gut

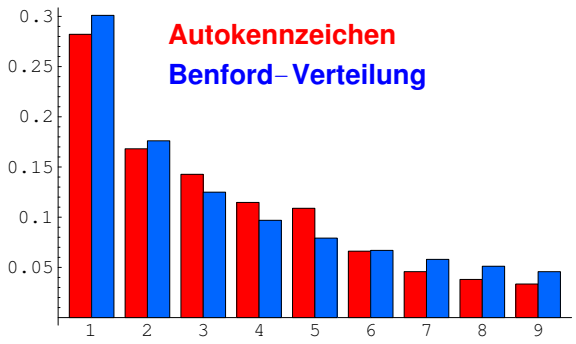


Abbildung 9: Verteilung führender Ziffern bei Autokennzeichen

dem Benfordschen Gesetz. Das Ergebnis erstaunt jedoch insofern, als dass das Benfordsche Gesetz für Autonummern aus *einem* Kanton ausdrücklich *nicht* gilt! Sehen wir uns die Sache genauer an:

Sei

$$\begin{aligned} E &:= \{x \in \mathbb{N} \colon x \text{ ist eine Autonummer im Kanton Z}\} \\ E_i &:= \{x \in E \colon x \text{ beginnt mit Ziffer } i\} \\ p_i &:= \frac{|E_i|}{|E|} \end{aligned}$$

Uns interessiert also, wie gross der Anteil p_i an Kennzeichen ist, die mit Ziffer i beginnen. Die Aufzählung der Zahlen, die mit 1 beginnen, liefert sofort den nötigen Überblick:

$$\underbrace{\underbrace{\underbrace{1}_{1}}_{11}, \quad 10, 11, \ldots, 19,}_{111} \quad 100, 101, \ldots, 199, \quad \ldots$$

Wir betrachten die beiden Extremfälle mit hohem respektive niedrigem Anteil an Kennzeichen, die mit Ziffer 1 beginnen. Einen besonders hohen Anteil bekommt man offenbar, wenn $|E|$ gerade am Anfang einer "Lücke" liegt: Für die führende Ziffer $i = 1$ und $|E| = 2 \cdot 10^n - 1$ ist dann nämlich

$$|E_1| \;=\; \frac{10^{n+1} - 1}{9}$$

also

$$p_1 \;=\; \frac{10^{n+1} - 1}{9(2 \cdot 10^n - 1)} \;=\; \frac{10 - \frac{1}{10^n}}{9(2 - \frac{1}{10^n})} \;\to\; \frac{5}{9} \quad \text{für } n \to \infty.$$

Andererseits ist der Anteil an Kennzeichen, die mit 1 beginnen am kleinsten, wenn $|E|$ gerade am Ende einer "Lücke" liegt: Dann ist $|E| = 10^n - 1$, und alle Mengen E_i weisen die selbe Mächtigkeit auf, d.h.

$$p_i = \frac{1}{9}\,.$$

Zusammenfassend halten wir fest:

$$\liminf_{|E|\to\infty} p_1(|E|) \;=\; \frac{1}{9} \;<\; \log(1 + \frac{1}{1}) \;<\; \frac{5}{9} \;=\; \limsup_{|E|\to\infty} p_1(|E|).$$

Das Strassenverkehrsamt des Kantons Zürich gab auf Anfrage bekannt, dass im Sommer 2005 die höchste vergebene Autonummer im Kanton zu einem bestimmten Zeitpunkt bei 782'500 lag. Die entsprechende Verteilung ist in Abbildung 10 wiedergegeben. Etwa so sähe daher auch die Verteilung aus, die unser Experiment an einer Dorfstrasse zum Beispiel in Volketswil liefern würde, wo fast ausschliesslich Anwohner verkehren. Die Abweichung vom Bendfordschen Gesetz ist offenkundig.

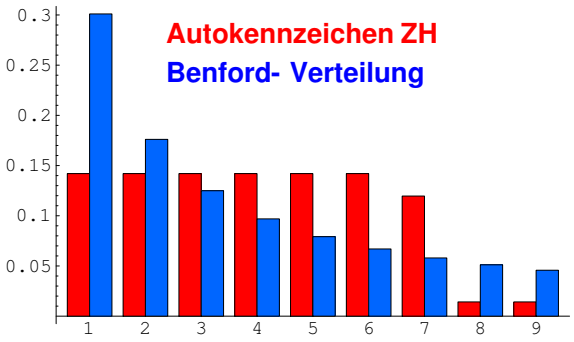


Abbildung 10: Verteilung führender Ziffern bei 782'500 Autokennzeichen im Kanton Zürich

Dass unser Experiment an der Autobahnraststätte dennoch eine angenäherte Benford-Verteilung ergab liegt daran, dass nicht nur Autos aus *einem* Kanton über die Autobahn fahren, sondern ein bunter Mix, der darüberhinaus auch noch einen Anteil ausländischer Fahrzeuge enthält. Diese Mischung liefert offenbar eine gute Näherung an die Benford-Verteilung. Hill konnte vor kurzem unter sehr allgemeinen Voraussetzungen zeigen, dass tatsächlich die Mischung von unterschiedlichen, je für sich nicht Benford-verteilten Zufallsgrössen, eine Benford-Verteilung ergibt. Wir werden im Abschnitt 4.2.3 noch kurz darauf zurückkommen.

3.3 Plouffe's Inverter

Simon Plouffe hat den wunderbaren "inverse symbolic calculator" entwickelt. Auf seiner Webseite[6] können über 200 Millionen Konstanten identifiziert werden. Gibt man beispielsweise die Zahl 22.2992216 ein so schlägt einem die Seite sofort vor, das dies wohl

$$\mathbf{22.29922164}797137\ldots = e^{\pi} - \sin 1$$

sei. Der Nutzen dieses Inverters liegt auf der Hand: Numerisch erhaltene Werte von Berechnungen können mit seiner Hilfe daraufhin überprüft werden, ob das Resultat eine bisher nicht erkannte Bedeutung hat. Plouffe hat 2001 die Verteilung seiner Konstanten nach den führenden vier Ziffern publiziert. Legt man die Kurve der theoretischen Benford-Verteilung darüber, zeigt sich eine verblüffende Übereinstimmung (siehe Abbildung 11).

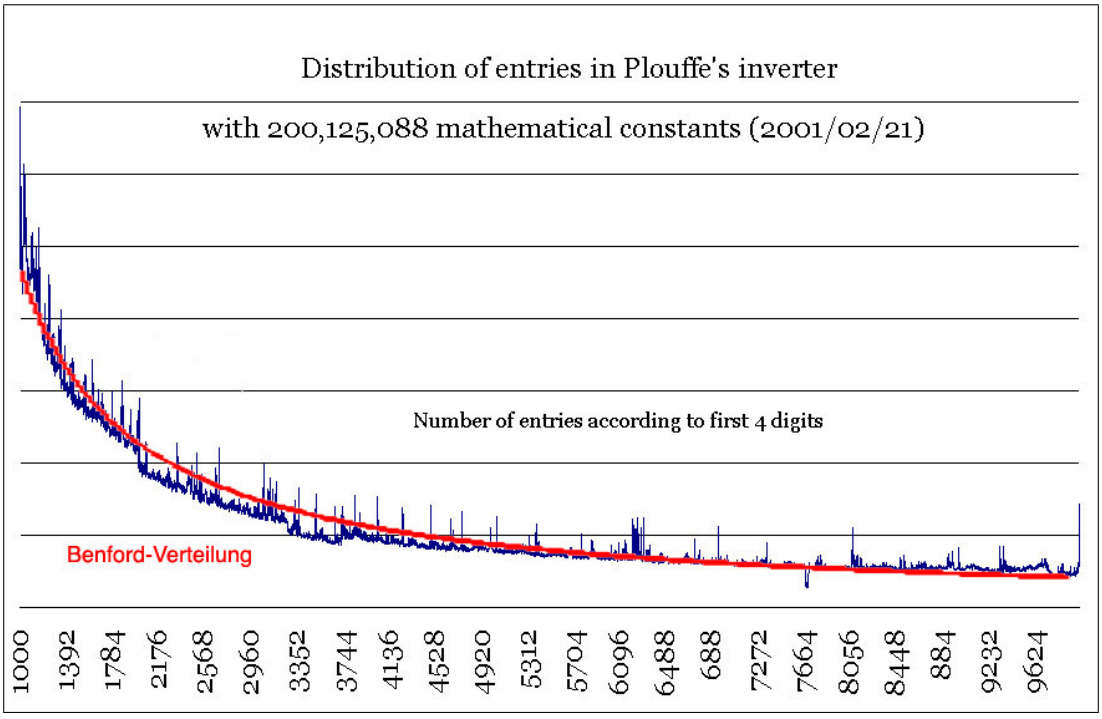


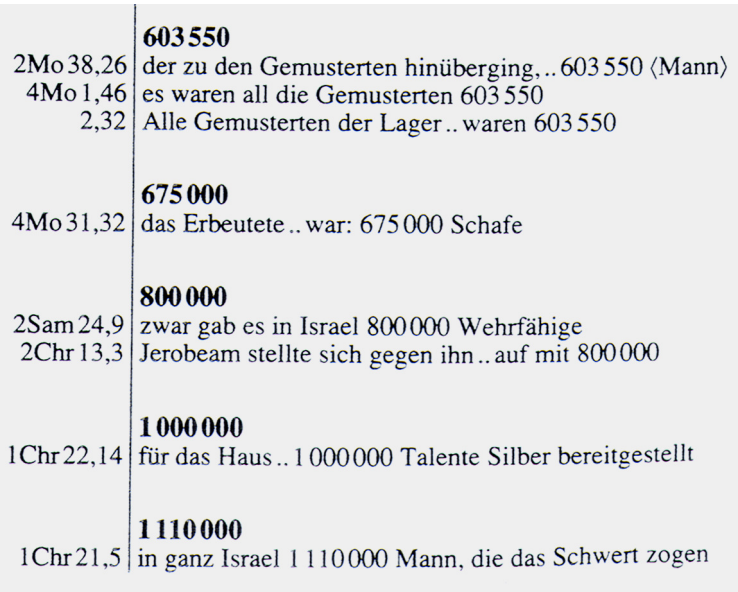
Abbildung 11: Verteilung der Konstanten in *Plouffe's Inverter* nach den vier führenden Ziffern

3.4 Findet man Benford in der Bibel?

Im Grunde erstaunt es, dass Benfords Gesetz nicht schon viel früher von Kabbalisten und Numerologen beobachtet worden ist. Kann das daran liegen, dass die Zahlen in der Bibel vielleicht

[6] http://pi.lacim.uqam.ca/eng

nicht Benford-verteilt sind? Die berühmte Elberfelder Konkordanz [27] besitzt einen eigenen Abschnitt über Zahlen. Dort sind also zu allen in der Bibel vorkommenden Zahlen die betreffenden Bibelstellen verzeichnet (siehe Abbildung 12). So kommt etwa die natürliche Zahl 603'550 insge-



603 550
2Mo 38,26 der zu den Gemusterten hinüberging, .. 603 550 ⟨Mann⟩
4Mo 1,46 es waren all die Gemusterten 603 550
2,32 Alle Gemusterten der Lager .. waren 603 550

675 000
4Mo 31,32 das Erbeutete .. war: 675 000 Schafe

800 000
2Sam 24,9 zwar gab es in Israel 800 000 Wehrfähige
2Chr 13,3 Jerobeam stellte sich gegen ihn .. auf mit 800 000

1 000 000
1Chr 22,14 für das Haus .. 1 000 000 Talente Silber bereitgestellt

1 110 000
1Chr 21,5 in ganz Israel 1 110 000 Mann, die das Schwert zogen

Abbildung 12: Ausschnitt aus der Elberfelder Konkordanz [27]

samt dreimal im Alten Testament vor. Mit Hilfe der Elberfelder Konkordanz lässt sich somit zum Beispiel leicht überprüfen, welches die kleinste natürliche Zahl ist, die nicht im neuen Testament vorkommt: Es ist die Zahl Dreizehn[7]. Dies ist einer von 13 Gründen, warum 13 als Unglückszahl gilt. Die kleinste natürlichen Zahl, die weder im alten, noch im neuen Testament erwähnt wird, ist übrigens 43. Anders ausgedrückt: Die natürlichen Zahlen von 1 bis 42 werden allesamt lückenlos aufgezählt. Dies mag erklären, wie die Antwort von "Deep thougt" auf Dougals Adams' "ultimate question of life, the universe and everything" zustande kam[8] (siehe [1, vol. 4]). Die grösste in der Bibel genannte Zahl ist übrigens 1'110'000. Vielleicht wird ja einem unserer Leser dies bei Günther Jauch einmal als Millionenfrage gestellt...

Aber zurück zu Benford: Mit Hilfe der Zahlenkonkordanz lassen sich nun recht einfach die in der Bibel vorkommenden Zahlen auf die Verteilung ihrer führenden Ziffern hin untersuchen. Abbildung 13 zeigt das Resultat. Mit zwei Ausnahmen ist die beobachtete Verteilung der Benford-Verteilung sehr ähnlich: Es gibt zu viele Zahlen, die mit 1 beginnen und solche, die mit 7 beginnen. Nun sind Abweichungen immer mindestens genauso interessant wie Übereinstimmungen. Wie lassen sich die vielen Einer erklären? Es bieten sich zwei Gründe an.

1. Es ist nicht auszuschliessen, dass bequemlichkeits- oder unwissenheitshalber manche Zahlen in der Bibel, die mit 8 oder 9 beginnen würden, auf die nächste Zehnerpotenz aufgerundet worden sind.

2. Im Deutschen stimmt der unbestimmte Artikel "ein" mit dem Zahlwort "ein" überein. In einem deutschen Text wäre es daher schwierig die beiden Bedeutungen auseinander zu halten. Im Hebräischen (und die Elberfelder Konkordanz basiert auf diesem Urtext), gibt es keine Artikel, sodass dieses Problem nicht besteht. Hingegen verweist die Zahlkonkordanz für die Zahl Eins auf den Eintrag "Ein" in der Wortkonkordanz. Dort aber wird unter "ein" sowohl auf das Wort "echad" (im Sinne von "einzig") als auch auf "jachid" (das Zahlwort "ein")

[7] Diese Beobachtung wurde dem Autor von Ernst Specker mitgeteilt.
[8] die andere Erklärung, "fourty two" meine eigentlich "tea for two" ist aber fast genauso überzeugend.

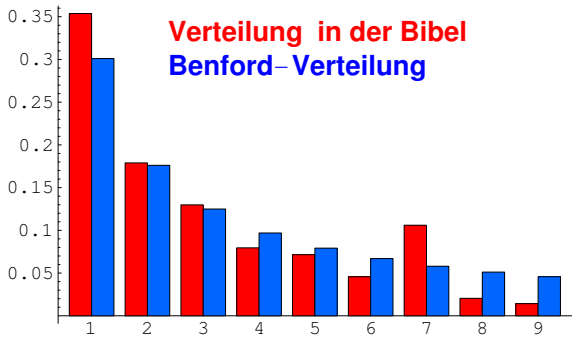


Abbildung 13: Verteilung der Zahlen in der Bibel nach führenden Ziffern

aufgenommen. Mit anderen Worten, um eine genaue Statistik zu haben, müsste man im hebräischen Urtext nachschauen und nur die Zahlwörter "jachid" zählen.

Dass die Sieben zu oft vorkommt erklärt sich aus der biblischen Zahlensymbolik, in welcher diese Zahl eine dominante Stellung einnimmt:

Biblische Symbolik der Zahl 7

- 7 bedeutet Vollkommenheit oder Vollständigkeit
- $7 \times 7 \times 7$ steht für Unendlichkeit
- 7 Tage der Schöpfungswoche
- 7 Bitten des Vaterunsers
- 7 Freuden der Maria
- 7 Gaben des heiligen Geistes
- 7 Worte Christi am Kreuz
- 7 Todsünden
- 7 Sakramente
- Passahfest und Laubhüttenfest dauern 7 Tage

Mit diesem kleinen Exkurs beschliessen wir vorerst die Reihe der Beispiele zum Benfordschen Gesetz ab.

4 Four roads to Benford

Bislang wurde in der Literatur das Benfordsche Gesetz aus vier verschiedenen Richtungen untersucht:

1. Heuristische Untersuchung von realen Datensätzen
2. Folgen (und Arrays)
3. Wahrscheinlichkeitsverteilungen
4. Strukturelle Analyse

Zum ersten Punkt haben wir im Abschnitt 3 verschiedene Beispiele gesehen, angefangen mit den Beobachtungen von Benford selber. Man betrachtet dort also in einem realen Datensatz die beobachtete relative Häufigkeit von Daten, die mit Ziffer i beginnen. Bei Folgen schaut man entsprechend, wie viele von den ersten n Folgengliedern mit der Ziffer i beginnen und betrachtet den Limes ihrer relativen Häufigkeit für $n \to \infty$ (falls dieser Limes in irgend einem Sinne existiert). Für eine gegebene Wahrscheinlichkeitsverteilung einer Zufallsvariable X schliesslich, berechnet man die Wahrscheinlichkeit des Ereignisses $X \in E_i$ (siehe Abschnitt 2.2). Mit struktureller Analyse meint man Eigenschaften, wie etwa die Skaleninvarianz am Ende des Abschnitts 2.2 oder im Abschnitt 4.3.

Mathematisch strenge Aussagen kann man nur zu den letzten drei Punkten machen und daraus allenfalls Einsichten darüber gewinnen, warum so viele reale Datensätze dem Benford-Gesetz gehorchen. Wir beleuchten daher nun nacheinander die Punkte 2 bis 4 und beginnen mit den Folgen.

4.1 Benford-Folgen

Die Idee ist hier ganz einfach: Man fragt sich, ob bei einer bestimmten gegebenen Folge der Anteil an Folgengliedern in einem Anfangsstück $a_1, a_2, \ldots, a_n$, die mit Ziffer i beginnen, für $n \to \infty$ asymptotisch gegen den Benford-Wert $\log(1 + \frac{1}{i})$ strebt. Statt nur Ziffern $i \in \{1, 2, \ldots, 9\}$ zu betrachten, kann man die Frage, wie wir gesehen haben, auf Ziffernfolgen ausdehnen. Das läuft dann aber, wegen der Bemerkung von Diaconis im Abschnitt 2.2, auf das Mantissengesetz hinaus. Wir wählen daher hier den Ansatz, der dem Mantissengesetz entspricht:

Definition *a_n ist eine Benford-Folge, wenn $u_n = \langle \log a_n \rangle$ eine Weyl-Folge ist.*

Weyl-Folgen u_n wiederum sind charakterisiert durch eine der folgenden äquivalenten Bedingungen:

- u_n ist gleichverteilt auf $[0, 1]$, d. h. für alle $0 \le a < b \le 1$ gilt
$$\lim_{N\to\infty} \frac{1}{N} \big|\{0 \le n \le N-1 \,:\, a \le u_n \le b\}\big| = b - a$$
- Für alle Funktionen $f \in C([0, 1])$ gilt
$$\int_0^1 f(x)dx = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(u_n)$$
- Für alle ganzen Zahlen $\ell \neq 0$ gilt
$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} e^{2\pi i \ell u_n} = 0$$

Mit der selben Rechnung wie im Abschnitt 2 folgt dann

Satz *Für eine Benford-Folge a_n gilt*

$$\lim_{N\to\infty} \frac{1}{N} \big|\{1 \le n \le N \,:\, \textit{erste signifikante Ziffer von } a_n \textit{ ist } i\}\big| = \log(1 + \frac{1}{i}).$$

Und ebenso erhält man die entsprechende Aussage für führende Ziffernfolgen. Natürlich könnte man nun alle Folgen aus Sloane's Online Encyclopedia of integer sequences[9] daraufhin untersuchen, ob es sich um Benford-Folgen handelt oder nicht. Interessanter ist es natürlich ganze Klassen von Folgen als Benfordsch zu entlarven.

Beispiel Sei $a_n = q^n$ mit $\xi = \log q$ irrational. Dann ist

$$u_n = \langle \log a_n \rangle = \langle n\xi \rangle.$$

[9]Siehe www.research.att.com/~njas/sequences. Die Web-Seite liefert nach Eingabe einiger Folgenglieder eine Liste von bekannten Folgen, welche die gegebenen Glieder als Teilstück enthalten. Dies ist nützlich, um Folgen zu identifizieren, und um die entsprechenden IQ-Testfragen ad absurdum zu führen.

Dass dies eine Weyl-Folge ist, lässt sich elementar nachprüfen. Ganz kurz geht es aber mit dem letzten Weyl-Kriterium oben: Für $\ell \neq 0$ ganz, ist $z := e^{2\pi i \ell \xi} \neq 1$ und

$$\frac{1}{N}\sum_{n=0}^{N-1} e^{2\pi i \ell u_n} = \frac{1}{N}\sum_{n=0}^{N-1} e^{2\pi i \ell n \xi} = \frac{1}{N}\sum_{n=0}^{N-1} z^n = \frac{1}{N}\cdot\frac{z^N-1}{z-1} \to 0 \text{ für } N \to \infty$$

Das heisst, a_n ist eine Benford-Folge. Da $\log 2$ irrational ist, ist insbesondere $a_n = 2^n$ eine Benford-Folge. Hingegen ist 10^n natürlich **keine** Benford-Folge, was zeigt, dass die Irrationalitätsbedingung nicht weggelassen werden darf.

Bei Arnold [2] ist es übrigens eine Übungsaufgabe, als Folgerung des Poincaréschen Wiederkehrsatzes zu zeigen, dass $a_n = 2^n$ eine Benford-Folge ist. Natürlich lassen sich die Aussagen auch in der Sprache der Ergodentheorie formulieren.

Eine wunderschöne Verallgemeinerung des obigen Beispiels ist kürzlich in [17] gefunden worden:

Satz (Jolissaint) *Soit $p(x) = x^q - c_1 x^{q-1} - \ldots - c_{q-1}x - c_q$ un polynôme de degré q qui possède une racine $\xi > 1$ de multiplicité 1 telle que $|\eta| < \xi$ pour toute autre racine η de $p(x)$. Soit $(a_n)_{n\geq 0} \subset [1, +\infty[$ une suite satisfaisant la relation de récurrence associée $a_{n+q} - c_1 a_{n+q-1} - \ldots c_q a_n = 0$ et telle que*

$$\inf\left\{ \frac{a_n}{\xi^n} \,\middle|\, n \geq 0 \right\} > 0.$$

Si $b \geq 3$ est un entier tel que $\log_b(\xi)$ est irrationel, alors la suite (a_n) satisfait la loi de Benford par rapport à la base b, et il en est de même de toute sous-suite $(a_{Q(n)})_{n\geq 0}$ où $Q(x)$ est un polynôme non constant à coefficients entiers tel que $Q(n) \geq 0$ pour tout entier $n \geq 0$.

Aus Jolissaints Satz folgt insbesondere sofort, dass die Folge der Fibonacci-Zahlen eine Benford-Folge ist.

Diaconis bewies in [5] mit Hilfe des Weyl-Kriteriums die lange gehegte Vermutung, dass auch die Folge der Fakultäten $n!$ eine Benford-Folge ist. Nun mag man langsam daran glauben, dass Nicht-Benford-Folgen eher die Ausnahme sind. Dagegen halten kann man jedoch, dass Benford-Folgen Eigenschaften besitzen, die sich als Kriterium benutzen lassen, um gewisse Folgen als nicht-Benfordsch zu identifizieren. Darum geht es im folgenden Abschnitt.

4.1.1 Bedingungen für Benford-Folgen

Um zu verifizieren, dass eine bestimmte Folge nicht Benfordsch ist, ist oft folgender Satz nützlich:

Satz (Kuipers-Niderreiter, Diaconis [5]) *Für eine Benford-Folge a_n gilt*

$$\limsup_{n\to\infty} n \log \frac{a_{n+1}}{a_n} = \infty.$$

Daraus lässt sich leicht ableiten, dass folgende Folgen nicht Benfordsch sind

- n^b für beliebiges reelles b
- Arithmetische Folgen beliebiger Ordnung
- $\log_b n$ für beliebiges reelles b
- Primzahlfolge p_n
- $\log_b p_n$

Vor allem bei arithmetischen Folgen stutzt man zunächst einen Moment, denn sie werden ja auch durch lineare Rekursionsgleichungen beschrieben. Der Satz von Jollisaint greift jedoch nicht, denn das entsprechende charakteristische Polynom besitzt keine dominante Wurzel. Andererseits hatten wir ja bereits beim Beispiel mit den Autonummern gesehen, dass für die ganz einfache arithmetische Folge $a_n = n$ die Dichte der Glieder, die mit 1 beginnt, keinen Limes besitzt. Nun kann man bekanntermassen den Limesbegriff auf nicht-konvergente Folgen und Reihen ausdehnen (siehe etwa das Standardwerk von Hardy [12] über divergente Reihen). Die entsprechenden Ergebnisse werden als Abel-Tauber-Theorie angesprochen. Wir stellen im Folgenden kurz dar, was die entsprechende Theorie im Zusammenhang mit dem Benfordschen Gesetz liefert.

4.1.2 Verallgemeinerte Limites von Dichten

Wir betrachten zunächst die arithmetische Folge $a_n = n$, und folgen dabei Knuth [18]. Für natürliche Zahlen n und ein festes reelles $r \in]1, 10]$ sei

$$P_0(n) \;:=\; \begin{cases} 1 & \text{falls } \langle \log n \rangle < \log r \\ 0 & \text{sonst.} \end{cases}$$

$P_0(n)$ liefert also eine 1, wenn die Ziffernfolge von n lexikographisch kleiner ist als diejenige von r. Anders gesagt, ist beispielsweise $r = 4$, so ist $P_0(n) = 1$ falls n mit Ziffer 1, 2 oder 3 beginnt. Die entsprechende Dichte (d. h. der Anteil an natürlichen Zahlen in $[1, n]$, die mit einer lexikographisch kleineren Ziffernfolge als diejenige von r beginnen) ist dann

$$P_1(n) \;:=\; \frac{1}{n} \sum_{k=1}^{n} P_0(k).$$

Wie wir uns beim Beispiel mit den Autonummern im Abschnitt 3.2 überlegt haben, existiert der Limes der Folge P_1 nicht. Der Übergang von der divergenten Folge P_0 zur Folge P_1 heisst Cesàro-Summation[10]. Da auch P_1 divergiert, kann man versuchen, die Cesàro-Summation zu iterieren:

$$P_{m+1}(n) \;:=\; \frac{1}{n} \sum_{k=1}^{n} P_m(k).$$

Es zeigt sich jedoch, dass keine der Folgen P_m einen Limes besitzt. Aber es existiert für $1 \leq s \leq 10$

$$\lim_{n\to\infty} P_m(10^n s) =: S_m(s).$$

Aber eben, keine der Funktionen $S_m(s)$ ist konstant. Allerdings zeigt es sich, dass sie sich mit wachsendem m immer mehr dem Benford-Wert $\log r$ nähern: Knuth zeigte in [18], dass in der Tat

$$S_m \to \log r \quad \text{gleichmässig.}$$

Statt konsekutive Cesàro-Mittel zu betrachten, hat es sich als fruchtbar erwiesen, die sogenannte **harmonische Dichte** (gelegentlich auch **logarithmische Dichte**) zu betrachten. Dabei geht bei der Mittelwertbildung ein Folgenglied a_n nicht wie üblich mit dem Gewicht 1 in die Berechnung ein, sondern mit dem Gewicht $\frac{1}{a_n}$. Allgemein ist also die harmonische Dichte einer Menge $A \subset \mathbb{N}$ definiert als

$$\delta(A) := \lim_{n\to\infty} \frac{\sum_{k\in A, k\leq n} \frac{1}{k}}{S(n)}\,, \tag{2}$$

wobei $S(n) := \sum_{k=1}^{n} \frac{1}{k}$, falls der Limes existiert. Aufgrund der Beziehung

$$\gamma := \lim_{n\to\infty} S(n) - \log n = 0.5772156\ldots \text{ (Euler Konstante)}$$

[10]Der divergenten Folge $a_n = (-1)^n + 1$ wird durch die Cesàro-Summation eine Folge zugeordnet, die gegen 1 konvergiert. Die Methode wird etwa bei den Fejérschen Mitteln in der Theorie der Fourier-Reihen verwendet.

kann man im Nenner in (2) genauso gut $\log n$ statt $S(n)$ schreiben (daher der Name "logarithmische Dichte").

Für die Menge E_Z der mit Ziffernfolge Z beginnenden natürlichen Zahlen (siehe (1)) hat Duncan in [10] gezeigt, dass ihre harmonische Dichte tatsächlich dem Benford-Wert entspricht:

$$\delta(E_Z) = \log(1 + \frac{1}{Z}).$$

Will man mit diesem Hilfsmittel allgemeine Folgen behandeln, so muss man die **relative harmonische Dichte** einführen: Ist $A \subset B \subset \mathbb{N}$, so ist die relative harmonische Dichte von A in B gegeben durch

$$\delta(A, B) := \lim_{n\to\infty} \frac{\sum_{k\in A, k\leq n} \frac{1}{k}}{\sum_{k\in B, k\leq n} \frac{1}{k}} = \frac{\delta(A)}{\delta(B)} \tag{3}$$

falls der Limes existiert. Man kann dann etwa nach der relativen harmonischen Dichte der Primzahlen, die mit der Ziffernfolge Z beginnen, innerhalb der Menge aller Primzahlen P fragen. Whitney tat genau das in [26] und fand wieder den Benford-Wert:

$$\delta(P \cap E_Z, P) = \log(1 + \frac{1}{Z}).$$

In der Literatur noch nicht behandelt worden ist zum Beispiel die Frage, wie sich allgemeine arithmetische Folgen (höherer Ordnung) verhalten, wenn man die relative harmonische Dichte der Glieder, die mit Ziffernfolge Z beginnen, betrachtet.

Der Vollständigkeit halber sei darauf hingewiesen, dass man mit Arrays von Zahlen genau dieselben Fragen stellen kann, wie mit Folgen: So kann man beispielsweise untersuchen, ob sich die Zahlen des Pascalschen Dreiecks Benfordsch verhalten. Solche Fragen werden in [5] behandelt.

4.2 Der Stochastische Ansatz

Dieser vielleicht natürlichste Ansatz, das Benfordsche Gesetz zu formulieren und zu untersuchen, wurde erstaunlicherweise erst ab 1995 durch eine entsprechende Frage Hills populär (siehe [14]).

4.2.1 Benfordsche Zufallsvariablen

Die Grunddefinition ist ganz natürlich:

Definition *Eine positive Zufallsvariable X heisst Benfordsch, wenn*

$$P(X \in E_Z) = \log(1 + \frac{1}{Z}).$$

Je nach Autor wird diese Gleichheit nur für die Ziffernfolge $Z = \{1, 2, \ldots, 9\}$ verlangt, oder aber für beliebige Anfangsziffernfolgen $Z \in \mathbb{N}$. Im ersten Fall spricht man gelegentlich von einer **schwachen** Benfordschen Zufallsvariable. Im zweiten Fall lässt sich die Definition wieder äquivalent umschreiben: X ist Benfordsch, wenn $\langle \log X \rangle$ auf $[0, 1]$ gleichverteilt ist.

Beispiel Ist eine exponentiell verteilte Zufallsvariable schwach Benfordsch?

Die Antwort kommt von Engel und Leuenberger [11]: Eine exponentiell mit Parameter $\lambda > 0$ verteilte Zufallsvariable hat die Dichte $f(t) = \lambda e^{-\lambda t}$. Man erhält für die Ziffern $d = \{1, 2, \ldots, 9\}$

$$g_d(\lambda) \;:=\; P(X \in E_d) \;=\; \sum_{k\in\mathbb{Z}} e^{-\lambda d 10^k}(1 - e^{\lambda 10^k}).$$

Es gilt offenbar $g_d(\lambda) = g_d(10\lambda)$. Man setzt daher naheliegender Weise

$$h_d(x) \;:=\; g_d(10^x).$$

Ein Plot der Funktionen h_d enthüllt, dass X zwar nicht schwach Benfordsch ist (und zwar für kein λ), aber auch, dass die Abweichungen zu den entsprechenden Benford-Werten nicht sehr gross sind (siehe Abbildung 14). Die Werte der Funktionen h_d oszillieren mit einer Abweichung von weniger als

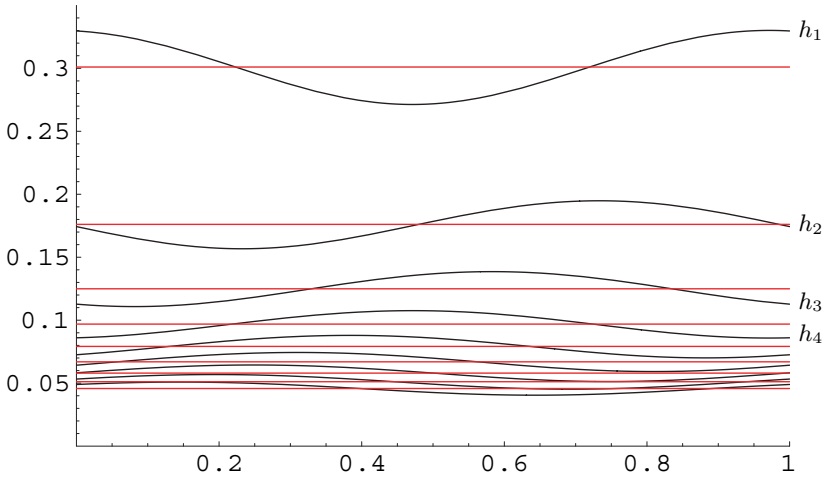


Abbildung 14: Die Funktionen h_d oszillieren um die entsprechenden Benford-Werte.

0.03 um die entsprechenden Benford-Werte. Engel und Leuenberger zeigten darüberhinaus, dass die Mittelwerte der Funktionen h_d, das heisst $\int_0^1 h_d(x)dx$, tatsächlich **genau** mit den entsprechenden Benford-Werten übereinstimmen.

Es zeigte sich, dass der oben definierte Begriff der Benford-Zufallsvariable zu eng gefasst ist, um wirklich nützlich zu sein, und um das Auftreten des Benfordschen Gesetzes in der realen Welt zu erklären. Hingegen hat sich die Betrachtung von Folgen von Zufallsvariablen als geeignet erwiesen.

4.2.2 Folgen von Zufallsvariablen

Die folgende Definition stammt von Duembgen und Leuenberger [9]:

Definition *Eine Folge X_n von positiven Zufallsvariablen heisst Benfordsch, wenn*

$$\lim_{n\to\infty} P(X_n \in E_Z) = \log(1 + \frac{1}{Z}).$$

Duembgen und Leuenberger haben dann folgendes gezeigt:

Satz (Duembgen & Leuenberger [9]) *Let $X_n > 0$ be a sequence of (non-degenerate) independent identically distributed random variables such that $\log X_n$ is not of lattice type with rational span*[11]*. Then*

$$P_n = \prod_{k=1}^{n} X_n$$

is a Benford sequence.

Wenn man davon ausgeht, dass reale Datensätze oftmals als Produkt vieler Einzelfaktoren zustandekommen, erklärt dieser Satz bis zu einem gewissen Grad die beobachteten Benford-Verteilungen. Eine weitere Aussage von Duembgen und Leuenberger in derselben Richtung lautet:

[11] Das heisst, das Mass besteht nicht aus Atomen, die auf einem Gitter mit rationaler Spannweite sitzen.

Satz (Duembgen & Leuenberger [9]) *Let $X_n > 0$ be a sequence of random variables with densities f_n. Suppose that the densities g_n of $\log X_n$ satisfy*

$$\lim_{n\to\infty} \mathrm{TV}(g_n) \;=\; 0 \quad {}^{12}$$

for $n \to \infty$. Then X_n is a Benford sequence.

Mit Hilfe dieses Satzes erklärt sich nun das eingangs verwendete Beispiel im Abschnitt 2 mit den Weibull-verteilten Zufallszahlen: Die Weibull-Dichte mit Parameter $\gamma > 0$ ist gegeben durch $f_\gamma(x) = \gamma x^{\gamma-1}\exp(-x^\gamma)$. Man findet $g_\gamma(y) = \gamma 10^{\gamma y}\exp(-10^{\gamma y})\ln 10$. Diese Verteilung ist unimodal mit Maximum in 0, also

$$\mathrm{TV}(g_\gamma) \;=\; \frac{2\gamma \ln 10}{e} \;\to\; 0 \quad \text{für } \gamma \to 0.$$

Somit liegt eine Benford-Folge vor, und für genügend kleines γ ist somit eine Weibull-verteilte Zufallsvariable fast Benfordsch.

Das nächste Beispiel von Duembgen und Leuenberger ist besonders im Hinblick auf die Anwendungen des Benfordschen Gesetzes interessant:

Beispiel (Duembgen & Leuenberger [9]) Eine Pareto-Verteilung mit Parameter $\lambda > 0$ hat die Dichte $f(x) = \lambda x^{-\lambda-1}$ auf $[1, \infty[$. Hier liefert eine direkte Berechnung

$$P(X \in E_d) \;=\; (d^{-\lambda} - (d+1)^{-\lambda})\frac{10^\lambda}{10^\lambda - 1} \tag{4}$$

$$\;=\; \log(1+\frac{1}{d})(1+\lambda\ln 10) + o(\lambda). \tag{5}$$

Somit ist aus (5) ersichtlich, dass auch hier für $\lambda \to 0$ eine Benford-Folge von Zufallsvariablen vorliegt. Allerdings können reale Pareto-Verteilungen natürlich mit einem grossen Parameter λ auftreten, womit zwar keine Benford-Verteilung mehr erkennbar ist, wo jedoch dank der expliziten Formel (4) trotzdem die Verteilung der führenden Ziffern exakt vorausgesagt werden kann. Duembgen und Leuenberger illustrieren dies anhand eines in der Statistik berühmten Datensatzes:

Anwendung Die *Danish fire insurance data* sind eine Sammlung von 2167 Schadensfällen von über einer Million Kronen (Werte 1985). Man hat empirisch festgestellt, dass diese Daten gut zu einer Pareto-Verteilung passen. Der Parameter λ kann dann mit einem gängigen Schätzer ermittelt werden. Die Abbildung 15, links, zeigt die gemäss (4) daraus resultierende Verteilung der führeden Ziffern und die hervorragende Übereinstimmung mit dem realen Datensatz. Rechts daneben ist zum Vergleich auch noch die Benford-Verteilung angegeben, die hier deutlich abweicht.

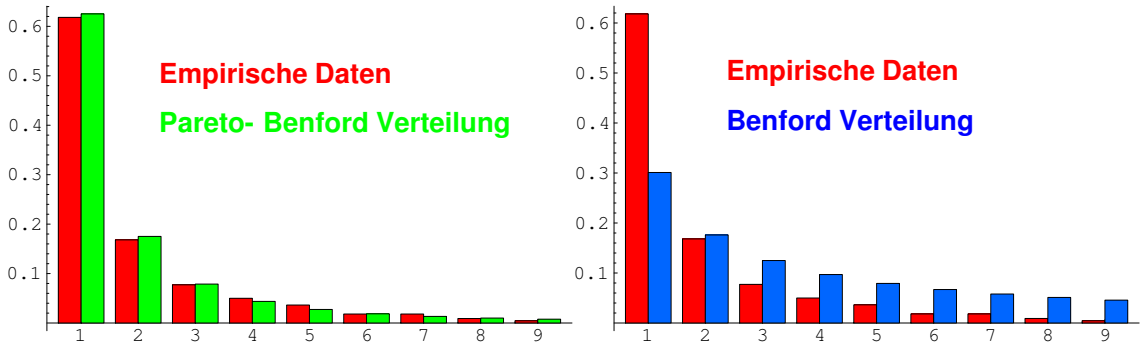


Abbildung 15: Verteilung nach führenden Ziffern in den Danish fire insurance data.

[12] TV bezeichnet hier die Totalvariation.

4.2.3 Zufällige Mischung von Wahrscheinlichkeitsverteilungen

Wir kommen nun zurück auf Benfords Beobachtung, dass die Mischung von verschiedenen Datensätzen sich besonders gut an die Benford-Verteilung hält (siehe Abschnitt 3.2). Dieses Phänomen konnte erst vor kurzem geklärt werden. Hill [14] bewies nämlich 1995 grob gesagt folgendes:

> *"Wählt man zufällig Wahrscheinlichkeitsverteilungen und dann entsprechend verteilte Daten so, dass der Gesamtprozess skalen-neutral ist, so folgt die resultierende Verteilung dem Benfordschen Gesetz."*

Der technische Rahmen ist etwas zu aufwändig, um hier im einzelnen dargestellt zu werden. Es muss unter anderem eine Wahrscheinlichkeitsverteilung auf einem Raum von Wahrscheinlichkeitsverteilungen eingeführt werden. Für die Details verweisen wir auf die Originalarbeit von Hill [14].

4.2.4 Statistische Tests

Ist die Verteilung einer Zufallsvariable X bekannt, so lässt sich daraus natürlich die Verteilung von $\langle \log X \rangle$ berechnen (siehe die Beispiele in den Abschnitten 4.2.1 und 4.2.2). Somit lässt sich dann überprüfen, ob $\langle \log X \rangle$ gleichverteilt auf $[0, 1]$ ist, das heisst, ob X dem starken Benfordschen Gesetz gehorcht. Entsprechend lassen sich die Wahrscheinlichkeiten $P(X \in E_i)$ für $i = 1, \ldots, 9$ berechnen und man kann damit verifizieren, ob Benford's first significant digit law erfüllt ist.

Bei einem realen Datensatz oder einer Stichprobe $X = \{X_1, \ldots, X_n\}$ lässt sich die empirische relative Häufigkeit $n_i = \frac{N_i}{n}$ berechnen, wobei N_i die Häufigkeit der führenden Ziffer i in der Stichprobe bezeichnet. Die empirischen Werte n_i können dann mit den Benford-Werten $\log(1 + \frac{1}{i})$ verglichen werden. An dieser Stelle bietet sich der Chi-Quadrat-Test an: Dieser Test erlaubt es, die Nullhypothese zu testen, dass der Stichprobe die Benford-Verteilung zugrunde liegt. Dazu betrachtet man

$$\sum_{i=1}^{9} \frac{(N_i - n \log(1 + \frac{1}{i}))^2}{n \log(1 + \frac{1}{i})} \,. \tag{6}$$

Stimmt die Nullhypothese, ist diese Grösse approximativ Chi-Quadrat-verteilt mit 8 Freiheitsgraden. Dies lässt sich dann für eine gegebene Irrtumswahrscheinlichkeit testen. Entsprechend kann man den Test auch für $k > 1$ Anfangsziffern durchführen. Dabei ist darauf zu achten, dass die Stichprobe genügend gross ist. Eine Faustregel sagt, dass auch in der kleinsten Klasse die zu erwartende Anzahl grösser gleich vier sein sollte. Will man etwa auf eine Ziffer testen, sollte $n \geq \frac{4}{\log(1+\frac{1}{9})} = 87.4\ldots$ sein. Wird die Nullhypothese verworfen, kann man die einzelnen Summanden in (6) untersuchen: Summanden, die grösser oder gleich zwei sind, bedeuten, dass die entsprechende Ziffer (im Vergleich zu Benford) zu oft oder zu selten vorkommt.

Wie wir gesehen haben, ergeben die klassischen Verteilungen nur angenäherte Benford-Verteilungen. Daher wird ein Chi-Quadrat-Test in solchen Fällen richtigerweise anzeigen, dass keine Benford-Verteilung vorliegt, und zwar umso eindeutiger, je grösser die Stichprobe ist, selbst wenn man nahe an einer Benford-Verteilung ist. Damit relativiert sich die Nützlichkeit derartiger Tests.

Als Alternative zum Chi-Quadrat-Test wird gelegentlich der Kolmogorow-Smirnow-Test verwendet.

Als Quintessenz kommen wir letzten Endes zu einem ähnlichen Schluss wie bei der Normalverteilung: Obwohl es eine theoretische Rechtfertigung durch Grenzwertsätze und durch empirische Untersuchungen gibt, findet man leicht Fälle, wo die postulierte Verteilung (sei dies die Benford- oder die Normalverteilung) offensichtlich nicht stimmt, und wenn man genauer hinschaut, findet man fast überall gewisse Abweichungen.

4.3 Strukturelle Aspekte

Wir hatten im Abschnitt 2.2 bereits festgestellt, dass eine Benford-Zufallsvariable gezwungenermassen skaleninvariant sein muss. Es gilt aber auch die Umkehrung:

Satz *X ist dann und nur dann eine Benfordsche Zufallsvariable, wenn X skaleninvariant ist, das heisst, wenn $\langle \log(\lambda X) \rangle$ eine von $\lambda > 0$ unabhängige Verteilung hat.*

Beweis Sei X eine positive Zufallsvariable. Dann gilt offenbar folgendes: $\langle \log(\lambda X) \rangle$ hat eine von $\lambda > 0$ unabhängige Verteilung F auf $[0, 1]$ dann und nur dann, wenn F die Gleichverteilung ist. Dies charakterisiert gerade eine (starke) Benford-Zufallsvariable. □

Wir hatten im Abschnitt 2.2 schon angedeutet, dass das Benfordsche Gesetz für beliebige Basen formuliert werden kann. Die genaue Definition einer **Baseninvarianz** ist jedoch subtiler als die der Skaleninvarianz. Hill ging dieser Frage im Detail nach (siehe [13] und [15]). Er zeigte darin, dass Baseninvarianz ebenfalls das Benfordsche Gesetz impliziert.

5 Anwendungen von Benfords Gesetz

Obwohl das Benfordsche Gesetz eine verhältnismässig einfache Beobachtung über unser Stellenwertsystem darstellt, hat es in jüngster Zeit überraschenderweise eine Reihe pfiffiger Anwendungen gefunden. Es ist damit ein anschauliches Beispiel dafür, dass, wer nur genau genug hinschaut und über die nötige Phantasie verfügt, Mathematik im Alltag nutzbringend anwenden kann.

5.1 Aufdeckung von Fälschungen

Durch Münzwurf kann man auf einfache Weise eine **echte** zufällige 0-1-Folge erzeugen. Andererseits ist es erstaunlich schwer, eine solche Folge zu **fälschen**, das heisst eine zu erfinden: Die meisten Leute, die man dazu auffordert, eine "zufällige" 0-1-Folge zu erfinden schaffen es zwar, dass Nullen und Einsen etwa mit der gleichen Häufigkeit auftreten, aber nur die wenigsten trauen sich, auch einmal längere Blöcke von Nullen oder Einsen zu schreiben. Echte Zufallsfolgen enthalten aber solche Cluster: So kommt in einer Zufallsfolge der Länge 200 mit etwa 95-prozentiger Sicherheit ein Block von Nullen oder Einsen der Länge 6 oder mehr vor.

Ähnlich verhält es sich beim Fälschen von Bilanzen, Resultaten von Medikamententestreihen, Laborbüchern, Krankenkassenabrechnungen von Praxen usw. Nigrini untersuchte in [20, 21] echte Bilanzen von amerikanischen Firmen und stellte fest, dass die Zahlen dem Benfordschen Gesetz folgen. Kaum ein Fälscher ist sich jedoch dieser Tatsache bewusst. Dies lieferte Nigrini den Ansatz, einen Test zu entwickeln, mit dem man verdächtige Bilanzen herausfiltern kann, indem systematisch und grossflächig nach Abweichungen vom Benfordschen Gesetz gefahndet wird. Die Methode wird von der US-Steuerbehörde IRS inzwischen mit Erfolg angewandt, und auch in Deutschland und in der Schweiz (siehe [24]) unternimmt man Versuche in dieser Richtung. Nigrini argumentiert, dass selbst Fälscher auffliegen, welche versuchen, Benford-Daten herzustellen, da sie nicht genau wissen, in welcher Weise und auf welche Teildaten der Test angewandt wird. Nigrini hat basierend auf seiner Idee inzwischen eine florierende Firma aufgebaut[13].

Übrigens scheint auch Benford selber in diese Falle getappt zu sein: Diaconis und Freedman [6] haben nach einer detaillierten Analyse den Verdacht geäussert, dass auch Benford in seiner Originaltabelle (siehe Abbildung 8) die eine oder andere Zahl zu seinen Gunsten gerundet hat, um das Ergebnis prägnanter erscheinen zu lassen.

Obwohl Nigrinis Idee natürlich bestechend ist, wollen wir ein paar kritische Anmerkungen nicht un-

[13] www.nigirini.com

terlassen. Bei seiner Methode können Fehler erster und zweiter Art auftreten: Ein echter Datensatz kann fälschlicherweise als manipuliert klassifiziert werden. So würden etwa ohne genauere Analyse der Situation die echten Danish fire data (siehe Abschnitt 4.2.2) als gefälscht entlarvt. Es ist aber auch möglich, dass die Daten (willentlich oder nicht) so gefälscht sind, dass sie trotzdem dem Benford-Gesetz folgen. Dieser Frage ist Andreas Diekmann, Soziologe an der ETH Zürich, nachgegangen. Er untersuchte, ob sich die Benford-Methode zur Aufdeckung von Fälschungen in wissenschaftlichen Daten eignet[14]. Er hat gefunden, dass je nach Versuchsanordnung auch gefälschte Daten Benford-verteilt sein können (obwohl die Testfälscher nichts vom Benford-Gesetz wussten). Eines von Diekmanns Resultaten ist, dass die Analyse der Verteilung der **zweiten** Ziffer erfolgversprechender ist, als die der ersten Ziffer (siehe [7] oder [8]).

Wir haben in diesem Zusammenhang ein eigenes Experiment durchgeführt: Einer Testperson wurden 500 Ortsnamen von chinesischen Ortschaften vorgelesen, zu denen sie Einwohnerzahlen erfinden sollte. Nigrini und Wood hatten festgestellt, dass Bevölkerungszahlen von US-Counties in der Tat dem Benfordschen Gesetz folgen [22]. Die 500 erfundenen Daten zeigen jedoch tatsächlich eine signifikante Abweichung von der Benford-Verteilung. Die resultierende Verteilung (siehe Abbildung 16)

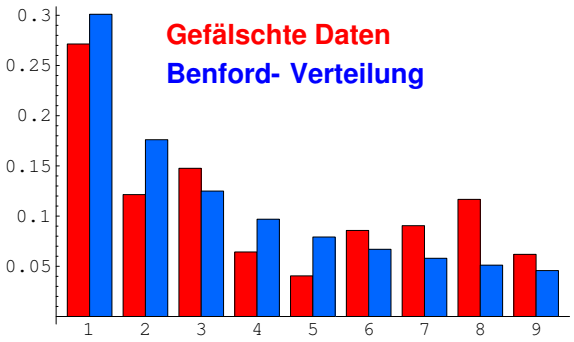


Abbildung 16: Verteilung nach führenden Ziffern von gefälschten Daten.

scheint persönliche Präferenzen für führende Ziffern widerzuspiegeln und damit Eigenschaften eines "Fingerprints" zu besitzen: Wird der Datensatz in zwei Hälften geteilt, so zeigen beide Hälften in guter Näherung das selbe Profil.

5.2 Optimierung von Algorithmen für floating point Operationen

Donald Knuth hat vorgeschlagen [18], das Benfordsche Gesetz zur Optimierung von Algorithmen von floating point Operationen im Computer heranzuziehen: Wenn man nämlich weiss, welche Ziffern häufiger als andere als führende Ziffern auftreten, so kommen bei Operationen wie Addition oder Multiplikation, gewisse Übeträge (und damit gewisse Registeroperationen) häufiger vor als andere. Durch geschickte Programmierung und Prozessor-Architektur lässt sich dieser Umstand zur Beschleunigung der Algorithmen ausnutzen. Knuth gibt ausserdem zu Bedenken, dass das Benfordsche Gesetz Auswirkungen auf die Analyse von Fehlern (durch Runden) hat.

5.3 Datenkompression

Wenn man Datensätze komprimieren möchte, lassen sich im Prinzip Strukturen jeder Art ausnutzen, um eine bessere Kompression zu erreichen. Je mehr Struktur (geringe Entropie) ein Datensatz

[14] Mehrere Betrugsfälle in der Physik machten in der jüngsten Vergangenheit Schlagzeilen (Jan Hendrik Schön, Victor Ninov).

aufweist, desto stärker lässt er sich komprimieren. Aus diesem Grund wurde vorgeschlagen, bei der Datenkompression die zusätzliche Struktur, welche das Benford-Gesetz liefert, auszunützen.

5.4 Test für Prognosemethoden

Wie oben schon angedeutet, haben Nigrini und Wodd in [22] festgestellt, dass reale Bevölkerungsdaten dem Benfordschen Gesetz folgen. Er hat daraufhin vorgeschlagen, Prognoseverfahren (zum Beispiel eben für die künftige Bevölkerungsentwicklung) daraufhin zu testen, ob sie ebenfalls Benford-verteilte Zahlen liefern. Wenn dies nicht der Fall sei, so müsse das Prognoseverfahren oder das zugrundeliegende Modell überprüft werden.

5.5 Lotto

Man kann zwar seine Gewinnchancen beim Lotto mit dem Benfordschen Gesetz nicht erhöhen, wohl aber den Gewinn, wenn er denn eintritt. Geht man davon aus, dass die meisten Leute kleine Zahlen als führende Ziffern unbewusst bevorzugen, sollte man beim Tippen selber diese Zahlen meiden, um beim Gewinn mit weniger Mitgewinnern teilen zu müssen. Bei der Verteilung der abgegebenen Tipps scheinen jedoch geometrische Muster auf dem Lottoschein eine entscheidendere Rolle zu spielen. Dies geht aus den Untersuchungen von Hans Riedwyl zur Mathematik des Zahlenlottos hervor.

6 Schlussbemerkung: Zipfs Gesetz

Neben Benfords Gesetz gibt es noch weitere (auf den ersten Blick) überraschende phänomenologische Gesetze ähnlicher Art: Als Beispiel sei hier nur noch Zipfs Gesetz erwähnt, benannt nach dem amerikanischen Linguisten George Kingsley Zipf. Es besagt folgendes: Wenn man die Zahlen eines Datensatzen der Grösse nach ordnet, sagen wir $a_1 \geq a_2 \geq \ldots$, so verhalten sich die Zahlen umgekehrt proportional zu ihrem Rang. Das heisst:

$$a_2 \approx \frac{1}{2}a_1, \quad a_3 \approx \frac{1}{3}a_1, \quad \ldots$$

Zipf "entdeckte" sein Gesetz als er die Häufigkeit des Vorkommens von Wörtern in der englischen Sprache untersuchte. Daneben existieren modifizierte Zipf-Gesetze (zum Beispiel von Mandelbrot) etwa der Form $a_n \approx \frac{a_1}{(n+c)^\alpha}$. In der Literatur wird üblicherweise sofort darauf hingewiesen, dass Zipfs Gesetz nichts anderes als eine spezielle Pareto-Verteilung ist (siehe zum Beispiel die Arbeit von Richard Perline [23]).

Dank

Ich danke Christoph Leuenberger, Daniel Stoffer und Hansruedi Künsch für wichtige Hinweise und Kommentare.

Literatur

[1] Douglas Adams: The hitchhiker's guide to the galaxy: a trilogy in four parts. London: Picador, 2002

[2] Vladimir I. Arnol'd: Mathematical methods of classical mechanics. New York: Springer, 1989

[3] Frank Benford: The law of anomalous numbers. Proc. Amer. Philos. Soc. 78, 551–572 (1938)

[4] Klaus Burde: Das Problem der Abzählreime und Zahlentwicklungen mit gebrochenen Basen. J. Number Theory 26, 192–209 (1987)

[5] Persi Diaconis: The distribution of leading digits and uniform distribution mod 1. Ann. Probab. 5, 72–81 (1977)

[6] Persi Diaconis, David Freedman: On Rounding Percentages. J. of the Amer. Statistical Association 74, 359–364 (1979)

[7] Andreas Diekmann: Datenfälschung. Ergebnisse aus Experimenten mit der Benford Verteilung. Manuscript, ETH Zürich, 2004

[8] Andreas Diekmann: Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data. Manuscript, ETH Zürich, 2004

[9] Lutz Duembgen, Christoph Leuenberger: Benford's law for random variables. Preprint

[10] R. L. Duncan: Note on the initial digit problem. Fibonacci Q. 7, 474–475 (1969)

[11] Hans-Andreas Engel, Christoph Leuenberger: Benford's law for exponential random variables. Stat. Probab. Lett. 63, No. 4, 361–365 (2003)

[12] Geoffrey H. Hardy: Divergent series. Sceaux: Gabay, 1992

[13] Theodore P. Hill: The significant-digit phenomenon. Am. Math. Mon. 102, No. 4, 322–327 (1995)

[14] Theodore P. Hill: A statistical derivation of the significant-digit law. Statistical Science 10/4, 354–363 (1995)

[15] Theodore P. Hill: Base-invariance implies Benford's law. Proc. Am. Math. Soc. 123, No. 3, 887–895 (1995)

[16] Thomas Jech: The logarithmic distribution of leading digits and finitely additive measures. Discrete Math. 108, No. 1–3, 53–57 (1992)

[17] Paul Jolissaint: Loi de Benford, relations de récurrence et suites équidistribuées. Elem. Math. 60, No. 1, 10–18 (2005)

[18] Donald E. Knuth: The art of computer programming. Reading, Massachusetts: Addison-Wesely, 1997

[19] Simon Newcomb: Note on the frequency of use of the different digits in natural numbers. Amer. J. Math. 4, 39–41 (1881)

[20] Mark J. Nigrini: The detection of income evasion through an analysis of digital distributions. Ph.D. thesis, Dept. of Accounting, Univ. Cincinnati, Cincinnati OH, 1992

[21] Mark J. Nigrini: A taxpayer compliance application of Benford's Law. J. of the Am. Taxation Assoc. 18, 72–91 (1996)

[22] Mark J. Nigrini, W. Wood: Assessing the integrity of tabulated demographic data. Preprint, University of Cincinnati and St. Mary's University

[23] Richard Perline: Strong, weak and false inverse power laws. Statist. Sci. 20, No. 1, 68–88 (2005)

[24] Reto U. Schneider: Das Rätsel der abgegriffenen Seiten. NZZ Folio 1/06. http://www-x.nzz.ch/folio/curr/articles/schneider_2.html

[25] Erwin Voellmy: Fünfstellige Logarithmen und Zahlentafeln für die 90°-Teilung des rechten Winkels. Zürich: Orell Füssli, 1970

[26] R. E. Whitney: Initial digits for the sequence of primes. Amer. Math. Monthly 79, 150–152 (1972)

[27] Grosse Konkordanz zur Elberfelder Bibel (revidierte Fassung): Wort- und Zahlenkonkordanz. Wuppertal, Zürich: Brockhaus, 1993