Beispiele und Schüleraktivitäten zum BENFORD-Gesetz[1]

JONATHAN R. BRADLEY AND DAVID L. FARNSWORT, NEW YORK

[1] aus: *Teaching Statistics* 31(2009), 1, S. 2-6
Übersetzung: KATJA KRÜGER, FRANKFURT

Zusammenfassung: *Die überraschende Eigenschaft einer Reihe von Datensätzen, deren Anfangsziffern dem sogenannten BENFORD-Gesetz genügen, liefert Beispiele, die nicht nur Interesse bei Schülern wecken, sondern auch aufrechterhalten können. Es werden einige Ideen für Schüleraktivitäten vorgestellt.*

1 Einleitung

Es gibt eine erstaunliche und zugleich der Intuition entgegensprechende Eigenschaft, die vielen Datensätzen gemeinsam ist, das sogenannte BENFORD-Gesetz. Es besagt, dass die relativen Häufigkeiten der Anfangsziffern[1] einer logarithmischen Gesetzmäßigkeit folgen. Für die Häufigkeitsverteilung der Anfangsziffern i = 1, 2, …, 9 gilt nach dem BENFORD-Gesetz (vgl. Tabelle 1)

$$\log_{10}\left(1+\frac{1}{i}\right) \qquad (1)$$

Das ist insofern erstaunlich, weil man wohl eher vermuten würde, dass die Ziffern 1 bis 9 nahezu gleich häufig als Anfangsziffern vorkommen und deren relative Häufigkeiten somit ungefähr 1/9 betragen sollten. Da diese Eigenschaft weitgehend unbekannt ist, kann sie sehr gut im Stochastikunterricht verwendet werden. Schülern kann diese Gesetzmäßigkeit einfach am Beispiel echter oder künstlich erzeugter Daten vermittelt werden.

Entdeckt wurde dieses Phänomen bereits 1881, als Simon Newcomb bemerkte, dass Logarithmentafeln auf den ersten Seiten stärker abgenutzt waren als auf den folgenden. Der Arzt Frank Benford entdeckte diese Gesetzmäßigkeit 1938 am gleichen Phänomen wieder. Er gab weitere 20 Beispiele für die nach ihm benannte Gesetzmäßigkeit an, für die er mehr als 20 000 Beobachtungen auswertete (vgl. Benford 1938). Darunter waren so verschiedene Beispiele wie Baseball-Statistiken, Hausnummern aus dem biografischen Verzeichnis des „*American Men of Science*" oder Oberflächen von Seen. Benford untersuchte auch Daten wie beispielsweise Molekulargewichte, spezifische Wärmen und Atomgewichte und fand heraus, dass diese nicht der nach ihm benannten logarithmischen Gesetzmäßigkeit genügen.

Das BENFORD-Gesetz ist ein *empirisches* Gesetz in dem Sinn, dass die Häufigkeiten der Anfangsziffern vieler natürlich gewonnener Daten mit der in (1) angegebenen Verteilung näherungsweise übereinstimmen. Bestimmte den Daten zugrunde liegende Strukturen, wie z. B. exponentielles Wachstum, können Zahlen erzeugen, deren Anfangsziffern der BENFORD-Verteilung genügen.

2 Ein Beispiel

Eine schnell zugängliche Datenquelle stellt heutzutage das Internet dar. Offizielle Statistiken, beispielsweise von Regierungsseiten, liegen dabei häufig in einem Format vor, das einfach zu kopieren ist, und sind darüber hinaus hervorragend dokumentiert. Als Beispiel für eine solche Datenquelle kann die folgende von dem *US Department of Labor's Bureau of Labor Statistics* gepflegte Seite dienen: *http://www.bls.gov/oes/2008/may/oes_ny.htm#b00-0000.* Auf dieser Seite werden Daten einer Vielzahl von Berufen in Tabellenform zugänglich gemacht. So sind hier die Beschäftigtenzahlen von insgesamt 694 verschiedenen Beschäftigungsarten im Staat New York aus dem

Anfangsziffer	1	2	3	4	5	6	7	8	9
BENFORD-Häufigkeiten	0,3010	0,1761	0,1249	0,0969	0,0792	0,0669	0,0580	0,0512	0,0458

Tab. 1: Relative Häufigkeit der Anfangsziffern nach dem BENFORD-Gesetz

Anfangsziffer	1	2	3	4	5	6	7	8	9
Beobachtete Häufigkeit Q_k	233	102	96	62	54	38	43	30	36
Erwartete Häufigkeit E_k	208,9	122,2	86,7	67,3	55,0	46,5	40,2	35,5	31,8

Tab. 2: Absolute Häufigkeiten der Anfangsziffern von Beschäftigtenzahlen im Staat New York im Vergleich mit den erwarteten Häufigkeiten nach der BENFORD-Verteilung

Jahr 2006 (Monat Mai) erfasst. Es gab beispielsweise 14 200 *„Marketing Manager“* oder 21 870 *„Computer and Information Manager“* usw. Wertet man die Anfangsziffern dieser 694 verschiedenen Beschäftigtenzahlen aus, so stimmen die beobachteten absoluten Häufigkeiten recht gut mit den erwarteten BENFORD-Häufigkeiten überein (vgl. Tabelle 2).

Mithilfe des χ^2-Anpassungstests lässt sich untersuchen, wie gut die beobachtete Häufigkeitsverteilung der Anfangsziffern mit der BENFORD-Verteilung übereinstimmt. Die χ^2-Prüfgröße

$$\sum_{k=1}^{9}(Q_k - E_k)^2 / E_k,$$

mit den beobachteten absoluten Häufigkeiten Q_k (in Zeile 2) und den erwarteten BENFORD-Häufigkeiten E_k (in Zeile 3) hat acht Freiheitsgerade und liefert für die vorliegende Häufigkeitsverteilung den Wert 10,69 mit einem p-Wert von 0,22. Die Verteilung der Anfangsziffern der Beschäftigungsdaten ist mit der BENFORD-Verteilung verträglich.

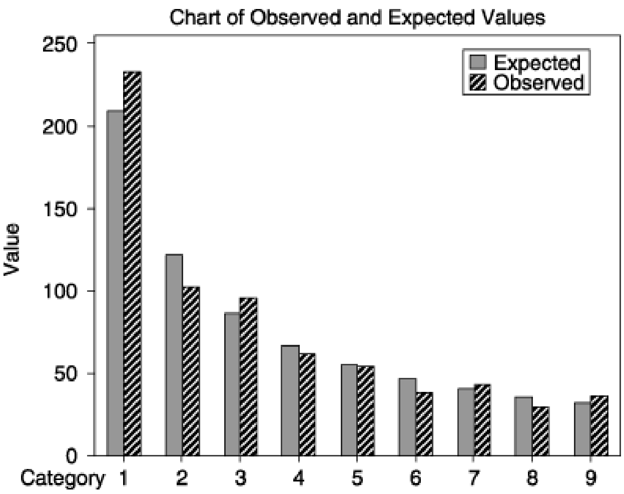

Abb. 1: Erwartete und beobachtete Häufigkeiten der Beschäftigungsdaten

3 Anwendungen

Wissenschaftler aus ganz verschiedenen Gebieten haben zahlreiche empirische Belege für das Auftreten des Phänomens des BENFORD-Gesetzes gefunden. Nicht nur Steuerdaten von Finanzämtern (Nigrini 1996) und Aktienindizes (Ley 1996) folgen dieser Gesetzmäßigkeit auch Preise bei eBay-Auktionen (Giles 2007).

Betrugsversuche aufzudecken ist die wohl wichtigste aktuelle Anwendung des BENFORD-Gesetzes z. B. im Rahmen von Rechnungs- und Wirtschaftsprüfungen (Kumar und Bhattacharya 2007). Besonders auf diese Möglichkeit sollten Schüler aufmerksam gemacht werden. Wenn wir einen Datensatz vorliegen haben, von dem wir erwarten dürfen, dass die darin enthaltenen Anfangsziffern der BENFORD-Verteilung genügen, können wir diesen auf mögliche Manipulationsversuche überprüfen. Zum Beispiel könnte es Absprachen zwischen Bietern bei eBay geben, sodass die Verkaufspreise nicht mehr dem BENFORD-Gesetz folgen. Die Suche nach Betrugsversuchen war auch der Schwerpunkt der Untersuchungen von Marchi und Hamilton (2006) über die Evaluation selbst erhobener Daten zur Freisetzung toxischer Chemikalien in der Umwelt. Swanson u. a. (2003) fanden heraus, dass viele Datensätze, die vom *United States Census Bureau* erhoben wurden, dem BENFORD-Gesetz genügen. Auch hier lässt sich diese Gesetzmäßigkeit anwenden, um gefälschte Umfragedaten zu entdecken, die beispielsweise durch die Praxis des *„curbstoning“* erzeugt wurden. Dabei werden Daten einfach frei erfunden, als würde man – bildlich gesprochen – im Rinnstein sitzen (*„sitting at the curb*“) anstelle Interviews durchzuführen.

4 Erklärungsmodelle

Es gibt verschiedene mathematische Modelle, die das BENFORD-Gesetz erklären. Da es sich etwas schwieriger gestaltet, in einer elementaren Weise Bedingungen für das Auftreten dieser Gesetzmäßigkeit anzugeben, beschränken wir uns im Folgenden darauf, Situationen zu skizzieren, in denen das Auftreten des BENFORD-Gesetzes allgemein bekannt ist.[2]

Zum Beispiel genügen Potenzen von Zufallszahlen, die auf (0,1) gleichverteilt sind, mit zunehmend höheren Exponenten dem BENFORD-Gesetz (Adhikari und Sarkar 1968). Ebenso weisen periodische Stichproben von geometrischen Prozessen wie z. B. $X_k = 3^k$ die typische BENFORD-Verteilung der Anfangsziffern auf (Diaconis 1977).

Wenn die Verteilung einer Zufallsgröße invariant gegenüber Skalenänderungen ist (z. B. beim Wechsel der Längeneinheit Meilen in Kilometer), genügen die beobachteten Werte (unter gewissen Einschränkungen) der BENFORD-Verteilung (Hill 1995 a, b). Diese Eigenschaft der Skaleninvarianz, wie sie auch bei Naturgesetzen vorkommt, stellt sicher, dass die Messskala nicht diktiert, ob das BENFORD-Gesetz in einem Datensatz auftritt oder nicht. Die Skaleninvarianz sorgt, nebenbei bemerkt, auch dafür, dass periodische Stichproben von z. B. $4(3^k)$ dem BENFORD-Gesetz genügen.

5 Aktivitäten

Wir möchten nun Schüleraktivitäten anregen, die selbstständige Erfahrungen „aus erster Hand“ mit dem BENFORD-Gesetz ermöglichen. Schüler sollten dazu ihre eigenen Daten sammeln, damit ein Gefühl von Eigenständigkeit sowie ein größeres Interesse an ihren eigenen Untersuchungsergebnissen geweckt werden.

Naturwissenschaftlich interessierte Schüler könnten eine Sammlung physikalischer Konstanten, wie sie beispielsweise in Schulbüchern vorkommen, näher untersuchen. Dabei zeigt sich, dass rund 30 % dieser Konstanten als erste signifikante Ziffer die 1 besitzen.

Die Anwendung des BENFORD-Gesetzes bei der Entdeckung von Betrug könnte als Einstiegsbeispiel dienen, wenn man im Unterricht in den χ^2-Anpassungstest einführen möchte. Bereits in der beschreibenden Statistik könnte ein solches Beispiel viel elementarer beim Vergleich von Verteilungen verwendet werden. Auch Projektarbeit wäre hier denkbar: Schüler suchen selbstständig nach Datensätzen[3] und überprüfen mithilfe eines Verteilungsvergleichs, eines Hypothesen- oder χ^2-Anpassungstests, ob die Häufigkeiten der Anfangsziffern in ihrem Datensatz mit denen der BENFORD-Verteilung übereinstimmen. Besonders interessant wäre es wohl für Schüler, hier mit Daten zu arbeiten, deren Anfangsziffern aufgrund von Fälschungsversuchen von der BENFORD-Verteilung abweichen.

Eine Alternative zur aufwendigeren Suche nach „echten“ Daten bietet die Möglichkeit, sich mithilfe eines Tabellenkalkulationsprogramms Daten künstlich selbst zu erzeugen und direkt auszuwerten. Die Tabellenkalkulation könnte dazu dienen, nicht nur das BENFORD-Gesetz, sondern auch dessen Invarianzeigenschaft sowie einige Excel-Funktionen zu demonstrieren. Zum Beispiel könnten Schüler mit Programmunterstützung die Potenzen von 3^k für $k = 1, \ldots, n$ für eine (genügend große) natürliche Zahl n berechnen und anschließend die relativen Häufigkeiten der führenden Ziffern ermitteln. Deren Verteilung liegt überraschend dicht an der BENFORD-Verteilung (vgl. Tabelle 3). Die Invarianz-Eigenschaft kann untersucht werden, indem beispielsweise alle Potenzen zur Basis 3 mit einem beliebigen konstanten Faktor multipliziert werden. In Tabelle 3 entsprechen die Anfangsziffern der Folgenglieder von $4 \cdot 3^k$ der BENFORD-Verteilung besonders gut.

In Abbildung 2 sind die wesentlichen Schritte angegeben, um mit Excel die Verteilung der Anfangsziffern für 100 Glieder der Folge 3^k zu bestimmen. Möchte man die anderen in Tabelle 3 gezeigten Häufigkeitsverteilungen ermitteln, so sind nur geringfügige Änderungen notwendig.

Im Anschluss an dieses Einführungsbeispiel könnten Schüler eigenständig Folgen eingeben und untersuchen, ob bei deren Folgengliedern die Anfangsziffern der BENFORD-Verteilung genügen. Wie bereits erwähnt, ist dies bei den meisten geometrischen Folgen der Fall. Sollten Schüler Folgen finden, die nicht dem BENFORD-Gesetz genügen, könnten diese Beispiele einen guten Anlass zur Vertiefung bieten. Beispielsweise wird die Folge $(k + 1)/k$ nicht funktionieren, weil alle Anfangsziffern dieser Folgenglieder 1 sind.

6 Zusammenfassung

Das BENFORD-Gesetz ist ein nützliches Hilfsmittel im Stochastikunterricht. Seine überraschende Existenz gepaart mit interessanten Beispielen kann Schülern einen Anreiz bieten, Daten genauer zu untersuchen. Das Gesetz zeigt, wie mächtig statistische Methoden bei der Erkennung von Betrugsversuchen sein können. Schüler können hier Detektivarbeit leisten oder im Rahmen von Einführungskursen in die Statistik selbst forschend tätig werden.

Anfangsziffer	BENFORD-Verteilung	Anteile der Anfangsziffern für 3^k, n = 100	Anteile der Anfangsziffern für 3^k, n = 500	Anteile der Anfangsziffern für $4 \cdot 3^k$, n = 100	Anteile der Anfangsziffern für $4 \cdot 3^k$, n = 500
1	0,301	0,280	0,300	0,2900	0,304
2	0,176	0,190	0,176	0,1900	0,174
3	0,125	0,120	0,124	0,1400	0,128
4	0,097	0,080	0,098	0,0800	0,092
5	0,079	0,090	0,080	0,0700	0,082
6	0,067	0,070	0,066	0,0700	0,066
7	0,058	0,070	0,058	0,0600	0,060
8	0,051	0,050	0,052	0,0500	0,048
9	0,046	0,050	0,046	0,0500	0,046

Tab. 3: Relative Häufigkeiten der Anfangsziffern für die Folgen 3^k und $4 \cdot 3^k$ für k =1, 2, … , n

1. Spalte A: Beginnend mit A2 werden fortlaufend die Zahlen von 1 bis 100 eingetragen.

2. Spalte B: Die Glieder der geometrischen Folge 3^k werden mittels relativen Zellbezug berechnet. In Zelle B2 schreibt man die Funktion =3^A2. Anschließend kopiert man den Zelleninhalt von B2 bis zur Zelle B101.

3. Spalte C: Nun werden die Anfangsziffern der Werte der Folgenglieder 3^k bestimmt. In C2 schreibt man dazu den geschachtelten Funktionswert =LINKS(TEXT(B2;"0,00");1). Die TEXT-Funktion konvertiert einen numerischen Wert in einen Textwert, die Angabe 0,00 wandelt das numerische Format in eine Zeichenfolge um. Die LINKS-Funktion gibt das erste Zeichen einer Zeichenfolge an.

4. Spalte E: Hier werden nun die Anfangsziffern von 1 bis 9 notiert.

5. Spalte F: Die absoluten Häufigkeiten der Anfangsziffern lassen sich mithilfe der Funktion ZÄHLENWENN ermitteln, in F1 zählt man mit =ZÄHLENWENN(C2:C101;1) die Häufigkeit der Anfangsziffer 1, in F2 analog mit =ZÄHLENWENN(C2:C101;2) die Häufigkeit der Anfangsziffer 2 usw.

6. Spalte G: Hier ermittelt man die relativen Häufigkeiten der Anfangsziffern wie üblich durch Quotienten. In G2 schreibt man =F2/100 und kopiert diese Zelle bis G101.

	A	B	C	D	E	F	G
1	k	3^k	Anfangsziffer			abs. H.	rel. H.
2	1	3	3		1	28	0,280
3	2	9	9		2	19	0,190
4	3	27	2		3	11	0,110
5	4	=LINKS(TEXT(B5;"0,00");1)			4	9	0,090
6	5	24	TEXT(Wert; Textformat) 2		5	9	0,090

Abb. 2: Erzeugung der Spalte 3 in Tabelle 3

Literatur

Adhikari, A. and Sarkar, B. (1968): Distribution of most significant digit in certain functions whose arguments are random variables. *Sankhyā*, *Series B*, **30**, 47–58.

Benford, F. (1938): The law of anomalous numbers. *Proceedings of the American Philosophical Society*, **78** (4), 551–572.

Diaconis, P. (1977): The distribution of leading digits and uniform distribution mod 1. *Annals of Probability*, **5** (1), 72–81.

Giles, D. E. (2007): Benford's Law and naturally occurring prices in certain eBay auctions. *Applied Economics Letters*, **14** (3), 157–161.

Hill, T. P. (1995 a): The significant-digit phenomenon. *American Mathematical Monthly*, **102** (4), 322–327.

Hill, T. P. (1995 b): A statistical derivation of the significant-digit law. *Statistical Science*, **10** (4), 354–363.

Kumar, K., Bhattacharya, S. (2007): Detecting the dubious digits: Benford's Law in forensic accounting. *Significance*, **4** (2), 81–83.

Ley, E. (1996): On the particular distribution of the US stock indices digits. *American Statistician*, **50**, 311–313.

Marchi, De S. and Hamilton, J. T. (2006): Assessing the accuracy of self-reported data: An evaluation of the toxics release inventory. *Journal of Risk and Uncertainty*, **13**, 57–76.

Newcomb, S. (1881): Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, **4** (1), 39–40.

Nigrini, M. (1996): A taxpayer compliance application of Benford's Law. *Journal of the American Taxation Association*, **18**, 72–91.

Swanson, D., Cho, M. J., Eltinge, J. (2003): Detecting possibly fraudulent or error-prone survey data using Benford's Law. *Proceedings of the Survey Research Section*. Washington DC: American Statistical Association, 4172–4177

http://www.amstat.org/sections/srms/proceedings/y2003/Files/JSM2003-000205.pdf

Anschrift der Verfasser:

Jonathan R. Bradley
David L. Farnsworth
Rochester Institute of Technology,
New York
USA
dlfsma@rit.edu

Anmerkungen der Übersetzerin

1 Mit „Anfangsziffer" sei im Folgenden die „erste" signifikante Ziffer einer Zahl gemeint, z. B. 2 bei 0,023.

2 Nachfolgend findet der interessierte Leser einen Überblick über deutschsprachige Veröffentlichungen, die auch Erklärungen für das Auftreten des BENFORD-Gesetzes enthalten:

HUMENBERGER, H. (1996): Das „Benford-Gesetz" über die Verteilung der ersten Ziffer von Zahlen. *Stochastik in der* Schule **16 (3)**, S. 2–17.

HUMENBERGER, H.(1997): Eine Ergänzung zum BENFORD-Gesetz - weitere mögliche schulrelevante Aspekte. *Stochastik in der Schule* **17**(3),S. 42–48.

HUMENBERGER, H. (2000): Das „BENFORD-Gesetz" – Warum ist die Eins als führende Ziffer von Zahlen bevorzugt? In: FOERSTER, F.; HENN, H-W.; MEYER, J. (2000): *ISTRON-Materialien für einen realitätsbezogenen Mathematikunterricht. Bd. 6. Computer-Anwendungen*. Hildesheim: Franzbecker. S. 138–150.

HUMENBERGER, H. (2008): Eine elementarmathematische Begründung des BENFORD-Gesetzes. *Der Mathematikunterricht* **54** (1), S. 24–34.

AURZADA, F. (2003): Benfords Gesetz. *Die Wurzel* **37** (7), S. 160–163.

FISCHER, U. (2006): Ein Gleichverteilungssatz und Benfords Gesetz. *MNU* **59** (1), S. 13–16.

3 Beispielsweise könnten Schüler die Anfangsziffern von zufällig ausgewählten Zahlen aus Tageszeitungen untersuchen, vgl. KRÄMER, W. (1990): Das Gesetz der abnormalen Zahl. *Stochastik in der Schule* **19** (3), S. 48–52.