

Kommen alle Ziffern gleich häufig vor?

Bei der Frage, ob alle Ziffern mit der gleichen Häufigkeit vorkommen, geht es um Benfords Gesetz. Mit diesem werden heute grosse Datensätze mittels Computern überprüft. Christian Kleiber

Quelle: Uni Nova Basel, September 2010

Kommen alle Ziffern gleich häufig vor? Zunächst eine genauere Formulierung der Frage: Stellen wir uns vor, wir nehmen die erste Seite einer Tageszeitung und notieren von jeder dort auftretenden Zahl nur die erste Ziffer. Die Null sei nicht zulässig, das heisst, es geht nur um die Ziffern 1 bis 9. Für die Zahl 190'364 notieren wir also eine 1, für die Zahl 274'673 eine 2. (Dies sind – Sie haben es vielleicht bereits vermutet – die Einwohnerzahlen der beiden Trägerkantone der Universität Basel, Ende 2009.) Mit welcher Häufigkeit treten nun die einzelnen Ziffern auf? Hier erwarten viele, alle neun möglichen Ziffern seien, vage ausgedrückt, «gleichberechtigt» und erscheinen also jeweils in einem Neuntel der Fälle. Doch es sei hier gleich verraten: Tatsächlich sind die Häufigkeiten unterschiedlich, und die Ziffer 1 tritt mehr als sechsmal so häufig (!) auf wie die Ziffer 9.

Abgenutzte Logarithmentafeln

Das Phänomen ist seit vielen Jahren unter der etwas altertümlichen Bezeichnung «Gesetz der abnormalen Zahl» bekannt. Entdeckt wurde es schon im späten 19. Jahrhundert vom Astronomen und Mathematiker Simon Newcomb, wiederentdeckt mehrere Jahrzehnte später von Frank Benford, einem Physiker bei General Electric, nach dem es nun oft «Benfords Gesetz» genannt wird. Newcomb hatte beobachtet, dass Logarithmentafeln – vor dem Aufkommen des Computers notwendige Hilfsmittel für aufwendigere Berechnungen in den Naturwissenschaften – unterschiedlich stark abgenutzt waren, nämlich auf den vorderen Seiten mehr als auf den hinteren. Er schloss daraus, dass niedrige Ziffern, in Logarithmentafeln vorne zu finden, «von Natur aus» häufiger auftreten. Wörtlich schrieb Newcomb: «The law of probability of the occurrence of numbers is such that all mantissae of their logarithms are equally probable.» Und er stellte die Formel auf: $P(z) = \log(1 + 1/z)$ $z = 1, \dots, 9$ wobei hier $P(z)$ die Wahrscheinlichkeit für Ziffer z bezeichnet; mit «log» ist zudem der Logarithmus zur Basis 10 – der

sogenannte dekadische Logarithmus – gemeint. Daraus ergibt sich für die Ziffer 1 eine Wahrscheinlichkeit von 30,1%, für die Ziffer 9 aber eben nur von 4,6%.

Benford untersuchte in den 1930er-Jahren grosse Mengen von Zahlenmaterial aus sehr unterschiedlichen Quellen, darunter Baseballstatistiken, Oberflächen von Flüssen oder auch Adressen der ersten 342 Personen aus dem Handbuch «American Men of Science». Er wies nach, dass die jeweilige Häufigkeitsverteilung der Anfangsziffern durch die obige Formel gut beschrieben werden kann. Als aktuelleres Beispiel zeigt die Grafik rechts die Verteilung der Anfangsziffern der Einwohnerzahlen aller Schweizer Städte und Gemeinden für 2007 (schwarz) im Vergleich zur Benford-Verteilung (rot); die ziemlich genaue Übereinstimmung ist augenfällig.

Erklärung erst seit 15 Jahren

Wie lässt sich nun dieses Phänomen erklären und wann ist mit seinem Auftreten zu rechnen? Für die erwähnten Bevölkerungszahlen kann man heuristisch so argumentieren: Viele Prozesse aus der Natur und dem sozialen Bereich lassen sich durch das sogenannte geometrische Wachstum recht gut beschreiben. Stellen wir uns vor, das Bevölkerungswachstum in einem Land beträgt überall konstant 2% pro Jahr. Dann braucht eine Stadt mit 100'000 Einwohnern 100% mehr Einwohner für die nächste Anfangsziffer 2, was bei 2% Wachstum rund 35 Jahre dauert. Eine Stadt mit 50'000 Einwohnern dagegen braucht nur 20% mehr Einwohner für die nächste Anfangsziffer 6, und dies dauert nur etwa neun Jahre. Die Verweildauer bei den niedrigen Ziffern ist also grösser, was motiviert, warum hier kleine Ziffern tendenziell häufiger auftreten.

Eine Erklärung für das Phänomen in allgemeineren Zusammenhängen ist allerdings erstaunlich neu: Erst 1995 konnte der amerikanische Mathematiker Theodore P. Hill zeigen, dass die Benford-Verteilung immer dann zu erwarten ist, wenn man grosse Mengen von Zahlen aus völlig unter-

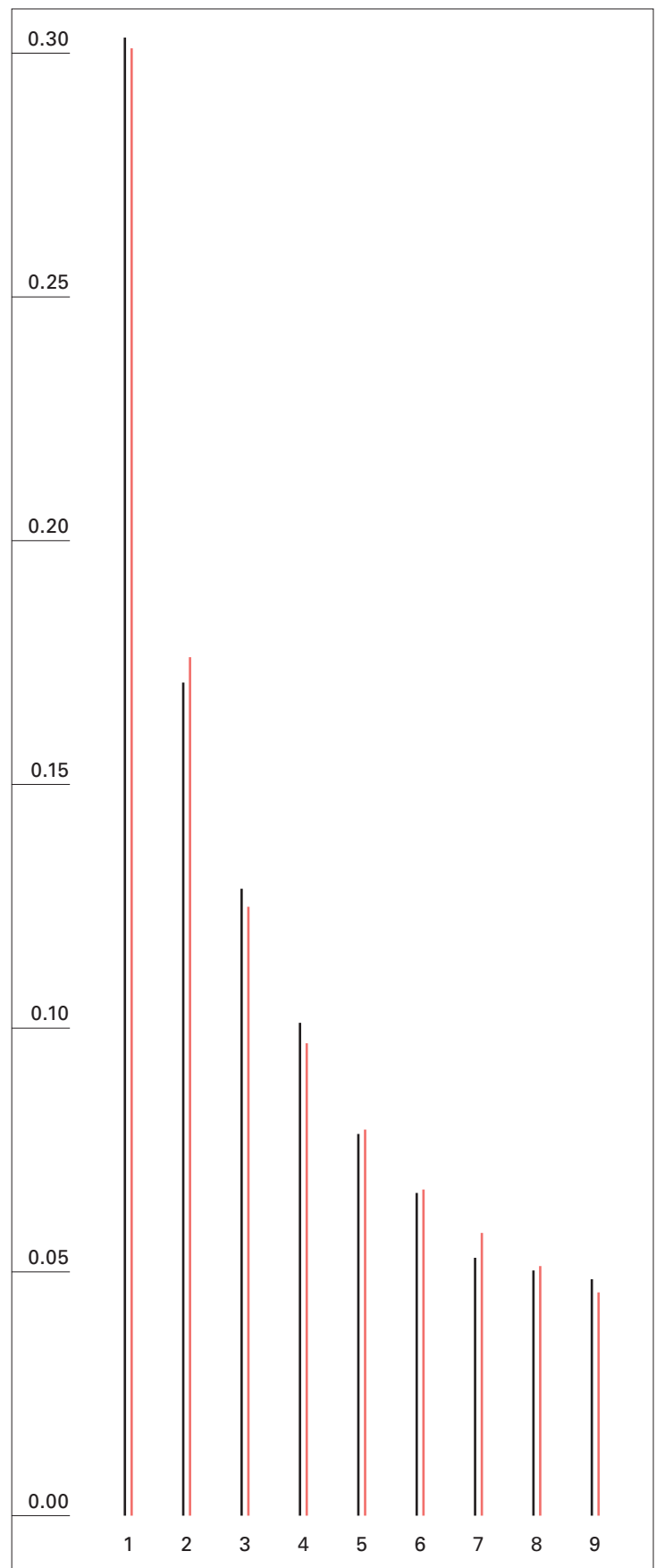
schiedlichen Quellen zusammenwirft. Und damit sind wir wieder bei der Tageszeitung, wo sich auf der ersten Seite vielleicht die Temperatur aus der aktuellen Wetterprognose, die Zahl der Verletzten bei einem Unglück, Sportresultate und diverse weitere Zahlen finden: Diese Zahlen kommen aus einer Vielzahl von ganz unterschiedlichen Quellen. Nicht zu erwarten ist die Benford-Eigenschaft dagegen bei künstlichen Zahlen wie etwa Telefonnummern: Diese beginnen in der Region Basel ja alle konstruktionsbedingt mit der Ziffer 6, wenn man den Spielregeln entsprechend die führende Null ignoriert.

Aktuelle Begeisterung

Über längere Zeit war die Anfangsziffern-Eigenschaft allerdings wenig mehr als eine empirische Kuriosität. Newcomb und Benford wären heute sicher überrascht von der Euphorie, mit der, getrieben durch die Möglichkeiten zur schnellen Auswertung von umfangreichem Zahlenmaterial am Computer, neuerdings zahlreiche Daten auf Vorliegen der oben beschriebenen Häufigkeitsverteilung überprüft werden. Dies begann Mitte der 1990er-Jahre im betrieblichen Rechnungswesen, wo man sich unter dem Etikett «forensic accounting» Hinweise auf Datenmanipulationen versprach und immer noch verspricht. Weitere Anwendungen untersuchen unter anderem die Qualität statistischer Daten in Entwicklungsländern oder Wahl- und Umfrageergebnisse, hier wiederum in der Hoffnung, Hinweise auf allfällige Manipulationen zu finden.

Es gibt aber nicht immer eine überzeugende Apriori-Erklärung, warum ein bestimmtes Datenmaterial «Benford-artig» sein sollte (oder auch nicht). Ausserdem wird manchmal vergessen, dass die benutzten statistischen Tests auf Vorliegen dieser Verteilung konstruktionsbedingt ebenfalls nicht fehlerfrei operieren können. Es dürfte also beispielsweise Daten geben, bei denen die Abweichung von der Benford-Verteilung ein Produkt des Zufalls ist – die Frage ist nur, auf welche Daten aus publizierten Arbeiten dies zutrifft. Eine gewisse Vorsicht kann angesichts der aktuellen Begeisterung für das Anfangsziffern-Phänomen gewiss nicht schaden.

Übrigens hat Louis Vladimir Furlan, ab 1932 Lehrbeauftragter und ab 1947 Professor für Versicherungsstatistik und Wirtschaftsmathematik an der Universität Basel (und damit in gewisser Weise ein Vorfahre des Verfassers), 1946 in einem lokalen Verlag ein ganzes Buch über die Häufigkeiten von Anfangsziffern in statistischem Zahlenmaterial publiziert. Er nannte das Phänomen das «Harmoniegesetz der Statistik». Und damit gibt es sogar einen Zusammenhang zwischen Benfords Gesetz und der Universität Basel.



Verteilung der Anfangsziffern der Einwohnerzahlen aller Schweizer Städte und Gemeinden für 2007 (schwarz) im Vergleich zur Benford-Verteilung (rot) [Daten: Christian Kleiber].

Prof. Christian Kleiber ist Ordinarius für Ökonometrie und Statistik an der Wirtschaftswissenschaftlichen Fakultät der Universität Basel.