Benfords Gesetz über die Verteilung der Ziffern

Habilitationsvorlesung am 9.11.01
im Rahmen des Habilitationsverfahrens
von Peter Mörters, Kaiserslautern und Bath

1 Die Geschichte von Benfords Gesetz

1881 Der Astronom **Simon Newcomb** macht die Beobachtung, daß in Logarithmentafeln die vorderen Seiten stärker abgenutzt werden als die hinteren. Er vermutet, daß sich die Häufigkeit der führenden Ziffern nach dem Gesetz

$$\text{Häufigkeit Ziffer } k \ = \log\Big(1 + \frac{1}{k}\Big)$$

verhalten, gibt aber keine weiterreichende Erklärung.

1938 Der Physiker **Frank Benford** macht unabhängig von Newcomb dieselbe Beobachtung und belegt sie mit verschiedenen Statistiken, zum Beispiel:

- Hausnummern zufällig ausgewählter Personen,
- Stromverbrauch zufällig ausgewählter Kunden,
- mathematische und physikalische Tabellen,
- Baseballstatistiken.

Nach ihm trägt die Verteilung den Namen *Benfords Gesetz.*

1961 Der Mathematiker **Roger Pinkham** zeigt, daß Benfords Gesetz die einzige *skalierungsinvariante* Verteilung auf $\{1, \ldots, 9\}$ ist.

Insgesamt gibt es bis heute über 100 Publikationen, die versuchen, Benfords Gesetz zu erklären. Außerdem gibt es Versuche, das Gesetz anzuwenden, zum Beispiel um Steuersünder zu überführen.

2 Wie kann man Benfords Gesetz erklären?

Natürlich kann man leicht Statistiken finden, die Benfords Gesetz *nicht* genügen. Ein guter Ansatz müßte also erklären, für welche Verteilungen die führenden Ziffern Benfords Gesetz genügen und insbesondere *aufdecken, daß es einen allgemeinen Mechanismus gibt, der zu Verteilungen führt, die Benfords Gesetz erfüllen.*

Man kann hier an das ähnliche Problem der *Gaußschen Normalverteilung* denken, die ebenfalls bei vielen Statistiken auftritt. Dies wird sehr befriedigend durch den zentralen Grenzwertsatz erklärt.

Eine ähnlich umfassende Erklärung wie im Fall der Gaußschen Normalverteilung ist für Benfords Gesetz leider nicht bekannt. Es gibt eine Vielzahl von Sätzen, die belegen, daß in verschiedenen Situationen Benfords Gesetz gilt, aber keine alles umfassende Erklärung.

3 Ein Grenzwertsatz

Ich möchte hier ein Beispiel für einen Grenzwertsatz geben, der einen allgemeinen Mechanismus beschreibt, der zu Benfords Gesetz führt. Der Satz folgt aus einem recht bekannten Resultat der Wahrscheinlichkeitstheorie über zufällige Rotationen eines Kreises.

Wir bezeichnen mit $D(x)$ die führende Ziffer von $x > 0$.

Satz 1 *Seien $Y_1, Y_2, \dots$ unabhängig identisch verteilte positive Zufallsvariablen mit Dichte. Dann gilt für alle Ziffern $k \in \{1, \dots, 9\}$, daß*

$$\lim_{n\to\infty} \mathbf{P}\Big\{ D\Big(\prod_{i=1}^{n} Y_i \Big) = k \Big\} = \log \Big(1 + \frac{1}{k} \Big).$$

Die Annahme, daß die Zufallsvariablen Dichten haben, kann abgeschwächt aber nicht ersatzlos weggelassen werden. Der Beweis wird dies noch aufklären.

Wir beginnen den Beweis mit einer einfachen Beobachtung von Pinkham, die von unabhängigem Interesse ist.

Lemma 2 (Pinkhams Lemma) *Wenn M_n Zufallsvariablen sind, so daß die Verteilung von $\log M_n \bmod 1$ gegen die uniforme Verteilung auf $[0,1)$ strebt, dann gilt*

$$\lim_{n\to\infty} \mathbf{P}\{ D(M_n) = k \} = \log \Big(1 + \frac{1}{k} \Big).$$

Beweis: Für alle $k \in \{1, \dots, 9\}$ gilt

$$\begin{aligned} \mathbf{P}\{D(M_n) = k\} &= \mathbf{P}\Big\{ \exists j \text{ mit } 10^j \cdot k \le M_n < 10^j \cdot (k+1) \Big\} \\ &= \mathbf{P}\Big\{ \log M_n \bmod 1 \in [\log k, \log(k+1)) \Big\} \\ &\longrightarrow \log(k+1) - \log(k) = \log \Big(1 + \frac{1}{k} \Big). \end{aligned}$$

■

Es genügt nun für $M_n = \prod_{i=1}^{n} Y_i$ die Voraussetzung von Pinkhams Lemma zu prüfen. Mit $X_n = \log Y_n$ folgt das aus dem folgenden klassischen Resultat, das vermutlich von Paul Lévy stammt.

Wir bezeichnen $S = \mathbf{R}/\sim$, wobei $x \sim y$ genau dann, wenn $x = y \bmod 1$. Man kann die Zufallsvariablen X_n also als Zufallsvariablen auf dem Kreis S, oder zufällige Rotationen des Kreises, auffassen.

Satz 3 *Seien $X_1, X_2, \dots$ unabhängig und identisch verteilte Zufallsvariablen auf S, deren Träger nicht in einem regulären Polygon enthalten ist. Dann konvergiert die Verteilung von $\sum_{i=1}^n X_i$ gegen die uniforme Verteilung auf S.*

Sei $u_0 : S \to [0,\infty)$ stetig und beschränkt. Definiere

$$u_n(x) = \mathbf{E}\Big\{u_{n-1}\Big(x + X_n\Big)\Big\} = \mathbf{E}\Big\{u_0\Big(x + \sum_{i=1}^n X_i\Big)\Big\}.$$

Zu zeigen ist, daß

$$\lim_{n\to\infty} u_n(0) = \int_0^1 u_0(x)\,dx. \tag{1}$$

Dafür genügt es zu zeigen, daß es eine Konstante m gibt, so daß für alle $x \in S$,

$$\lim_{n\to\infty} u_n(x) = m. \tag{2}$$

Tatsächlich gilt dann

$$\begin{aligned} m &= \int_0^1 \lim_{n\to\infty} u_n(x)\,dx = \lim_{n\to\infty}\int_0^1 u_n(x)\,dx \\ &= \lim_{n\to\infty} \mathbf{E}\int_0^1 u_0\Big(x + \sum_{i=1}^n X_i\Big)\,dx = \int_0^1 u_0(x)\,dx. \end{aligned}$$

Wir wollen zunächst zeigen, daß jede Folge eine konvergente Teilfolge besitzt. Es ist leicht zu prüfen, daß die Familie $\{u_n : n \in \mathbb{N}\}$ gleichgradig stetig und beschränkt ist: Tatsächlich ist u_0 auf dem Kompaktum S automatisch gleichmäßig stetig, das heißt, zu jedem $\varepsilon > 0$ existiert ein $\delta > 0$ so, daß für alle $x, y \in S$ mit $|x - y| < \delta$ gilt $|u_0(x) - u_0(y)| < \varepsilon$. Dann gilt auch

$$|u_n(x) - u_n(y)| \le \mathbf{E}\left\{\left|u_0\Big(x + \sum_{i=1}^n X_i\Big) - u_0\Big(y + \sum_{i=1}^n X_i\Big)\right|\right\} < \varepsilon.$$

Also gibt es nach dem Satz von Arzelà–Ascoli zu jeder Folge eine Teilfolge (n_j) und ein $v_0 : S \to [0,\infty)$ mit

$$\lim_{j\to\infty} \sup_{x\in S} |u_{n_j}(x) - v_0(x)| = 0.$$

Daraus folgt, für alle $n \in \mathbb{N}$, gleichmäßig in $x \in S$,

$$u_{n_j+n}(x) \quad = \quad \mathbf{E}\left\{u_{n_j}\left(x + \sum_{k=1}^n X_i\right)\right\} \longrightarrow \mathbf{E}\left\{v_0\left(x + \sum_{k=1}^n X_i\right)\right\} =: v_n(x).$$

Sei $m_n = \max_{x\in S} u_n(x)$. Offensichtlich ist m_n monoton fallend und beschränkt. Also existiert

$$m := \lim_{n\to\infty} m_n.$$

m ist das Maximum von v_0 und von allen v_n. Wir zeigen nun, daß $v_0(x) = m$ für alle $x \in S$, was den Beweis abschließt.

Angenommen, die Aussage sei falsch. Dann gibt es ein offenes Intervall I mit $v_0(x) < m$ für alle $x \in I$. Wir wollen zeigen, daß es dann ein n gibt mit $v_n(x) < m$ für alle x. Sei $\lambda := |I|$.

Der Träger von X_i enthält entweder unendlich viele Punkte von S oder zwei Punkte mit irrationalem Abstand. Im ersten Fall hat die Trägermenge einen Häufungswert und wir finden zwei Punkte im Träger deren Abstand b kleiner als $\lambda/2$ ist. Da Rotationen keinen Einfluß auf die Maxima haben, dürfen wir annehmen, daß die Punkte 0 und b im Träger von X_i liegen.

Im zweiten Fall können wir ebenso annehmen, daß der Träger von X_1 die 0 und einen irrationalen Punkt q enthält. Nun ist die Menge $\{nq : n \in \mathbb{N}\}$ unendlich und hat einen Häufungswert. Man findet insbesondere ein $r \in \mathbb{N}$ mit $b := rq \in (0, \lambda/2)$. Dann liegen 0 und b im Träger von $\sum_{i=1}^{r} X_i$.

Setzt man im ersten Fall $r = 1$ und wählt man $m > 1/b$ und $n = mr$, so liegen die Punkte

$$0, b, 2b, \ldots, mb$$

im Träger von $\sum_{i=1}^{n} X_i$ und jedes Intervall der Länge λ enthält einen dieser Punkte.

Zu jedem $x \in S$ findet man also ein y im Träger von $\sum_{i=1}^{n} X_i$ mit $x + y \in I$. Daraus folgt

$$v_n(x) = \mathbf{E}\Big\{v_0\Big(x + \sum_{i=1}^{n} X_i\Big)\Big\} < m \text{ für alle } x \in S.$$

Dies widerspricht der Tatsache, daß v_n das Maximum m annimmt. ■

4 Literatur

J. Albrecht (2000) Die Eins vom Planeten Zob *Die Zeit* vom 28.9.2000, Seite 35.

F. Benford (1938) The law of anomalous numbers. *Proc. Amer. Philos. Soc.* **78** 551–572.

P. Diaconis (1977) The distribution of leading digits and uniform distribution mod 1. *Ann. Probab.* **5** 72–81.

W. Feller (1966) An introduction to Probability Theory II. *Wiley, New York.*

T.P. Hill (1995) The significant digit phenomenon. *Amer. Math. Monthly* **83** 322–326.

P. Lévy (1939) L'addition des variables aléatoires des variables définies sur un circonférence. *Bull. Soc. Math. France* **67** 1–41.

S. Newcomb (1881) Note on the frequency of use of the different digits. *Amer. Journal Math.* **4** 39–40.

R. Pinkham (1961) On the distribution of the first significant digits. *Ann. Math. Statist.* **32** 1223–1230.

R.A. Raimi (1976) The first digit problem. *Amer. Math. Monthly* **83** 521–538.

J. Walthoe, R. Hunt and M. Pearson (1999) Looking out for number one. *Plus* **9** http://pass.maths.org.uk/issue9/features/benford/.