First Digit Law

W.A. Kreiner
Abteilung Chemische Physik
Arbeitsgruppe Laseranwendungen
Universität Ulm
D-89 069 ULM Germany

Dated: 2002-05-03

Key words: Newcomb, Benford, First digit law, Statistisches Fraktal

Historie

In vielen Tabellen statistischer und naturwissenschaftlicher Daten fällt auf, dass die meisten Zahlen mit einer Eins beginnen, die Zwei noch wesentlich häufiger als Anfangsziffer vorkommt als die Drei und die neun das Schlusslicht bildet. Aufgrund dieser seit langem bekannte Tatsache vermutete Newcomb (1), dass diesen Verteilungen ein logarithmisches Gesetz zugrunde liege und leitete daraus die Wahrscheinlichkeit für das Auftreten einer bestimmten Anfangsziffer ab. Benford (2) prüfte dies anhand einer umfangreichen Datensammlung.

Fragestellung

Wovon hängt die Häufigkeit einer bestimmten Anfangsziffer bei Zahlenensembles ab?

Die Häufigkeit für das Auftreten einer bestimmten Anfangsziffer x = 1,2, ..., 9 kann aus der Verteilungs- oder Dichtefunktion der Menge bestimmt werden.

Man beobachtet, dass sich unter den Bruchstücken eines Steins viel mehr kleine Massen befinden als große und dass die Zahlenwerte der Gewichte X über mehrere Größenordungen verteilt sind, wobei die Verteilungsfunktion sehr ähnlich bleibt (statistisches Fraktal). Die Anzahl der Bruchstücke als Funktion ihres Gewichts - die Dichtefunktion - kann als Potenz von X gewählt werden:

$$D(X) = \frac{P1}{X^P} \qquad (1)$$

Man betrachtet alle N Steine, deren Gewichte X über das Intervall [1,10) verteilt sind, also gerade über die ganze erste Dekade. Die Messwerte werden zu Gruppen z = 1,2,3, …9 zusammen-gefasst, die durch die Anfangsziffern x = 1,2, ..., 9 charakterisiert sind. Das statistische Gewicht $W_z(x)$ einer Gruppe z erhält man aus dem Integral über die Verteilungsfunktion D(X) in den Grenzen des rechtsseitig offenen Intervalls [x, x+1):

$$W_z = \int_z D(X)dX = P1 \cdot \int_x^{x+1} X^{-P} dX \quad (\text{für } x = 1,2 \ldots\ldots 9)$$

$$= \frac{P1}{1-P}\left[(x+1)^{1-P} - x^{1-P}\right] \qquad \text{mit } \sum_{x=1}^{9} W_x = N \qquad (2)$$

Daraus kann der Parameter P1 bestimmt und in Gl. (2) eingesetzt werden:

$$W_z(x) = \frac{N}{10^{1-P} - 1}\left\{(x+1)^{1-P} - x^{1-P}\right\} \qquad (3)$$

Das Verhältnis der statistischen Gewichte der Anfangsziffern 1 und 9 lautet:

$$\frac{W(1)}{W(9)} = \frac{2^{1-P} - 1^{1-P}}{10^{1-P} - 9^{1-P}} \qquad (4)$$

Gl. (3) gilt für alle Dekaden, und das Verhältnis von Einsen zu Neunen ist in allen Dekaden dasselbe. Aus diesem Grund kann die Häufigkeit der einzelnen Ziffern aus den unterschiedlichen Gewichtsdekaden zusammengefasst und aus der Anpassung von Gl. (3) an die beobachteten statistischen Gewichte W(1) ..... W(9) ein mittlerer Wert für den Exponenten P bestimmt werden.

Beispiele

In einem Garten wurden aus einem Bereich von 0,7 m^2 alle Steine an der Oberfläche zwischen 10 mg und 49,42 g (Kalkstein, Oberer Jura Schwäbische Alb) gewogen und die Häufigkeitsverteilung der ersten Ziffern des Gewichts bestimmt (Bild 1). In allen vier Größenordnungen, über die sich die 310 Messwerte verteilen, sind die Zahlen mit einer Eins vorn wesentlich häufiger als die mit einer Neun. Es ist anzunehmen, dass Erosionsvorgänge die ursprüngliche Größenverteilung stark verändert haben und z. B. die kleineren Steine auf Grund von Verwitterung unterrepräsentiert sind.

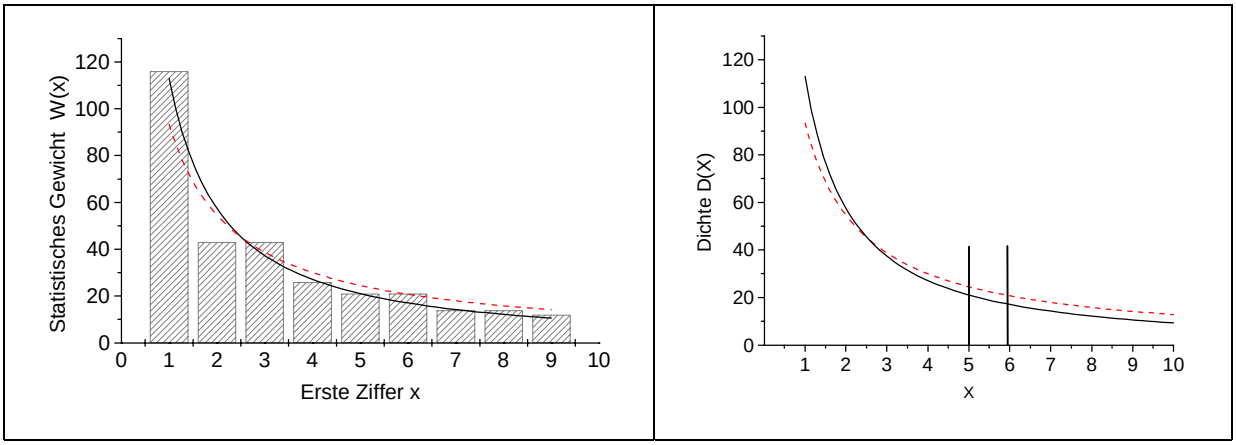

Bild 1. Links: Die Häufigkeitsverteilung der ersten Ziffern des Gewichts von 310 Steinen zwischen 10 mg und 49,42 g ergibt ein P = 1,253. Die gestrichelte Kurve entspricht P = 1 und zeigt die Verteilung, die man nach dem Newcomb/Benford-Gesetz erwarten würde. Rechts: Die daraus abgeleitete Dichtefunktion $X^{-1,253}$ im Vergleich zur Newcomb-Funktion X^{-1}. Um z. B. das statistische Gewicht der Messwerte mit der Anfangsziffer 5 zu erhalten, wird zwischen den eingezeichneten Intervallgrenzen integriert. Das Ergebnis stimmt mit der Höhe der Säule bei x = 5 im linken Diagramm überein..

Aus der Anpassung nach Gl. (3) erhält man P = 1,253(67). Damit sollte - nach Gl. (4) – die Eins 10,66 Mal so häufig an erster Stelle stehen wie die Neun. Nach Newcomb sollte dies Verhältnis nur 6,58 betragen. Dieses Verhältnis ist auch unabhängig von der Wahl des

Maßstabs, so lange er linear bleibt. In der folgenden Dichtefunktion wird die Größe X erst in inch und dann in cm gemessen. Die Funktion lautet:

$$D(X) = N.X^{-2} \qquad \text{(beliebiger Faktor N; X in inch)} \qquad (5)$$

Eine solche Funktion kann z. B. die Häufigkeit von Objekten beschreiben, die um so seltener werden, je größer sie sind, und zwar proportional zum Quadrat ihres Durchmessers. Aus dem statistischen Gewicht

$$W(x) = N \int_{x}^{x+1} \frac{1}{X^2}\, dX \qquad \text{(x = erste Ziffer)}$$

erhält man die Messwerte, die

mit einer 1 beginnen:
$$N \int_{1inch}^{2inch} \frac{1}{X^2}\, dX = -\frac{1}{X}\Big|_1^2 = N \cdot 0{,}5$$

Das Verhältnis lautet 45.

und die mit einer 9 beginnen:
$$N \int_{91nch}^{10inch} \frac{1}{X^2}\, dX = N\left[\frac{1}{9} - \frac{1}{10}\right] = \frac{N}{90}.$$

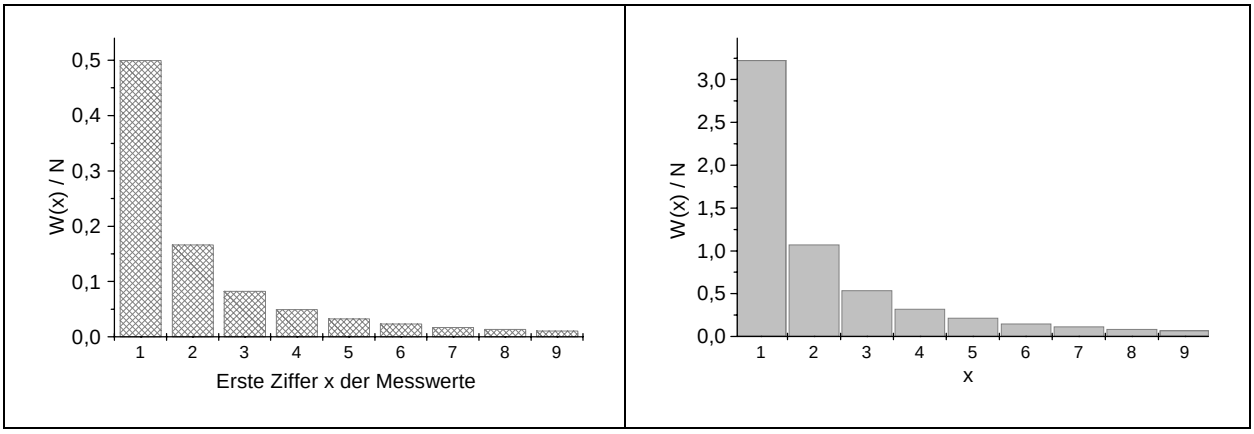

Bild 2. Statistisches Gewicht der Anfangsziffern für die Funktion $N \cdot D^{-2}$. Links: Gewichte von Messwerten (in inch) im Intervall [1,10), die mit einer bestimmten Ziffer 1, 2, ... , 9 beginnen. Rechts: Statistische Gewichte im Intervall [1,10) auf der cm-Skala. Da die Anzahl der Objekte zu kleineren X-Werten hin stark zunimmt, sind die Ordinatenwerte höher, auch wenn das cm-Intervall kleiner ist als das inch-Intervall.

Dieselbe Dichtefunktion über der cm-Skala lautet:

$$D(X) = N \cdot 2.54^2\, X^{-2}.$$

Aus den statistischen Gewichten $W(x) = N \int_{x}^{x+1} \frac{2.54^2}{X^2}\, dX$ für z_1 und z_9

(X in cm; x = erste Ziffer des entsprechenden Intervalls z) ergibt sich dasselbe Verhältnis:

$$\frac{1 - 1/2}{1/9 - 1/10} = 45.$$

Um zu festzustellen, wie sich eine Änderung des Maßstabs auf die Verteilung der Anfangsziffern auswirkt, muss geprüft werden, wie sich eine Gruppe z von Messwerten, die eine gemeinsame erste Ziffer haben, in einer anderen Messbasis auf Gruppen mit unterschiedlichen Anfangsziffern aufteilt.

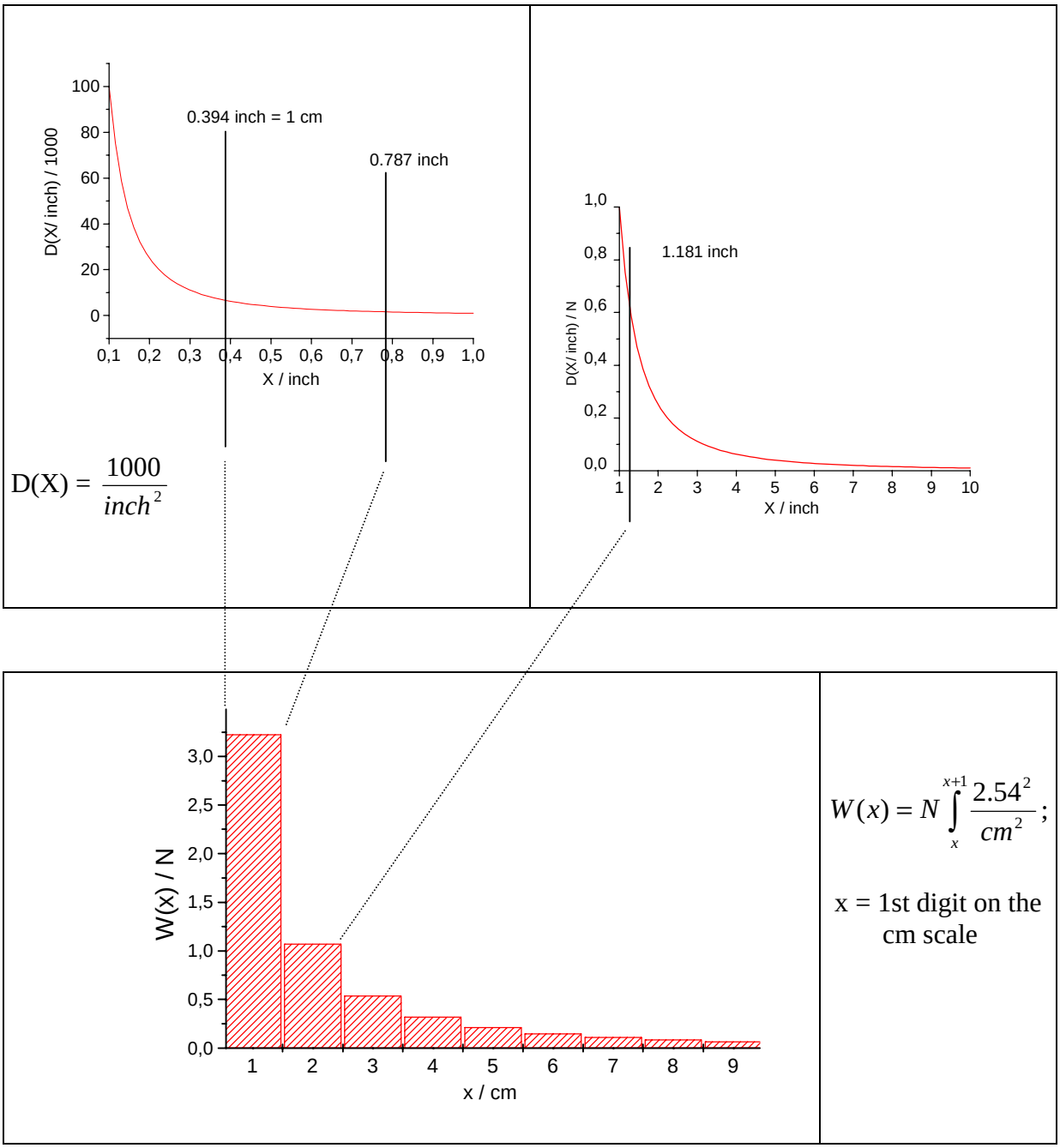

Bild 3. Unten: Statistisches Gewichte von Messwerten (in cm), die mit einer bestimmten Ziffer x = 1,2, …., 9 beginnen. Die Messwerte mit einer 1 vorn entsprechen Werten zwischen 0,394 und 0,787auf der inch-Skala, also Zahlen, die dort mit 3, 4, 5, 6, oder 7 beginnen können (oben links). Umgekehrt werden die Messwerte zwischen 1 inch und 2 inch (oben rechts) bei der Umrechnung in cm auf Zahlen mit den Anfangsziffern 2 bis 5 verteilt. Das Verhältnis von Einsen zu Neunen an der ersten Stelle der Messwerte bleibt dabei erhalten.

In Bild 3 sind zwei aufeinander folgende Dekaden der Dichtefunktion D(X/inch) gezeichnet. Sie geben die Wahrscheinlichkeit an, mit der Objekte in Abhängigkeit von ihrer Größe in inch vorkommen. Darunter ist dieselbe Funktion gezeigt, jedoch über einer cm-Skala. 1 cm entspricht 0,394 inch. Alle Objekte mit einer Größe zwischen 0,349 und 0,787 inch finden sich über der cm-Skala in einer Gruppe wieder, deren Größe mit 1 beginnt. Die Eins kommt also auch auf der cm-Skala sehr häufig vor.

Der Fall P =1 ist identisch mit der von Newcomb (1) postulierten Häufigkeitsverteilung der Anfangsziffern:

$$W(x) = \ln\frac{x+1}{x} \qquad (6)$$

Die Dichtefunktion $D \sim X^{-1}$ gilt z, B. für Mengen, deren Elemente exponentiell wachsen, wie Bakterienkulturen oder Einkommen und Steuern. Angenommen, ein angelegter Geldbetrag H(t) wachse exponentiell mit der Zeit und habe am Anfang den Wert H(0) = 1:

$$H(t) = e^{at} \qquad (7)$$

H bleibt lange Zeit bei niedrigen und zahlenmäßig sehr ähnlichen Werten, um dann immer schneller immer größere Wertebereiche zu durcheilen. Betrachtet man zu einem späteren Zeitpunkt t eine Reihe von Kapitalen, die zu unterschiedlichen Zeitpunkten angelegt wurden, dann findet man, dass ein überproportionaler Anteil kleine Werten mit niedrigen Anfangsziffern aufweist und die Werte, die mit einer Neun beginnen, selten sind. Werden f Beträge pro Jahr angelegt, dann sind es nach t Jahren n = f.t Beträge. Wie viele Beträge haben zur Zeit t eine Höhe zwischen H1und H2 ?

Aus dn = f.dt und dH = a.exp(a.t)dt folgt $\frac{dn}{dH} = \frac{f}{a}\frac{1}{H}$ und daraus

$$n(H1, H2) = \frac{f}{a}\int_{H1}^{H2}\frac{dH}{H} = \frac{f}{a}\ln\frac{H2}{H1} \qquad (8)$$

Es handelt sich um ein Integral über eine Dichtefunktion $D \sim H^{-1}$, entspricht also dem von Newcomb diskutierten Fall. In allen Intervallen, deren Grenzen im selben Verhältnis zueinander stehen, findet man die gleiche Anzahl von Kapitalen. Dazu braucht deren Startgröße nicht den Werte eins zu haben. An erster Stelle der Beträge findet sich die Eins 6,58 Mal so häufig wie die Neun vorausgesetzt, die Beträge erstrecken sich über mehrere Größenordnungen. Auch ein anderer Maßstab (Euro oder Dollar) ändert nichts an der relativen Häufigkeit der einzelnen Anfangsziffern, auch nicht eine Multiplikation aller Werte mit einem endlich großen Faktor.

In Bild 4 sind die Einnahmen der Gemeinden Baden-Württembergs über einen längeren Zeitraum zusammengefasst, wobei wieder nur die Häufigkeit der ersten Ziffern berücksichtigt wurde. Der Exponent P = 0,974(95) ergibt innerhalb der Fehlergrenzen eine mit der Newcomb-Funktion (P = 1)übereinstimmende Kurve.

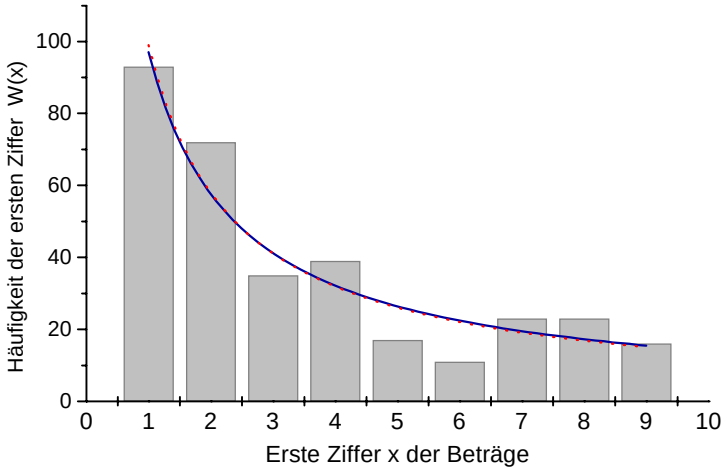

Bild 4. Einnahmen der Gemeinden Baden-Württembergs 1953-1999, aufgeschlüsselt nach sieben verschiedenen Einnahmearten (insgesamt 329 Zahlen). Aufgetragen ist die Häufigkeit der ersten Ziffer, unabhängig von der Größe der Beträge, die sich über drei Größenordnungen erstrecken. Die gestrichelte rote Kurve zeigt die Voraussage nach dem Newcomb-Benford-Gesetz (P = 1), die durchgezogene Kurve ist den Daten angepasst (P = 0,974). Quelle: Statistisches Landesamt Baden-Württemberg.

Auch Aktienkurse können eine sehr ähnliche Verteilungsfunktion aufweisen:

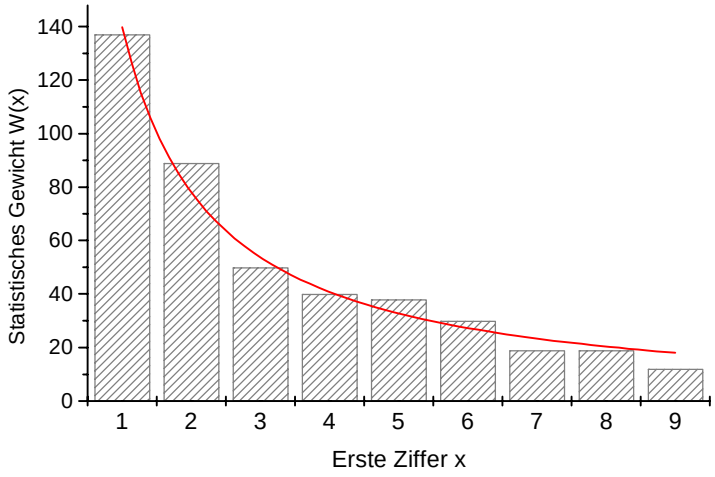

Bild 5. Häufigkeit der ersten Ziffern bei 282 Kurswerten Deutscher Aktien sowie von 152 Werten des Neuen Markts vom 03. 05. 2002. Man erhält ein P = 1,086(46), also einen geringfügig höheren Wert als bei der Newcomb-Verteilung (P = 1).

Die relative Häufigkeit der Anfangsziffern ändert sich, wenn der lineare Maßstab (1,2,3, ..) z. B. durch eine exponentielle Basis ersetzt wird: (b^1, b^2, b^3, ...). Darin besteht eine Möglichkeit, die Ungleichverteilung der Anfangsziffern aufzuheben. Ein Beispiel dafür ist in der Herstellung elektronischer Bauelemente verwirklicht: Die Größeneinteilung bei Widerstandswerten ist von einer geometrischen Reihe abgeleitet.

Hausnummern, Paragraphen, Seitenzahlen in Büchern sowie Atomgewichte beginnen deshalb so häufig mit einer Eins, weil Aufzählungen und Nummerierungen immer bei eins beginnen, aber oft nicht 9, 99 oder 999 erreichen und stets am Anfang einer Dekade die meisten Zahlen mit einer eins vorn stehen. Je kleiner der Wertebereich der Zahlenmenge ist, desto geringer ist natürlich auch die Streuung der ersten Ziffern.

Zusammenfassung:

In den meisten Fällen statistischer und technischer Datensammlungen sind die Anfangsziffern x = 1,2,3, ...9 nicht gleich häufig. Ihre Verteilung hängt von Dichtefunktion ab, also von der Funktion, die beschreibt, bei welchen Zahlenwerten Daten häufig oder selten vorkommen. Im Fall von Daten, die die Größe exponentiell wachsender Objekte wiedergeben, wird die Verteilung der Anfangsziffern durch das Newcomb/Benford-Gesetz beschrieben. Es sind jedoch auch Mengen mit anderen Verteilungsfunktionen möglich, die dann zu einer anderen Statistik in der Häufigkeit der ersten Ziffern führen.

Literatur:

1. Simon Newcomb, „Note on the Frequency of Use of the Different Digits in Natural Numbers“, American Journal of Mathematics **4**, 39-40 (1881)

2. Frank Benford, „The Law of Anomalous Numbers“, Proceed. American Phil. Society **78**, 551-572 (1938)