

# Benford's Law

<http://www.fh-fulda.de/~fd9006/dnkfln.htm#Benford>

Dies ist ein verblüffendes Gesetz über Verteilungen der ersten signifikanten Ziffer, welches verblüffend häufig für *beliebigen* Zahlen in *beliebigem* Zusammenhang erfüllt ist, z.B.

- Alle Zahlen auf der ersten Seite einer Tageszeitung (ich meine wirklich ALLE Zahlen, es ist gar nicht so leicht, sie alle zu finden!),
- Flächen von Ländern, Seen, etc (egal ob in  $m^2$ , Quadratmeilen, Quadratfuß oder Lichtjahre<sup>2</sup>)
- Einwohnerzahlen, Aktienkurse, Firmengrößen etc (egal, welche Währung)
- Größe der Files auf der Festplatte.

“Signifikant” bedeutet dabei die erste von Null verschiedene Zahl: “01.05. 2003” ergibt z.B. die signifikanten Ziffern 1, 5 und 2, 4.25 Milliarden die Ziffer 4 und 0.0001 die Ziffer 1. Als einzige Zahl wird die 0 selbst ignoriert.

Grundlage ist die Beobachtung, dass bei vielen dieser Zahlen  $X$  der *Logarithmus*  $Y = \ln X$  annähernd gleichverteilt ist. Allgemein ist dies häufig bei Zahlen der Fall, die mehrere Größenordnungen annehmen können oder bei denen der Wert von i.A. willkürlichen Einheiten (km, Inch, Wochen) abhängt. Das Gesetz ist sehr robust in Hinblick darauf, wie “gut” die Annahme der Gleichverteilung erfüllt ist. Häufig ist Benford's Gesetz bereits bei beliebigen Verteilungen angenähert erfüllt, sofern die Ausprägungen der Zufallsvariable mehrere Größenordnungen umfassen und nicht aus systematischen Gründen eine Zahl bevorzugt ist (Beispiel Files auf Festplatte: Häufig werden für ein Verzeichnis als solches 1024 oder 4096 Bytes belegt. Ist die Zahl der Verzeichnisse nicht wesentlich kleiner als die Zahl der gewöhnlichen Dateien, wird systematisch die 1 bzw. die 4 bevorzugt.

## Herleitung

Gesucht ist zunächst die Dichtefunktion  $f(x)$ , so dass für den Logarithmus  $Y = \ln X$  gilt:  $g(y) = C_1 = \text{const.}$  ( $1/C_1$  gibt die Zahl der Dekaden bzw.  $e$ -Potenzen an für die die Gleichverteilung gilt)

$$f(x)dx = g(y)dy = g(y(x))\frac{dy}{dx}dx = C_1\frac{dy}{dx}dx = \frac{C_1}{x}dx, \quad (1)$$

also

$$f(x) = \frac{C_1}{x}. \quad (2)$$

Bei Zahlen im Dezimalsystem geht die erste signifikante Ziffer von  $i = 1$  bis 9 (bei “Null komma” wird ja die erste Ziffer  $\neq 0$  genommen). Benford's Law ist die

Verteilung  $p_i$  dieser ersten Ziffer. Die Berechnung ist einfach aus den Bedingungen

$$p_i = C_2 \int_i^{i+1} f(x) dx \quad \text{mit } C_2 \text{ aus } \sum_{i=1}^9 p_i = 1 \quad (3)$$

Also (mit  $C = C_1 C_2$ )

$$p_i = C \ln\left(\frac{i+1}{i}\right), \quad C = \frac{1}{\ln\left(\frac{10}{1}\right)} \quad (4)$$

und damit

$$\boxed{p_i = \frac{\ln\left(\frac{i+1}{i}\right)}{\ln(10)} = \log_{10}\left(1 + \frac{1}{i}\right)} \quad (5)$$

## Bemerkung

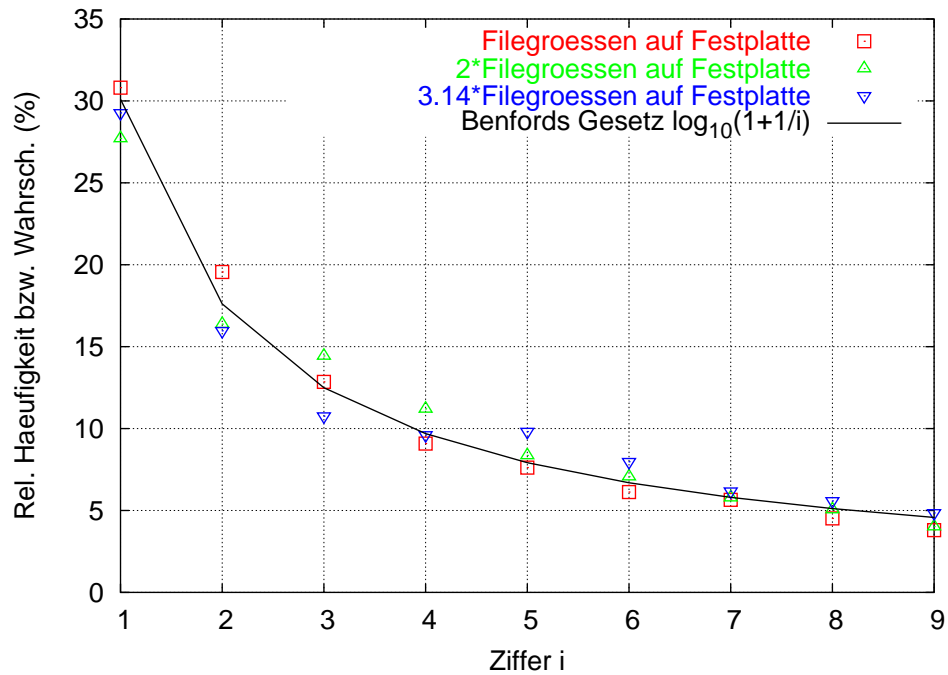
Wie der zentrale Grenzwertsatz ist Benford's law erstaunlich robust und gilt verblüffend gut, auch wenn seine Voraussetzung (ein gleichverteilten Logarithmus) *nicht* erfüllt ist. Wenn die Werte um mehrere Größenordnungen schwanken (z.B. bei einer Steuererklärung), gilt es de-facto fast immer. Nur Zahlen, die explizit auf eine Dekade aufgeteilt werden, wie Telefonnummern und Wertpapierkenn-Nummern muss man ausschließen.

## Vorschläge für Aufgaben (zu nichtparametrischen Tests)

In der Übung davor Studenten auffordern, in die nächste Übung beliebige aktuelle Tages- oder Wochenzeitung mitzubringen (ohne preiszugeben, was man damit vor hat) In der Übung eine Zeitung geben lassen (zur Sicherheit eigene dabei haben), alle Zahlen auf erster Seite suchen und jeweils erste Ziffer in Strichliste darstellen.

- Nichtparametrischer Test von Benford's Law
- Nichtparametrischer Test auf Gleichverteilung mit denselben Zahlen (letzterer sollte signifikant abgelehnt werden, wenn es mehr als 30 Zahlen sind; falls es weniger sind, genauer gucken! man hat sicher einige Zahlen übersehen!)
- Nichtparametrischer Test der *letzten* Ziffer  $\neq 0$  jeder Zahl mit mindestens zwei Ziffern  $\neq 0$  auf Gleich- und Benford-Verteilung. Hier sollte das Ergebnis genau andersrum (zugunsten der Gleichverteilung) ausfallen (evtl. ist allerdings die 5 dominant).

## Beispiel: Größe von Dateien auf der Festplatte



Dargestellt ist die Verteilung der ersten Ziffern der Größen von etwa 200 000 Dateien auf einer Festplatte, im Vergleich mit Benfords Gesetz (5). Um die Unabhängigkeit von z.B. Einheiten (Bit, Byte, etc...) zu zeigen, wurden die Dateigrößen zusätzlich mit 2 und mit  $\pi \approx 3.14$  multipliziert und nochmals analysiert.