DAS BENFORDSCHE GESETZ

Saskia Brech, Finn Heydemann, Andreas Kruff

Technische Hochschule Köln

Technology
Arts Sciences
TH Köln

Geschichte und Einleitung

Das Benfordsche Gesetz (in engl. Benford's Law) beschreibt die Verteilung von **führenden Ziffern**, also den ersten Stellen einer Zahl. Im Falle des Benfordschen Gesetzes werden allerdings führende Nullen nicht betrachtet, weshalb man auch von **signifikanten Ziffern** spricht [5]. Es sagt aus, dass die erste bzw. die ersten beiden führenden Ziffern nicht gleichmäßig, sondern logarithmisch verteilt sind. Demnach kommen niedrige Ziffern häufiger vor als hohe Ziffern.

Das Benfordsche Gesetz wird auch Newcomb-Benford's Law (NBL) genannt. Grund hierfür ist, dass die Gesetzmäßigkeit bereits 1881 von **Simon Newcomb** entdeckt wurde. Er wunderte sich, dass die Logarithmentafeln auf den ersten Seiten und besonders die der Eins abgenutzter waren als die anderen und demnach häufiger verwendet würden [9]. Seine Entdeckung blieb allerdings weitestgehend unentdeckt, was dazu führte, dass **Frank Benford** 1938 die Gesetzmäßigkeit der Verteilung nochmals entdeckte und veröffentlichte [2].

Mathematischer Hintergrund

Die **Wahrscheinlichkeit** für die **erste Ziffer** im Benfordschen Gesetz ist definiert durch:

$$P(d_1) = \log(d_1{+}1) - \log(d_1) = \log(\frac{d_1+1}{d_1}) \quad (d_1 = 1, ..., 9)$$

Die Gesamtwahrscheinlichkeit für eine bestimmte Zahl an der **2ten bis nten Ziffer** ist definiert durch:

$$P(d) = \sum_{k=10^{(n-1)}}^{10^{(n-1)}-1} \log(1 + \frac{1}{10 * k * d}) \quad (d = 0, ..., 9)$$

Die **statistische Signifikanz** der Verteilung der ersten beiden Ziffern wird dabei im experimentellen Teil des Posters mithilfe des **Chi-Quadrat-Tests** wie folgt berechnet:

$$\chi^2 = \sum_{j=1}^{m} \frac{(N_j - n_{0j})^2}{n_{0j}}$$

Die folgenden Eigenschaften, sind weitere Indizien dafür, dass ein Datensatz möglicherweise nach dem Benfordschen Gesetz verteilt sind:

1. Die Bruchteile ihres dekadischen Logarithmus sind gleichmäßig zwischen 0 und 1 verteilt.
2. Die Verteilung ihrer signifikanten Ziffern bleibt unter Änderungen des Maßstabs invariant.
3. Die Verteilung ihrer signifikanten Ziffern ist kontinuierlich und bleibt unter Änderungen der Basis invariant.

Eine ausführliche Zusammenfassung von Spezialfällen und Beispiele, die die oben erwähnten Charakteristiken zeigen, finden sie in [3]. Die dazugehörigen mathematischen Beweise in [1]

Eigenschaften, die häufig für eine Benfordsche Verteilung der Daten sprechen:

1. Zahlen im Datensatz haben eine Bandbreite über mehrere Zehnerpotenzen (**Dispergierung**).
Mithilfe von

$$\log_{10}(X) = \log_{10}(r * 10^n) = \log_{10}(r) + n$$

kann gezeigt werden, dass die Benford Verteilung in den Intervallen widergespiegelt wird. Eine wichtige Eigenschaft dafür ist, dass eine hinreichend glatte Verteilung vorliegt, die sich über mehrere ganze Zehnerpotenzen erstreckt.[4]

2. Stabilität gegenüber gemischten Populationen [6]
3. Konstantes geometrisches Wachstum in Bezug zum Benfordschen Gesetz (Beispielrechnung: [7])

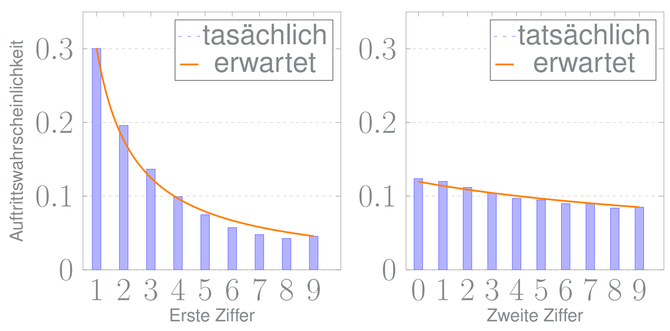
Einsatz in der Praxis

Das Benfordsche Gesetz kann dazu verwendet werden, **Betrugsfälle** zu **identifizieren**. Dabei werden Datensätze darauf überprüft, ob ihre führenden Ziffern im Wesentlichen der erwarteten Benfordschen Verteilung entsprechen. **Abweichungen** von der erwarteten Verteilung können darauf hinweisen, dass der Datensatz möglicherweise **manipuliert** wurde. Während der Hauptaspekt in der Detektion von manipulierten Daten liegt, findet das Gesetz auch Verwendung im Bereich der Detektion von "natürlichen Phänomenen" wie z.B. der Erdbebenfrüherkennung oder im Bereich der *Computer Science*.[1]

Beispiel A: Bevölkerung in Städten

Als erstes Beispiel wird ein Datensatz herangezogen, der die Bevölkerung von 50.000 Städten [8] beinhaltet. Die tatsächliche Verteilung der **ersten Ziffer** (blaue Balken) folgt dabei mit statischer Unschärfe der nach Benford erwarteten Verteilung (orange Linie). Um einen **statistischen Zusammenhang** zu beweisen, wird die Nullhypothese (H_0), dass die Verteilung der ersten signifikanten Ziffer der Bevölkerung ausgewählter Städte dem Benfordschen Gesetz entspricht, aufgestellt. Die H_0 wird angenommen (vgl. Tabelle):

$$\chi^2 = 8,26 \times 10^{-3} < \chi^2_{krit}(0,005; 8) = 1,34$$

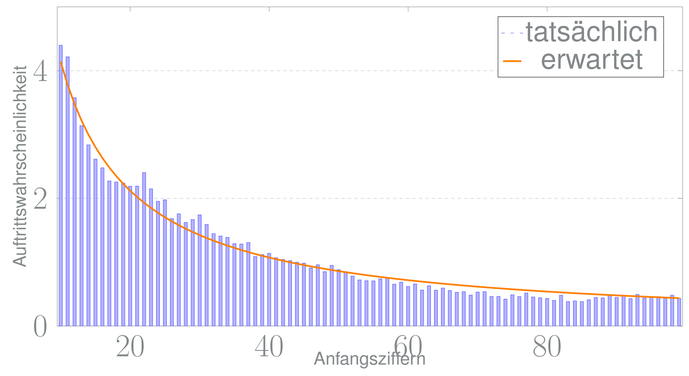


Ziffer	tatsächlich	erwartet	χ^2
1	3.0027×10^{-1}	3.0103×10^{-1}	1.9033×10^{-6}
2	1.9590×10^{-1}	1.7609×10^{-1}	2.2287×10^{-3}
...	...	...	...
9	4.5546×10^{-2}	4.5757×10^{-2}	9.7332×10^{-7}
$\sum$	1	1	8.2647×10^{-3}

Auch bei der Betrachtung der jeweils **zweiten signifikanten Ziffer** (0 tritt hier ebenso auf) zeigt sich, dass die Verteilung der vom Benfordschem Gesetz vorausgesagten Verteilung folgt. Die vorausgesagte und tatsächliche Verteilung näheren sich der Gleichverteilung an. H_0 wird hier ebenso angenommen:

$$\chi^2 = 9,76 \times 10^{-4} < \chi^2_{krit}(0,005; 9) = 1,73$$

Mithilfe des Benfordschen Gesetzes können auch Wahrscheinlichkeiten für das Auftreten von **Ziffernkombinationen** berechnet werden. Die Auftrittswahrscheinlichkeiten für die Anfangsziffern 10 bis 99 verhalten sich in diesem Datensatz nahezu vollständig wie vom Benfordschem Gesetz vorausgesagt.



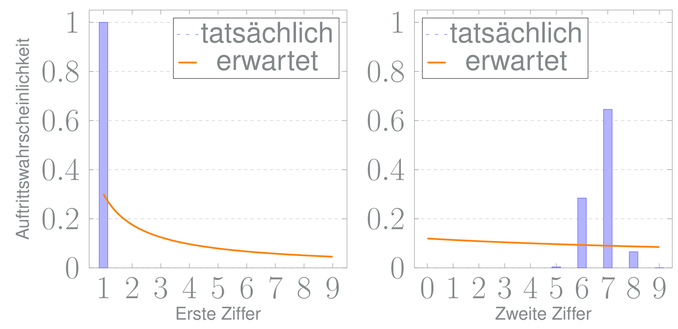
Beispiel B: Körpergrösse Menschen

Als weiterer Datensatz wird die Körpergröße ausgewählter 18-jähriger Menschen betrachtet [10]. Die tatsächliche Verteilung der **ersten Ziffer** (ausschließlich die Ziffer 1) folgt jedoch keineswegs der erwarteten Verteilung. Der visuelle Eindruck wird statistisch durch **Ablehnung** der H_0 bestätigt:

$$\chi^2 = 2,32 > \chi^2_{krit}(0,005; 8) = 1,34$$

Da die Körpergröße der Personen mit $\mu = 1,7m$ und $\sigma = 4,83cm$ annähernd normalverteilt sind, treten die Ziffern 7, 6 und 8 an **zweiter Stelle** besonders häufig auf. Damit folgen sie nicht der vom Benfordschem Gesetz erwarteten Verteilung. H_0 wird **verworfen**:

$$\chi^2 = 4,3487 > \chi^2_{krit}(0,005; 9) = 1,7349$$



Evaluation

Im Gegensatz zum ersten Beispiel folgen die Ziffern im zweiten Beispiel nicht der von Benford vorausgesagten Verteilung, da 18-Jährige sich bereits nahe ihrer endgültigen (normalverteilten) Körpergröße befinden. Die sich innerhalb **natürlicher Grenzen** bewegenden Körpergrößen **dispergieren** somit anders als die Daten im Städtedatensatz nicht ausreichend. Auf eine Manipulation der Körpergrößen lässt sich trotz fehlender Benfordverteilung aufgrund der fehlenden Erfüllung von Anforderungen **nicht schließen**. Die Bevölkerung von Städten weist dagegen nach Benford **keine Unregelmäßigkeiten** auf.

Fazit

Obwohl das Benfordsche Gesetz zum ersten Mal vor mehr als 100 Jahren entdeckt wurde, bleibt die Thematik weiterhin Gegenstand intensiver **wissenschaftlicher Untersuchungen** und **praktischer Anwendungen** in verschiedenen Disziplinen. Auch in den für die Ausarbeitung gewählten Datensätzen lassen sich die Gründe für das Auftreten der Benfordschen Verteilung finden und erklären. Durch die **zugrundeliegenden Gesetzmäßigkeiten** lassen sich Datensätze identifizieren, die potenziell der hier behandelten Verteilung folgen. Die Analyse von Abweichungen dieser Verteilung lassen sich, wie in Kapitel "Einsatz in der Praxis" beschrieben, nutzen.

[1] Theodore P. Berger Arno Berger. *An Introduction to Benford's Law*. Princeton University Press, 2015. DOI: https://doi.org/10.1515/9781400866588.
[2] Frank Benford. "The Law of Anomalous Numbers". In: *Proceedings of the American Philosophical Society (Proc. Amer. Phil. Soc.)* (1938). URL: http://www.jstor.org/stable/984802..
[3] Arno Berger u. a. *The Mathematics of Benford's Law – A Primer*. 2020. arXiv: 1909.07527 [math.ST].
[4] R. Fewster. "A Simple Explanation of Benford's Law". In: *The American Statistician* 63 (Feb. 2009), S. 26–32. DOI: 10.1198/tast.2009.0005.
[5] D.c. Harris. "Experimenteller Fehler". In: *Lehrbuch der Quantitativen Analyse* (1998). DOI: https://doi.org/10.1007/978-3-642-37788-4.
[6] Theodore P. Hill. "A Statistical Derivation of the Significant-Digit Law". In: *Statistical Science* 10.4 (1995), S. 354–363. DOI: 10.1214/ss/1177009869. URL: https://doi.org/10.1214/ss/1177009869.
[7] Steven Miller. *Benford's Law: Theory and Applications*. Juni 2015. ISBN: 9781400866595.
[8] Max Mind. *World Cities Database*. Kaggle, 2008. URL: https://www.kaggle.com/datasets/max-mind/world-cities-database (Zuletzt besucht am 22. 12. 2023).
[9] Simon Newcomb. "Note on the Frequency of the Use of different Digits in Natural Numbers". In: *American journal of mathematics (Amer. J. Math.)* (1938). URL: https://www.jstor.org/stable/2369148.
[10] Smit Partel. *Heights and Weights Dataset*. Kaggle, 2008. URL: https://www.kaggle.com/datasets/burnoutminer/heights-and-weights-dataset (Zuletzt besucht am 22. 12. 2023).