Benfords Gesetz und seine Anwendung in der Betrugsaufdeckung

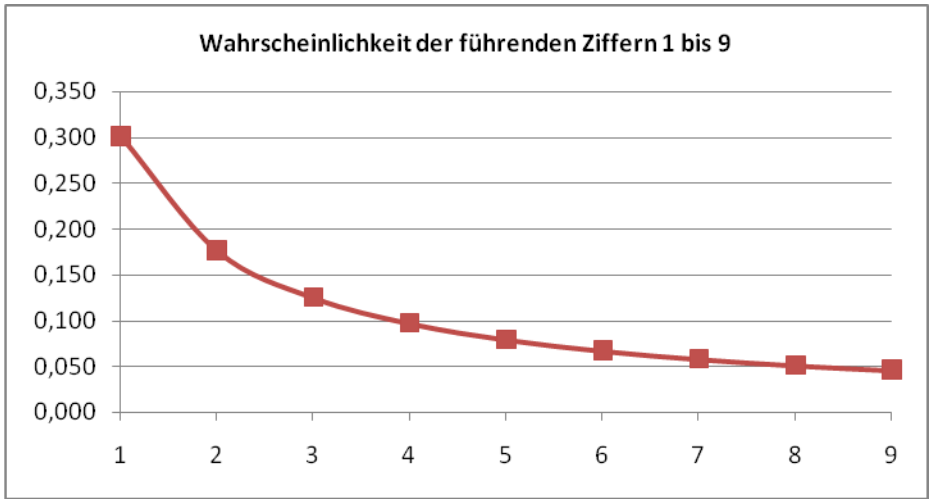
Ansgar Dorneich, 2014-06-19

Benfords Gesetz

In den 1920er Jahren entdeckte Frank Benford, ein bei General Electric arbeitender Physiker, dass in vielen numerischen Datensammlungen aus verschiedensten Bereichen des täglichen Lebens die Vorkommenshäufigkeiten der führenden Ziffern eine deutlich unterschiedliche Häufigkeitsverteilung aufweisen: Zahlenwerte mit ‚1' als führender Ziffer tauchen viel häufiger auf als solche mit ‚9' als führender Ziffer. Mathematisch präziser formuliert besagt Benfords Gesetz: die Wahrscheinlichkeit P, dass ein Zahlenwert mit der Ziffernkombination z beginnt, ist

$$P(z) = \log_{10}(1+ 1/z).$$

Die Wahrscheinlichkeit, dass ein Zahlenwert mit 1 beginnt, ist also $\log_{10}(2) = 30{,}1\%$, und die Wahrscheinlichkeit für 9 als führende Ziffer ist $\log_{10}(10/9) = 4{,}6\%$. Die Wahrscheinlichkeit für die führende Ziffernkombination 10 ist $\log_{10}(11/10) = 4{,}1\%$ und die für die Ziffern 99 ist $\log_{10}(100/99) = 0{,}4\%$.



Voraussetzungen für die Gültigkeit von Benfords Gesetz:

- Jeder Zahlenwert in der betrachteten Wertemenge beschreibt dieselbe Größe (z.B. den Preis, das Gewicht, die Anzahl, die Dauer, das Volumen) eines Dinges oder Individuums aus einer Menge gleichartiger Dinge oder Individuen.
- Es sollte keine fest vorgegebenen Werteobergrenzen und –untergrenzen geben.
- Die Werte sollten keine nach einem bestimmten Schema künstlich zugewiesenen Werte sein, wie zum Beispiel Kundennummern oder Artikelnummern.
- Der betrachtete Zahlenwert oder ein mit diesem stark korreliertes Merkmal sollte nicht für die Einteilung und Abgrenzung der einzelnen Entitäten oder Individuen in dem Datensatz verwendet worden sein. So folgt zum Beispiel die Verteilung der Bevölkerungszahlen in den Wahlkreisen zur Bundestagswahl in Deutschland nicht dem Benfordschen Gesetz, weil die Wahlkreise so eingeteilt wurden, dass in jedem der 299 Wahlkreise etwa 1/299 der in Deutschland Wahlberechtigten wohnt, also etwa 205000 Wahlberechtigte. Da die Anzahl der Wahlberechtigten fast hundertprozentig mit der gesamten Bevölkerungszahl korreliert, wird jeder Wahlkreis etwa eine Bevölkerung von 230000 bis 280000 Personen umfassen, so dass (fast) alle Zahlenwerte mit der Ziffer 2 beginnen.

Beispiele für die Anwendbarkeit von Benfords Gesetz

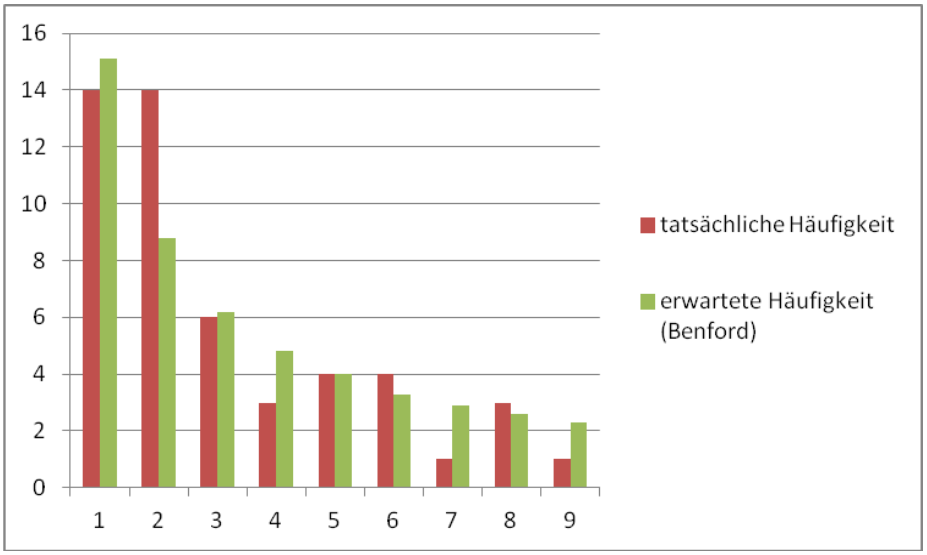
Benfords Gesetz ist insbesondere gültig für zwei Klassen von Datensammlungen:

1. Daten, die sich auf ‚Fraktale' (also skaleninvariante, selbstähnliche Strukturen) wie zum Beispiel geologische Strukturen beziehen.
2. Daten, die sich auf exponentielle Wachstumsprozesse (z.B. ökonomische oder biologische Prozesse) beziehen

Nachfolgend soll für diese beiden Arten von Daten je ein Beispiel gegeben werden. Die folgende Tabelle enthält Daten zu den 17 größten Binnenseen der Welt (Fläche, größte Tiefe, Wasservolumen; Quelle: http://www.taschenhirn.de/geografie/seen-der-erde/):

Binnensee	**Land**	**Fläche (km²)**	**Tiefe (m)**	**Volumen (km³)**
Kaspisches Meer	Russland, Kasachstan, Iran	386.500	1025	78200
Oberer See	USA, Kanada	82.103	405	12100
Victoriasee	Kenia, Tansania, Uganda	69.484	81	2760
Huronsee	USA, Kanada	59.570	229	3550
Michigansee	USA	57.866	281	4920
Tanganjikasee	Sambia, Tansania, Kongo	32.893	1435	18900
Großer Bärensee	Kanada	31.795	446	2240
Baikalsee	Russland	31.490	1637	23600
Großer Sklavensee	Kanada	28.570	614	2000
Eriesee	USA, Kanada	25.667	64	500
Winnipeggsee	Kanada	24.390	18	285
Malawisee	Malawi, Mosambik	23.300	695	8400
Ontariosee	USA, Kanada	19.011	244	1640
Balchaschsee	Kasachstan	18.428	26	105
Ladogasee	Russland	17.703	230	910
Wostoksee	Antarktis (unter dem Eis)	15.700	1000	540
Maracaibosee	Venezuela	13.512	35	

Wenn man für die 50 in der Tabelle enthaltenen Zahlenwerte die Häufigkeit der führenden Ziffern zählt, erhält man eine im Rahmen der erwartbaren statistischen Unsicherheit (Faustregel: Abweichnungen von weniger als 5 Fällen zwischen Erwartung und Ergebnis sind nie statistisch signifikant) gut mit der Benford-Verteilung übereinstimmende Häufigkeitsverteilung:

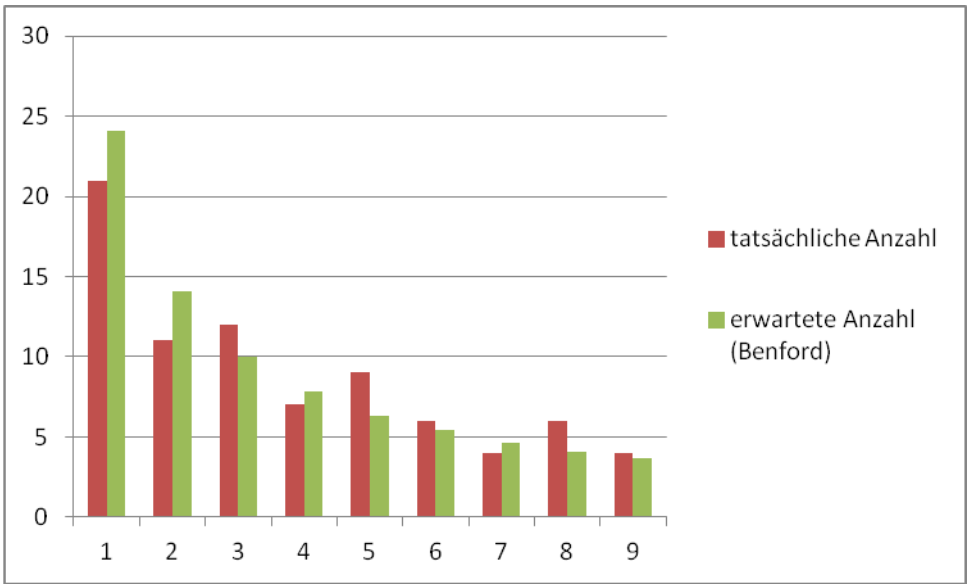


Die nächste Tabelle enthält die Aktienkurse der 80 größten deutschen börsengehandelten Unternehmen (DAX und MDAX) am 19.06.14, sortiert nach den führenden Ziffern ihres Aktienkurses in EUR:

Aktie (DAX+MDAX)	**Kurs [EUR] am 19.06.14**
BAYER AG	103,6
FIELMANN	104,25
Hugo Boss AG	109,05
FRESENIUS SE & CO KGAA	109,7
KABEL DT.	109,85
KLOECKNER	11,34
TUI	11,96
COMMERZBANK AG	12,5
DEUTSCHE TELEKOM AG	12,735
GAGFAH	12,87
ALLIANZ SE	124,15
MERCK KGAA	128,75
Brenntag AG	136,9
SÜDZUCKER	14,12
E.ON SE	14,915
DEUTSCHE LUFTHANSA AG	15,79
DT. WOHNEN	15,84
LINDE AG	156,8
MUENCHENER RUECKVERSICHERUNGS-GESELLSCHAFT AG	161,8
CONTINENTAL AG	172,1
VOLKSWAGEN AG VZ	197,4
THYSSENKRUPP AG	21,615
RATIONAL	235,3
RHÖN-KLINIKUM	24,49
K+S AG	24,695
DMG MORI SEIKI	25,28
SGL CARBON	25,29
Talanx AG	25,5
CELESIO	26,08
DEUTSCHE POST AG	26,59
DEUTSCHE BANK AG	27,7
Evonik Industries AG	29,58
METRO	30,62
RWE AG	30,945
SALZGITTER	31,42
ELRINGKLINGER	31,88
FUCHS PETRO.	33,82
STADA	34,42
ProSiebenSat.1 Media AG	34,85
GEA GROUP	34,86
OSRAM Licht AG	35,31
AAREAL BANK	36,29
DT. EUROSHOP	36,48
GERRY WEBER	37,18
AURUBIS	40,26
SYMRISE	40,67
Norma Group SE	41,46
KUKA	42,88
WINCOR NIXD.	42,95
AXEL SPRINGER	46,5
FRESENIUS MEDICAL CARE AG & CO KGAA	47,42
AIRBUS GROUP	50,52
LEG Immobilien AG	50,82
LANXESS AG	50,88
RHEINMETALL	52,04
GERRESHEIMER	52,35
FRAPORT	52,76
DEUTSCHE BOERSE AG	56,09
SAP AG	57,77
LEONI	59,68
SKY	6,86
HEIDELBERGCEMENT AG	64,69
DÜRR	65,48
HANNO. RÜCK	65,68
HOCHTIEF	65,75
MTU AERO	69,29

DAIMLER AG	70,22
BEIERSDORF AG	72,131
KRONES	75,01
ADIDAS AG	77,11
TAG IMMOBILIEN	8,86
RTL Group	83,4
BILFINGER	83,44
HENKEL AG & CO KGAA	84,01
BASF SE	87,36
WACKER CHEM.	88,57
INFINEON TECHNOLOGIES AG	9,411
MAN	90,14
BMW AG	92,4
SIEMENS AG	99,61

Auch hier passt die tatsächliche Häufigkeit der führenden Ziffern gut zu der erwarteten:



Herleitung von Benfords Gesetz aus der Betrachtung exponentieller Wachstumsprozesse

Die Wahrscheinlichkeits-Formel $P(z) = \log_{10}(1+ 1/z)$ kann man sich folgendermaßen herleiten, wenn man von einem exponentiellen Wachstumsprozess ausgeht:

Betrachten wir ein Wertpapier, dessen Wert sich kontinuierlich nach einem exponentiellen Wachstumsprozess vergrößert. Zum Zeitpunkt t=0 sei der Wert des Papiers 1.0, zum Zeitpunkt t=1 sei der Wert 10.0. Der Wert W(t) zu einem beliebigen Zeitpunkt t ist also $W(t)=10^t$ (Wir könnten zum Beispiel in 30-Jahres-Einheiten rechnen, dann hätte sich der Wert des Papiers also nach 30 Jahren verzehnfacht, was ein für eine Aktie realistischer Wert ist).

Da wir genau den Zeitraum einer Zeiteinheit betrachten und genau eine 10er-Dekade 1.0…10.0 der Wertentwicklung, ist die Wahrscheinlichkeit P(z), dass die Ziffer z als führende Ziffer des Wertes auftaucht, genau gleich der Zeitdifferenz $dt(z) = t(z+1) – t(z)$, wobei $t(z)$ der Zeitpunkt ist, bei dem gilt: $10^{t(z)}=z$ und $t(z+1)$ der Zeitpunkt, bei dem gilt: $10^{t(z+1)}=z+1$. Nach Logarithmieren wird daraus: $t(z) = \log_{10}(z)$ und $t(z+1) = \log_{10}(z+1)$

Nun ergibt sich leicht die Benford-Formel: $P(z) = t(z+1) – t(z) = \log_{10}(z+1) – \log_{10}(z) = \log_{10}(1+1/z)$.

Anwendung des Benfordschen Gesetzes zur Erkennung von Problemen und Betrugsfällen

Wer im Zusammenhang mit Geldwäsche, Betrug oder Unterschlagung größere Zahlungsströme in vielen kleinen unauffälligen Einzelbeträgen verpacken will, der versucht normalerweise, viele zufällig ausgewählte kleine Buchungsbeträge zu erzeugen. Wenn aber die korrekten Buchungen dem Benfordschen Gesetz folgen, dann genügt oft schon eine vergleichsweise geringe Zahl an künstlich hinzugefügten 'Zufallsbuchungen', um die Gesamtstatistik signifikant von der Benford-Kurve zu entfernen – eine Tatsache, die durch automatisierte, Software-basierte Buchungsüberwachung aufgedeckt werden kann.

Ein ganz anderes Anwendungsgebiet der Benford-Analyse ist die Aufdeckung von systematischen Fehlerquellen und Stockungen in Unternehmensprozessen und Arbeitsabläufen. Viele Zeitdauer-Verteilungen von Arbeitsabläufen entsprechen den Benford-Kriterien. Jede gemessene signifikante Abweichung der Zeitdauern von der Benford-Verteilung kann ein Indiz für ein systematisches Problem sein. So kann eine oberhalb der Benford-Verteilung liegende Häufung von Reklamationen, Hotline-Anfragen oder Reparaturen zu einem bestimmten Zeitpunkt nach dem Kauf eines Produkts auf systematische Fehler im Produkt oder der begleitenden Benutzerdokumentation hinweisen. Ein weiteres Beispiel ist die Benford-Analyse von Wartezeiten und Gesprächsdauern bei Telefon-Hotlines. Ein signifikantes Absinken der Wartezeit-Häufigkeit oberhalb bestimmter Werte kann bedeuten, dass diese Werte die vom Kunden maximal akzeptierte Verweildauer in der Warteschleife darstellen und viele Kunden nach Ablauf dieser Zeit verärgert den Kontaktversuch beenden.

Kombination der Benford-Analyse mit Data-Mining-Methoden

Beim Einsatz der Benford-Analyse in der Praxis wird man immer wieder mit der Frage konfrontiert, ob die an einem konkreten Datensatz gemessene Abweichung von der Benford-Verteilung nun als signifikant einzustufen ist oder nicht. Dabei sind folgende Fragen zu klären:

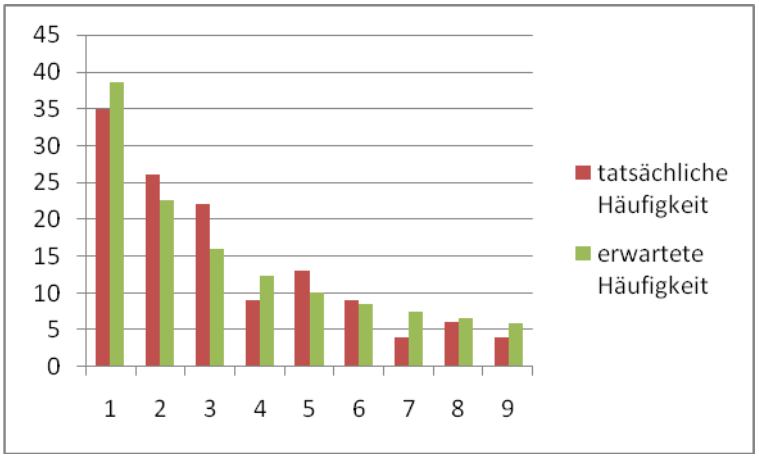
- Liegt die gemessene Abweichung noch im Rahmen der für diese Datensatz-Anzahl zu erwartenden statistischen Fluktuationen?
- Gibt es Korrelationen und Kausalzusammenhänge, die auf die betrachtete Größe einwirken und für gewisse Teilmengen der Datensätze eine systematische Abweichung von der Benford-Statistik erwarten lassen.

Bei der Beantwortung beider Fragen hilft die Kombination der Benford-Analyse mit Verfahren wie der Assoziationsanalyse, welche zusätzliches Regelwissen bereitstellen, das in die Benford-Analyse eingehen kann, um Toleranzbereiche festzulegen oder systematisch zu erwartende Abweichungen heraus zu rechnen.

Im letzten Teil dieses Textes soll die Kombination von Benford- und Assoziationsanalyse an einem konkreten Anwendungsbeispiel demonstriert werden: der Analyse der Buchungsdaten auf einem Girokonto.

Beispiel: Analyse der Buchungsdaten auf einem Girokonto

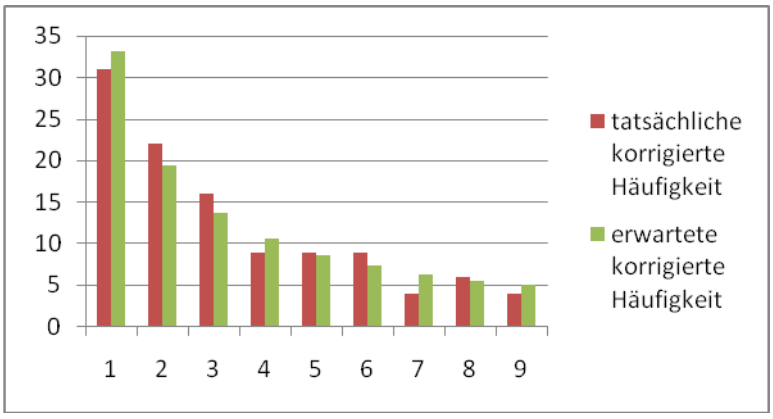
Das nachfolgend dargestellte Diagramm enthält eine Benford-Analyse der führenden Ziffer in den Euro-Werten der Buchungen auf einem Girokonto während eines Quartals (128 Buchungen).



Es wird ersichtlich, dass die Häufigkeitsverteilung insgesamt recht gut der Benford-Verteilung entspricht. Bei der Frage, ob die erhöhten Häufigkeiten der Ziffern 2, 3 und 5 sowie der zu niedrige Wert bei der Ziffer 4 schon eine Auffälligkeit darstellen, hilft das Hinzuziehen von Regeln, welche ein Assoziationsregel-Verfahren durch die gemeinsame Analyse von Kunden-Stammdaten und Buchungs-Statistiken über alle Girokonteninhaber hinweg ermittelt hat. Wenn man nun weiß, dass das im obigen Diagramm analysierte Konto das Gemeinschaftskonto eines Ehepaars mit zwei kleinen Kindern ist, dann kann man in der Sammlung aller gefundenen Regeln die für diese Situation passenden finden. Diese könnten z.B. sein:

- Wer ein Gemeinschaftskonto hat und zwei unter 18-jährige Kinder und in Deutschland wohnt, der hat mit 90%-iger Wahrscheinlichkeit 3 Buchungen ‚Kindergeld 308 €' pro Quartal
- Von Gemeinschaftskonten werden bei Bargeld-Abhebungen über Geldautomaten bevorzugt 100 oder 50 € abgehoben. Dies erhöht die Häufigkeiten der Ziffern 1 und 5 um durchschnittlich je 4 Buchungen pro Quartal.
- Von Gemeinschaftskonten, deren Inhaber 25-50 Jahre alt sind, werden überproportional viele EC-Karten-Zahlungen durch Drogeriemärkte oder Lebensmittelmärkte/Supermärkte in der Größenordnung 20-40€ abgebucht, so dass die Häufigkeiten der Ziffern 2 und 3 im Schnitt das 1.1-fache der Benford-Häufigkeit betragen.

Rechnet man diese Regeln aus der Häufigkeitsverteilung im obigen Schaubild heraus, so ergibt sich die korrigierte Darstellung



Aus dieser korrigierten Darstellung geht hervor, dass es für das betrachtete Konto im betrachteten Quartal keinerlei Hinweise auf ungewöhnliches Buchungsverhalten gibt.

Kontakt:

Dr. Ansgar Dorneich
Fichtenstraße 7
71088 Holzgerlingen
Email: ansgar@dorneich.de