32. GIL-Jahrestagung in Weihenstephan

Benfords Gesetz:
Ein Qualitätstest für statistische Reihen angewendet auf Handelsdaten für Agrarprodukte

Stefan Güttler
Franziska Thiemann
Rolf A.E. Müller

Institut für Agrarökonomie
Christian-Albrechts-Universität zu Kiel

Einleitung

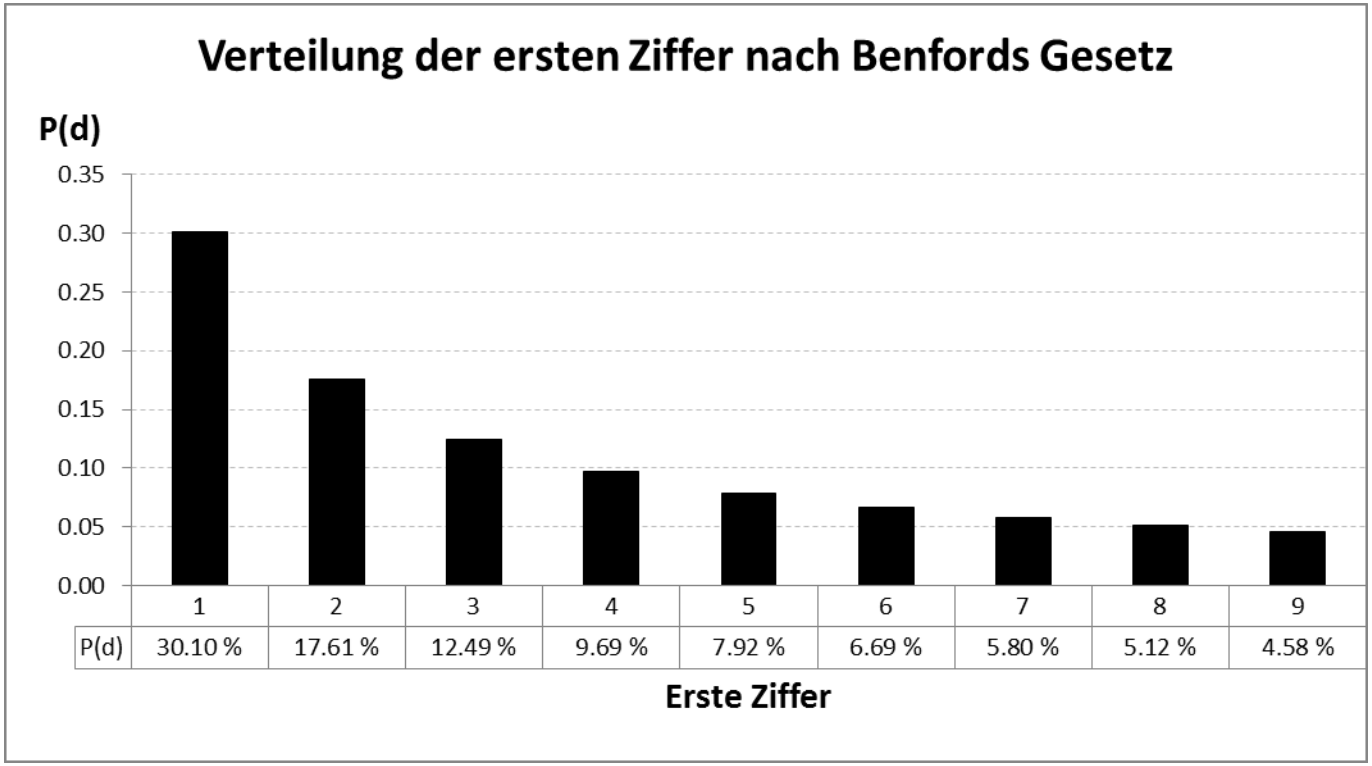
- Statistiken sind Artefakte
- Fehler
  - zufällig
  - systematisch (Bias)
- Datenqualität

→ Überprüfung der Verteilung der ersten Ziffer von Handelsdaten mit Benfords Gesetz

Benfords Gesetz

Verteilung der ersten Ziffer

$$\mathrm{P}(d) = log_{10}(1 + \frac{1}{d}) \text{ für } d = 1, 2, \ldots, 9$$ (Newcomb 1881; Benford 1938)

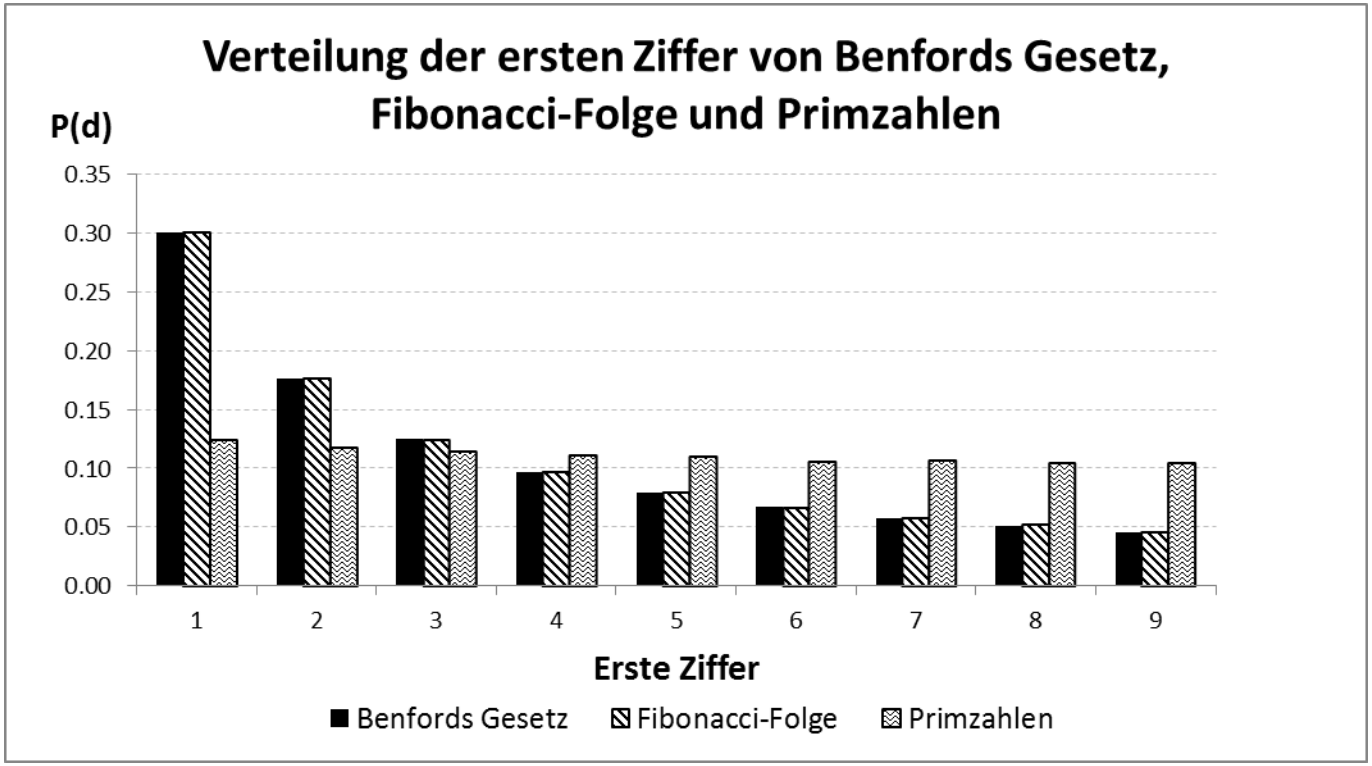


	1	2	3	4	5	6	7	8	9
P(d)	30.10 %	17.61 %	12.49 %	9.69 %	7.92 %	6.69 %	5.80 %	5.12 %	4.58 %

Benfords Gesetz

Verteilung der ersten Ziffer

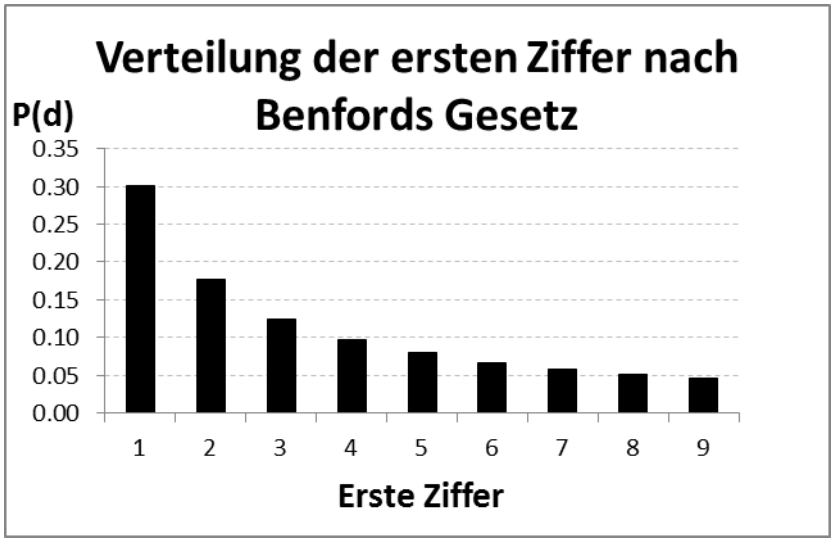
$$\mathrm{P}(d) = log_{10}(1 + \frac{1}{d}) \text{ für } d = 1, 2, \ldots, 9$$ (Newcomb 1881; Benford 1938)



Benfords Gesetz

Verteilung der ersten Ziffer

$$P(d) = log_{10}(1 + \frac{1}{d}) \text{ für } d = 1, 2, \ldots, 9$$ (Newcomb 1881; Benford 1938)

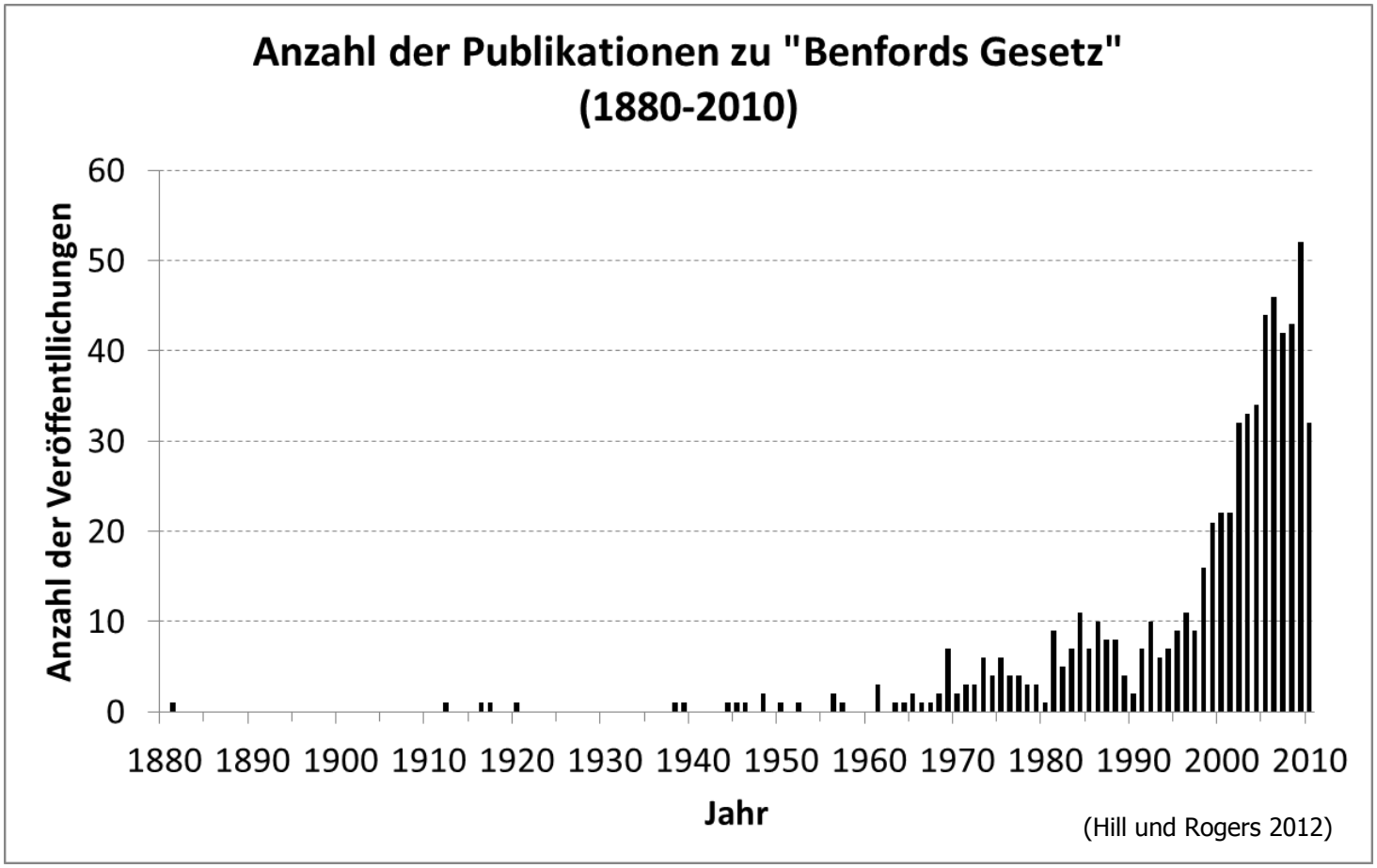


- Eigenschaften (Pinkham 1961; Hill 1995)
  - Baseninvarianz
  - Skaleninvarianz

- Nicht anwendbar bei: (Hill 1995; Nigrini und Mittermaier 1997; Durtschi et al. 2004)
  - zugeordneten Zahlen, z.B. Bestellnummern
  - psychologisch beeinflussten Zahlen, z.B. Preise im LEH

Benfords Gesetz

- Anzahl der Publikationen steigt



Benfords Gesetz

- Anzahl der Publikationen steigt
- Anwendung
  - betrügerische Datenmanipulation (Steuer- oder Bilanzfälschung) (Nigrini 1996, 1999; Durtschi et al. 2004)
  - Plausibilität von Prognosemodellen (Ley 1996; Tödter 2007)
  - publizierte Analyseergebnisse (Diekmann 2007)
  - Umfragedaten (Judge und Schechter 2009; Schräpler 2010)
- Tests
  - visuelle Vergleiche
  - Chi-Quadrat-Test
  - z-Statistiken, K-S-Test, Bayes‘sche Methoden

Datenauswahl

- Datenbank: UN Comtrade
- Warenklassifikation: SITC Rev. 3

- 20 ausgewählte Agrar- u. Ernährungsprodukte
- Zeitraum: 1995-2009
- > 90% der globalen Exporte erfasst
- Handelswert (in US$)
- bilaterale Handelsdaten

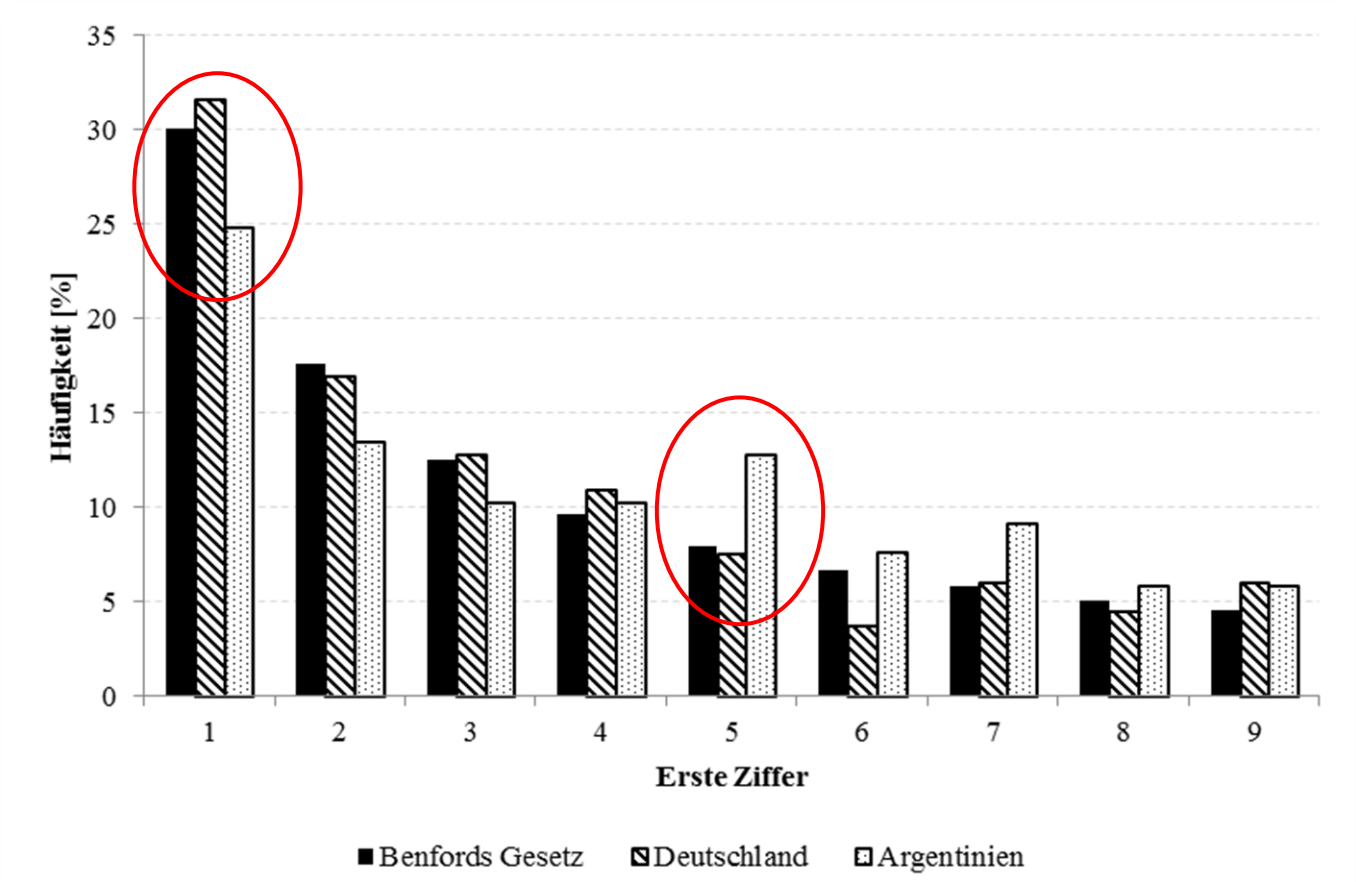
Ergebnisse: Agrar- & Ernährungsprodukte

SITC-Code	Warengruppe	N	Jahr														
		(95-09)	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	**95-09**
0	Nahrungsmittel	16695							10									**5**
01112	Rindfleisch	3180																
034	Fisch	12813							10					10				**5**
04	Getreide	15192	10					5					1					
041	Weizen	2974			5								10					**10**
042	Reis	6390			5													
0451	Roggen	1471											10	5				
05	Gemüse & Früchte	16191											5					
0544	Tomaten	2815																
05711	Orangen	4819				5									1			
0573	Bananen	4016												10				
059	Fruchtsäfte	7294						5							10			
1	Getränke & Tabak	14979			10													
11	Getränke	9495																
11101	Wasser	2656																
1121	Wein	4336		1	5							5						
1123	Bier	6932		10				1							1			**10**
2	pflanzl. Rohstoffe	13468															10	
263	Baumwolle	5945					5											
29271	Schnittblumen	3556								1		5			10			**5**

1, 5, 10: Signifikanzniveau des Chi-Quadrat-Tests (in %)

Ergebnisse: Länder & Agrarprodukte

Verteilung der ersten Ziffer der Weizenexportdaten von Deutschland und Argentinien im Vergleich zu Benfords Gesetz, 1995-2009



Ergebnisse: Länder & Agrarprodukte

	Warengruppe																		
Exporteure	Nahrungsmittel	Rindfleisch	Fisch	Getreide	Weizen	Reis	Roggen	Gemüse & Früchte	Tomaten	Orangen	Bananen	Frucht- & Gemüsesäfte	Getränke und Tabak	Getränke	Wasser	Wein	Bier	pflanzl. Rohstoffe	Baumwolle	Schnittblumen
Argentinien					1		.		.		.	5	10		.		.			.
Australien	1	1	10	10				1	.		.				.		.	1		.
Brasilien			.		.	.	.		.		1			.	.	.	.			.
China		.			.		.		.	5	.			1	1	.				.
Deutschland	1			5		.	10								5		5		1	5
Frankreich	1	1						1				10			1	1			.	.
Großbritannien	.						.	1	.	.	.	.	5		5			1	.	.
Italien					.	5			5	10								1	.	
Kanada	1	10	5				5			.	.					.	1		.	.
Mexiko	1	.	.		.		.	5	5	.						.				10
Niederlande	1			10	.			.				5				.	1		.	1
Russland		.	1					.	.	.	.	.		.	.	.			.	.
Spanien	5	.			.			5				5			.	1		5		
USA	1			10	10		.	10		5				10				5	5	1

1, 5, 10: Signifikanzniveau des Chi-Quadrat-Tests (in %)

". ": Datensatz nicht untersucht, da das Land nicht zu den größten Exporteuren gehört

" " (leere Zelle): Datensatz unterscheidet sich nicht signifikant von Benford`s Gesetz

Zusammenfassung & Diskussion

- Ergebnisse: keine systematische Verzerrung
  - aber: möglicherweise Hinweis auf atypische Prozesse in der Datenentstehung (Posch 2010)

- Gründe fehlerhafter Handelsdaten (UN 2004)
  - Fehlklassifikation
  - Unterschiede bei Datenerhebung und -verarbeitung
  - Bewertung (Wechselkurse, cif - fob)
  - kriminelle Aktivitäten (Schmuggel)
  - bewusste Manipulation

Ausblick & Fazit

- Import- und Produktionsdaten
- weitere Datenquellen
- weitere Teststatistiken berechnen

Benfords Gesetz:

→ einfache Überprüfung der Datenqualität

→ nützliches Tool zur Identifizierung fehlerverdächtiger Datenreihen zur weiteren Analyse

→ Datenqualität vor weiteren Analysen prüfen!

Vielen Dank!

Stefan Güttler
Franziska Thiemann
Rolf A.E. Müller

Institut für Agrarökonomie
Abteilung Innovation & Information

Christian-Albrechts-Universität zu Kiel

E-Mail: stefan.guettler@ae.uni-kiel.de

Backup: Benfords Daten (1938)

TABLE 1: *The distribution of leading digits in Benford's data sets in percentages (Benford 1938)*

Group	Title	1	2	3	4	5	6	7	8	9	Count
A	Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
B	Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3,259
C	Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
D	Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
E	Spec. Heat	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1,389
F	Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
G	H.P.Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
H	Mol. Weight	27.7	25.3	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1,800
I	Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
J	Atomic Wgt.	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
K	n^{-1}, $\sqrt{n}$,...	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5,000
L	Design	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
M	Digest	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
N	Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
O	X-Ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
P	Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1,458
Q	Black Body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1,165
R	Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
S	$n^1, n^2, \ldots, n!$	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
T	Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
	Average	30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1,011
	Predicted	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6	

Backup: Generalisierung von Benford's Gesetz

Erweiterung auf alle weiteren Ziffernstellen:

$$P(D_1 = d_1, \dots, D_k = d_k) = log_{10}\left[1 + \left(\sum_{i=1}^{k} d_i \times 10^{k-i}\right)^{-1}\right],$$

für d_1=1,2,…9; d_j=0,1,2,…,9; j=2,…,k,
wobei d_1 die erste Ziffer darstellt und d_j für die zweite bis k-te Ziffer einer Zahl steht.

Die Wahrscheinlichkeit für das Auftreten der Zahl 189 ist danach:

$$P(D_1 = 1, D_2 = 8, D_3 = 9) = log_{10}[1 + (189)^{-1}] \cong 0{,}0023.$$

Backup: Tests auf signifikante Abweichung

Chi-Quadrat-Test $n\sum_{i=1}^{9}\frac{(e_i-b_i)^2}{b_i}$, für i = 1, 2,…, 9

Kuiper-Test (Modifikation des K-S-Test):

$$V_n^* = V_n\left(n^{1/2} + 0.155 + 0.24n^{-1/2}\right)$$

wobei $V_n = \max_x[F_e(x) - F_b(x)] + \max_x[F_b(x) - F_e(x)]$

m-Test $m^* = n^{1/2} \times \max_{i=1,2,\ldots,9}\{|b_i - e_i|\}$

b_i: erwartete Wahrscheinlichkeit durch Benford's Gesetz

e_i: empirisch beobachtete relative Häufigkeit

Test statistic	Test level		
	α=0.1	α=0.05	α=0.01
Chi-square test (χ^2)	13.36	15.51	20.09
Kuiper test (V_n^*)	1.19	1.32	1.58
m test (m^*)	0.85	0.97	1.21

Exporteure	Warengruppe																			
	Nahrungsmittel	Rindfleisch	Fisch	Getreide	Weizen	Reis	Roggen	Gemüse & Früchte	Tomaten	Orangen	Bananen	Frucht- & Gemüsesäfte	Getränke und Tabak	Getränke	Wasser	Wein	Bier	pflanzl. Rohstoffe	Baumwolle	Schnittblumen
Argentinien					1		.		.		.	5	10		.		.			.
Australien	1	1	10	10				1	.		.				.		.	1		.
Belgien					.		.	10	10		5			10		.			.	10
Belgien-Luxemburg					.		.		5						5	.			.	
Brasilien			.		.	.	.		.		1			.	.	.	.			.
China		.			.		.		.	5	.			1	1	.				.
Dänemark		10	5		.	.		.	.	.	.	.			.	.	1		.	.
Deutschland	1			5		.	10								5		5		1	5
Finnland	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			.	.
Frankreich	1	1						1				10			1	1			.	.
Griechenland	.	.			.	.	.	1	.	5	.	.	10		.	.	.	.		.
Großbritannien	.						.	1	.	.	.	.	5		5			1	.	.
Indien		.	5		.	5	.		.	.	.	.	.	.	.	.	.		10	.
Irland			5	.	.		.	.	.	.	.		1	1		.		.	.	.
Italien					.	5			5	10								1	.	
Kanada	1	10	5				5			.	.					.	1		.	.
Luxemburg					.	1	.					10			5	.		5	.	
Mexiko	1	.	.		.		.	5	5	.						.				10
Niederlande	1			10	.			.				5				.	1		.	1
Österreich		.	.		.	.	5		.	.	.				.	.			.	.
Portugal	.	.		.	.	.	.	.	.	.		.	1	10	.		1	.	.	.
Russland		.	1					.	.	.	.	.		.	.	.			.	.
Schweden	.	.		5	.	.		.	.	.	10	.			.	.	.		.	.
Spanien	5	.			.			5				5			.	1		5		
USA	1			10	10		.	10		5				10				5	5	1

"1, 5, 10": Signifikanzniveau (in %)

". ": Datensatz wurde nicht untersucht, da das Land nicht zu den größten Exporteuren gehört

" " (leere Zelle): Datensatz unterscheidet sich nicht signifikant von Benford`s Gesetz