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
Olshausenstr. 40
24118 Kiel
stefan.guettler@ae.uni-kiel.de
fthiema@ae.uni-kiel.de
raem@ae.uni-kiel.de

Abstract: Die Qualität von Exporthandelsdaten ausgewählter Agrarprodukte wurde mit Benfords Gesetz überprüft. Die Ergebnisse zeigen, dass einige Daten von Produkten und Ländern statistisch signifikant von Benfords Gesetz abweichen. Handelsdaten sollten daher nicht ungeprüft in weitere Analysen eingehen.

1. Einleitung

Statistische Daten sind von Menschen produzierte Artefakte. Wie bei allen menschlichen Produkten ist die Qualität statistischer Daten nicht homogen, manche Statistiken sind fehlerhaft und andere wurden bewusst verfälscht [Mo50]. Dem Qualitätsmanagement in der Statistik stehen grundsätzlich zwei Gruppen von Maßnahmen zur Qualitätssicherung zur Verfügung: Maßnahmen zur Sicherung der Prozessqualität und Maßnahmen zur Überprüfung der Produktqualität. Hier befassen wir uns mit Benfords Gesetz, einem Ansatz zur Überprüfung der Qualität statistischer Daten.

Die Zahl der Anwendungen von Benfords Gesetz, das wir im nächsten Abschnitt vorstellen, ist in letzter Zeit rasch gewachsen. Die Anwendungsbereiche reichen vom Versuch der Aufdeckung von betrügerischen Datenmanipulationen [Ni96] [Ni99] [DHP04], über die Überprüfung der Plausibilität mathematischer Modelle [Le96] [Tö07] und publizierter Analyseergebnisse [Di07] bis hin zur Überprüfung der Vertrauens-würdigkeit von Umfragedaten [Sc10]. Allerdings sind uns keine Anwendungen zur Überprüfung der Qualität landwirtschaftlicher statistischer Daten bekannt.

In unserer Untersuchung haben wir Benfords Gesetz auf die Exporthandelswerte von 20 Agrarprodukten verschiedener Klassifizierungsebenen des SITC-Codes angewendet. Die ausgewählten UN Comtrade Datensätze decken mindestens 90% des Welthandels des entsprechenden Produktes ab, umfassen den Zeitraum von 1995-2009 und bestehen im Mittel aus ca. 7.700 Beobachtungen.

2. Benfords Gesetz

Benfords Gesetz wurde, wie viele eponyme Gesetze, nicht nach seinem ursprünglichen Entdecker - das war Newcomb [Ne81] - sondern nach seinem Zweitentdecker Frank Benford [Be38] benannt, der 1938 gezeigt hat, dass die Wahrscheinlichkeit der ersten Ziffer *d* der Daten in einem Datensatz einem einfachen logarithmischen Gesetz folgt:

$$P(d)=log_{10}(1+1/d), \qquad \text{für } d=1,2,\ldots,9$$

Nach Benfords Gesetz ist die Wahrscheinlichkeit P(1), dass eine zufällig gezogenen Zahl eines Datensatzes mit der Ziffer „1“ beginnt, P(1) = 0,301; diese Wahrscheinlichkeit fällt monoton bis zu P(9) = 0,046. Hill [Hi95] [Hi99] hat Benfords Gesetz für die nachfolgenden Ziffern erweitert. Benfords Gesetz ist dagegen nicht anwendbar auf zugeordnete Zahlen, wie z.B. Bestellnummern, und psychologisch beeinflusste Zahlen, wie z.B. Preise im Lebensmitteleinzelhandel [Hi95] [DHP04] [NM97].

Methoden, die zur Überprüfung der empirischen relativen Häufigkeiten der ersten Ziffern mit den entsprechend Benfords Gesetz erwarteten Wahrscheinlichkeiten verwendet werden, sind visuellen Vergleiche, z-Statistiken, Chi-Quadrat-Tests und Bayes’sche Methoden. In unserer Analyse haben wir die Beurteilung anhand des Chi-Quadrat-Tests vorgenommen. Bei der Ergebnisinterpretation muss der Einfluss der Stichprobengröße *N* auf den Prüfwert des Tests berücksichtigt werden, da bei großen Stichproben bereits kleine Abweichungen statistisch signifikant sein können [Sc10].

3. Datenauswahl und Ergebnisse

Die Datensätze dieser Untersuchung stammen aus der UN Comtrade Datenbank (http://comtrade.un.org), die über 1,7 Mrd. Datensätze ab dem Jahr 1962 in verschiedenen Warenklassifikationen enthält. Die ausgewählten Exportdaten der Agrarprodukte umfassen den Zeitraum von 1995-2009 und stammen aus der SITC (Revision 3)-Klassifikation. Die Daten decken mindestens 90% des Gesamthandels in der entsprechenden Produktkategorie durch die Länderauswahl ab.

Die Ergebnisse in Tabelle 1 zeigen, dass die Exportdatensätze für Nahrungsmittel, Fisch, Weizen, Bier und Schnittblumen signifikant von der Benford Verteilung abweichen. Dagegen entspricht die Verteilung der ersten Ziffer in den Handelsdaten von Rindfleisch, Tomaten, Getränken und Wasser in jedem Jahr der Benford Verteilung. Für andere Produkte konnte in einigen, aber nicht in allen Jahren, eine signifikante Abweichung von Benfords Gesetz festgestellt werden. Bei einigen untersuchten Produkten besteht also der Verdacht, dass diese Daten verfälscht worden sein könnten. Betrachtet man jedoch die Exportstatistiken der einzelnen Länder, lassen sich keine systematischen Abweichungen erkennen. Die Ergebnisse für die Auswertung der Exportdaten nach Ländern und Warengruppen in Tabelle 2 zeigen ebenfalls, dass signifikante Abweichungen von Benfords Gesetz für eine Vielzahl der untersuchten Datensätze vorliegen.

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**

Tabelle 1: Signifikanzniveaus des Chi-Quadrat-Tests beim Test auf Abweichung von Benfords Gesetz für Exportstatistiken ausgewählter Agrarprodukte, 1995-2009 [in %]

	SITC-Code																				
Exporteure	0	1112	34	4	41	42	451	5	544	5711	573	59	1	11	11101	1121	1123	2	263	29271
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 (in %)

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

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

Tabelle 2: Signifikanzniveaus des Chi-Quadrat-Tests beim Test auf Abweichung von Benfords Gesetz für die Exportstatistiken ausgewählter Länder und Agrarprodukte [in%]

4. Diskussion und Zusammenfassung

Benfords Gesetz kann schnell einen Überblick über die Qualität von Datensätzen schaffen. Neben der Möglichkeit, Daten mit frei verfügbarer Software zu überprüfen (www.checkyourdata.com) ermöglichen dies auch viele Statistikprogramme wie Stata, R, SAS, Python und Excel (kostenpflichtiges Makro: DATAS 2009).

Es gibt viele Gründe, warum Handelsdaten fehlerhaft oder verfälscht sein können; dazu zählen unter anderem: (i) Fehlklassifikation, (ii) Unterschiede bei der Datenerhebung und im Umgang mit den Daten, (iii) volatile Wechselkursraten, (iv) Erfassung bzw. Nichterfassung von Transithandel, (v) kriminelle Aktivitäten wie Schmuggel und (vi) bewusste Manipulation um z.B. Steuervorteile zu erzielen [UN04]. Benfords Gesetz lässt jedoch keine Schlussfolgerungen über die Ursachen von Datenunvollkommenheiten zu.

Die Ergebnisse zeigen, dass viele Handelsdaten von Agrarprodukten von Benfords Gesetz abweichen. Es empfiehlt sich daher, Datensätze vor einer aufwendigen ökonometrischen Analyse mit Benfords Gesetz auf ihre Qualität zu überprüfen, wenn man dem Rubbish-in – Rubbish-Out-Syndrom der statistischen Datenanalyse entgehen möchte.

Literaturverzeichnis

[Be38] Benford, F.: The law of anomalous numbers. In: Proceedings of the American Philosophical Society 78: 551 – 572, 1938.

[DHP04] Durtschi, C.; Hillison, W.; Pacini, C.: The effective use of Benford's law to assist in detecting fraud in accounting data. In: Journal of Forensic Accounting 5: 17-34, 2004.

[Di07] Diekmann, A.: Not the first digit! Using Benford's law to detect fraudulent scientific data. In: Journal of Applied Statistics 34 (3): 321-329, 2007.

[Hi95] Hill, T.P. (1995): A statistical derivation of the significant-digit law. Statistical Science 10 (4): 354-363.

[Hi99] Hill, T.P.: The difficulty of faking data. In: Chance 26: 8-13, 1999.

[Le96] Ley, E.: On the peculiar distribution of the U.S. stock indexes' digits. In: The American Statistician 50: 311 – 313, 1996.

[Mo50] Morgenstern, O.: On the accuracy of economic observations. Princeton University Press, Princeton, 1950.

[Ne81] Newcomb, S.: Note on the frequency of use of the different digits in natural numbers. In: American Journal of Mathematics 4 (1): 39-40, 1881.

[Ni96] Nigrini, M.J.: A taxpayer compliance application of Benford's law. In: The Journal of the American Taxation Association 18: 72 – 91, 1996.

[Ni99] Nigrini, M.J: I've got your number. In: Journal of Accountancy 187 (5): 79-83, 1999

[NM97] Nigrini, M.J.; Mittermaier, L.J.: The use of Benford's law as an aid in analytical procedures. In: Auditing: A Journal of Practice & Theory 16 (2): 52–67, 1997.

[Sc10] Schräpler, J.P.: Benford's law as an instrument for fraud detection in surveys using the data of the Socio-Economic Panel (SOEP). SOEPpapers on Multidisciplinary Panel Data Research at DIW Berlin 273, February 2010, Berlin.

[Tö07] Tödter, K.H.: Das Benford-Gesetz und die Anfangsziffern von Aktienkursen. In: Wirtschaftswissenschaftliches Studium 36(2): 93 – 97, 2007.

[UN04] UN: International Merchandise Trade Statistics – Compilers Manual. United Nations Publication, New York, 2004.