INSTITUT FÜR TECHNIKFOLGEN-ABSCHÄTZUNG

ITA manu:script

Diagnose von Fehlerquellen und methodische Qualität in der sozialwissenschaftlichen Forschung

Andreas Diekmann

http://www.oeaw.ac.at/ita/pdf/ita_02_04.pdf

ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

Wien, 06/2002
ITA-02-04
ISSN 1681-9187

http://www.oeaw.ac.at/ita/pdf/ita_02_04.pdf

Juni/2002
(ITA-02-04)

Diagnose von Fehlerquellen und methodische Qualität in der sozialwissenschaftlichen Forschung

Andreas Diekmann

Institut für Soziologie der Universität Bern, E-Mail: diekmann@soz.unibe.ch

Keywords

Methodenprobleme, Gütekriterien, Fehlerquellen, Ausschöpfungsquote, Datenfälschung, Benford-Gesetz

Abstract

Umfragen liefern heute das Zahlenmaterial für wirtschaftliche und administrative Entscheidungen sowie sozialwissenschaftliche Untersuchungen. Fehler können dabei teuer zu stehen kommen. Der per Befragung erhobene IFO-Geschäftsklima-Index zum Beispiel hat nach Bekanntgabe unmittelbaren Einfluss auf den Kurs des Euro. Media-Analysen entscheiden über die Verteilung umfangreicher Werbebudgets. Und Irrtümer bei der Ermittlung des Preisindex wirken sich auf zahlreiche Verträge aus, die direkt oder indirekt mit dem Index verkoppelt sind. Die Sicherung methodischer Qualität und Transparenz bezüglich der Gütekriterien von Erhebungen ist daher eine nachdrückliche Forderung.

In dem Vortrag werden Fehlerquellen bei der Befragung, der Stichprobenziehung und der Datenanalyse aufgezeigt sowie Möglichkeiten zur Behebung zumindest einiger typischer Fehlerquellen vorgeschlagen. Dabei wird auch das Thema betrügerischer Datenmanipulationen angesprochen. Zur Diagnose von Datenfälschungen wurden in jüngster Zeit Testverfahren entwickelt, die sich allerdings noch in der Erprobungsphase befinden. Diese Verfahren und ihre Anwendungsmöglichkeiten werden abschließend vorgestellt.

ISSN 1681-9187

Inhalt

Dieses ITA-manu:script ist die überarbeitete Fassung eines öffentlichen Vortrags, den der Verfasser am 19.2.2002 am ITA gehalten hat.

Für Auswertungen mit der Schweizerischen Arbeitskräfteerhebung 1991 und mit dem Mikrozensus Familie 1994/95 bedanke ich mich bei Ben Jann und Kurt Schmidheiny.

Einleitung

Umfragen liefern heute das Zahlenmaterial für wirtschaftliche Entscheidungen, administrative Planung und für sozialwissenschaftliche Untersuchungen. Die Resultate haben oftmals weitreichende soziale und wirtschaftliche Konsequenzen. Media-Analysen zur Reichweite von Zeitungen und Zeitschriften z. B. entscheiden über die Verteilung umfangreicher Werbebudgets.

Der Geschäftsklima-Index des Münchner ifo-Instituts für Wirtschaftsforschung basiert auf einer schriftlichen Befragung mit dreizehn Fragen, die einer Stichprobe von 7.000 Unternehmen im monatlichen Rhythmus vorgelegt werden. Am 20. September 2.000 berichtete die deutsche Presseagentur dpa, dass der deutsche Aktienindex u. a. infolge der neuesten ifo-Zahlen deutliche Einbußen erlitten hatte. Der ifo-Index war im Vergleich zum Vormonat von 99,1 auf 99 Punkte gefallen. Der erneute Rückgang des Index hatte milliardenschwere Auswirkungen auf die Devisen- und Aktienmärkte.

Wenn Umfrageergebnisse teure Folgen haben, können Umfragefehler teuer zu stehen kommen. Deshalb ist es eine wichtige Aufgabe, die methodische Qualität von Umfragen zu sichern und Transparenz darüber herzustellen, inwieweit Umfragen die methodischen Gütekriterien erfüllen.

Zunächst werde ich auf einige Qualitätsmerkmale sozialwissenschaftlicher Erhebungen eingehen. Sodann werde ich mich mit einem Thema befassen, das gerne vernachlässigt und verdrängt wird: dem Problem der Fälschung von Daten.

1 Gütekriterien von Umfragen

Als Verbraucher wird man heute über alle möglichen Produkteigenschaften aufgeklärt. Jeder Beipackzettel einer Kopfschmerztablette enthält eine lange Liste von Risiken und Nebenwirkungen. Auf der Verpackung eines Joghurtbechers werden die Inhaltsstoffe in der Terminologie der Lebensmittelchemie minutiös aufgelistet. Bei Umfrageergebnissen erhält man dagegen, oftmals selbst als Auftraggeber, nur spärliche Informationen über die Qualität des Produkts. In Veröffentlichungen heißt es häufig lapidar: Repräsentativumfrage mit z. B. 1.200 Befragten. Dann werden in der Regel noch der Fragetext und die Ergebnisse mitgeteilt.

Was ist eine Repräsentativumfrage? In der Statistik gibt es streng genommen überhaupt keine Repräsentativumfragen. Eine Stichprobe kann niemals in allen Merkmalen ein repräsentatives Abbild der Population sein. Allenfalls kann man sagen, dass „repräsentativ“ ein bildhaftes Kürzel für bestimmte Verfahren der Stichprobenziehung darstellt. Bislang gibt es aber keine in der Profession allgemein akzeptierte Definition von Repräsentativumfragen.

Wichtiger als das schwammige Etikett „Repräsentativumfrage“ ist die Angabe der zentralen Merkmale der Erhebungsprozedur. Nur anhand solcher Kriterien kann die Qualität von Umfragedaten beurteilt werden. Wie wir noch anhand von Beispielen sehen werden, steckt dabei das Problem oftmals im Detail.

Ohne Anspruch auf Vollständigkeit sei hier eine Liste wichtiger Gesichtspunkte bzw. Fragen zur Beurteilung der Qualität der Umfragemethodik aufgeführt:

1. *Art der Stichprobenziehung.* Zufallsstichprobe oder anderes Verfahren? Wie genau sieht der Stichprobenplan aus?
2. *Bei Zufallsstichproben* und Face-to-Face-Interviews: Auswahl der Haushalte per Random Route (der Interviewer ermittelt den Haushalt selbst nach vorgegebenen Begehungsregeln) oder mittels Adressrandom (der Interviewer erhält vorgegebene Haushaltsadressen, die zuvor per Zufallsauswahl ermittelt wurden). Wie erfolgt die Auswahl der Zielperson im Haushalt? Wird die zu befragende Person im Haushalt zufällig ausgewählt, z. B. nach der „Geburtstagsmethode" oder per „Schwedenschlüssel"[1]?
3. Umfang der Stichprobe.
4. *Ausschöpfungsquote.* Wie hoch ist die Ausschöpfungsquote und nach welchem Schema wurde sie berechnet?
5. *Erhebungsmethode.* Face-to-face, telefonisch, schriftlich. Mit oder ohne Computerunterstützung (CAPI, CATI)?
6. *Interviewer.* Wie setzt sich der Interviewerstab zusammen und wie wurden die Interviewer und Interviewerinnen geschult? Wie werden die Interviewer bezahlt?
7. *Feldkontrolle.* Wie erfolgt die Kontrolle der Interviewer und wie hoch ist der Anteil kontrollierter Interviews?
8. *Pretest.* Gab es mindestens einen oder mehrere Pretests?
9. *Omnibuserhebung.* Handelt es sich um eine eigenständige Umfrage oder stammt das Ergebnis aus einer „Omnibus-Erhebung"? Wenn ja, wo wurden die Fragen platziert?
10. *Reliabilität und Validität* der Fragen bzw. Indizes.
11. *Gewichtung.* Basieren die Ergebnisse auf gewichteten Daten? Wenn ja, erfolgte die Gewichtung mit Designgewichten gemäß Stichprobenplan und Stichprobentheorie der Statistik? Oder wurden einfach demographische Merkmale an bekannte Randverteilungen angepasst (Nachgewichtung oder so genanntes Redressment)? Redressment ist höchst umstritten, basiert nicht auf statistischer Theorie, wird aber dennoch fast überall praktiziert.
12. *Datenanalyse.* Wurden die dem Forschungsproblem und der Datenqualität angemessenen statistischen Verfahren angewandt? Wie wurde mit fehlenden Werten („missing values") umgegangen? Wurden die Schätzungen korrekt interpretiert?

Welche Probleme bei der Beurteilung der Datenqualität auftreten können und welche oftmals vernachlässigten Detailfragen dabei eine Rolle spielen, möchte ich jetzt an einigen Beispielen illustrieren.

[1] Bei der „Geburtstagsmethode" wird das zur Grundgesamtheit zählende Haushaltsmitglied befragt, das zuletzt Geburtstag hatte. Der „Schwedenschlüssel" oder „Kish-Auswahlschlüssel" ist eine auf den Fragebogen gedruckte Zufallszahlenkombination, die dem Interviewer eindeutig vorgibt, welche Person im Haushalt befragt werden soll. Vgl. z. B. Diekmann (2001).

Beispiel 1: Größe kann täuschen!

Nach der schweizerischen Volkszählung von 1990 beträgt der Anteil teilzeitarbeitender Erwerbstätiger an allen Erwerbstätigen 19 Prozent. Mit der Stichprobe der schweizerischen Arbeitskräfteerhebung (SAKE, Zufallsstichprobe von ca. 16.000 telefonisch befragten Personen) ergab sich 1991 bei gleicher Definition eine Teilzeitquote von 26 Prozent. Sieben Prozentpunkte Differenz bei einer zentralen Arbeitsmarktstatistik ist schon eine bedenkliche Unschärfe. Geht man nach dem Umfang der Erhebung – hier Vollerhebung versus Stichprobe – würde man dem Ergebnis der Volkszählung Glauben schenken. Buhmann et al. (1994) vom Bundesamt für Statistik haben nun mit Hilfe weiterer Datenquellen die Unterschiede zwischen den verschiedenen Erhebungen genauer unter die Lupe genommen. Ihr Befund lautet, dass die Anzahl von Personen mit geringer Erwerbstätigkeit in der Volkszählung unterschätzt wurde. Die Fehlerquelle sei die Selbstdeklaration im Volkszählungsfragebogen. Geringfügig erwerbstätige Personen stufen sich selbst oftmals als nicht erwerbstätig ein. Ob dies der Grund für die Verzerrung ist, sei hier dahingestellt. (Im Schweizerischen Arbeitsmarktsurvey hatten wir eine ähnliche Frage wie in der Volkszählung gestellt, ohne dass sich eine gravierende Abweichung zum SAKE-Resultat ergab.) Gehen wir aber einmal davon aus, dass die Erklärung von Buhmann et al. zutrifft und dass die SAKE die tatsächliche Teilzeitquote besser trifft als die Volkserzählung. Es zeigt sich dann wieder einmal, dass Größe täuschen kann (wie in Abbildung 1 auf der nächsten Seite).

Wichtiger als die Aufstockung von Stichproben ist die Vermeidung von gravierenden Verzerrungen. Häufig sind selektive Stichproben die Ursache für eine verzerrte Schätzung. Aber auch die Frageformulierung kann der Grund für einen „Bias“ sein. Man spart jedenfalls am falschen Ende, wenn man auf sehr genaue Tests der Fragen in eventuell mehreren Pretests verzichtet.

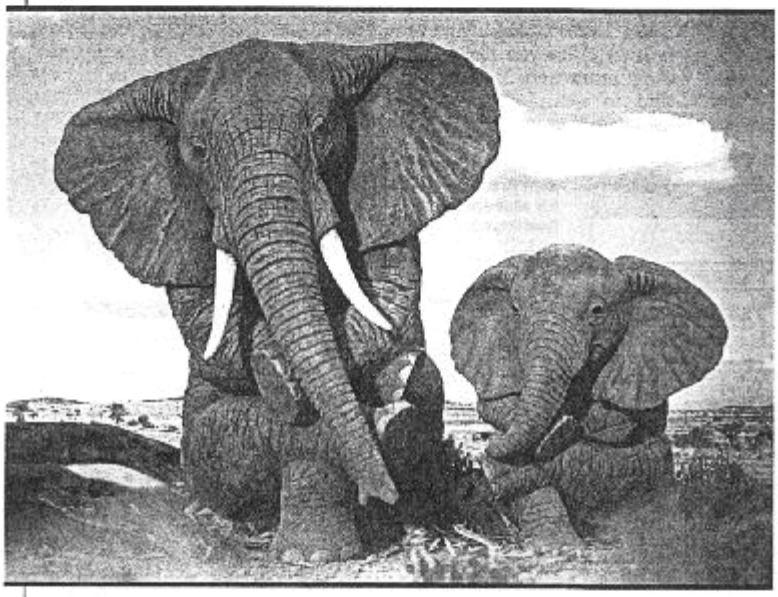
Wenn ein systematischer Fehler vorliegt, kann auch eine noch so große Stichprobe nicht für die Verzerrung kompensieren. Vergrößert man den Stichprobenumfang, trifft man den falschen Wert sozusagen mit größerer Genauigkeit. Eine Erhöhung des Stichprobenumfangs bewirkt bei einem starken Bias nur, dass sich die Masse der Stichprobenverteilung enger um den falschen Wert konzentriert. Bei einem knappen Budget ist man also gut beraten, einen Teil der Mittel zur Abhilfe von Verzerrungen zu investieren (z. B. durch eine Steigerung der Ausschöpfungsquote; dazu weiter unten), statt die Fallzahl zu maximieren.

GET YOUR MIND WORKING WITH PARIBAS

Zwei Elefanten sitzen auf einem Baumstamm.
Der kleine Elefant ist der Sohn des großen.
Aber der große Elefant ist nicht sein Vater.
Wie ist das möglich?

Fig. 1
"Wer ist mein Vater, wenn Du's nicht bist?"

Fig. 2
"Ja, aber.. wer bin ich dann?"



PARIBAS
EINE KOMPETENZ IM GLOBALEN CORPORATE BANKING

Aktiva
US-$ 290 Milliarden

Eigenmittel
US-$ 12 Milliarden

70% der Erträge außerhalb Frankreichs erwirtschaftet

GRÖSSE KANN TÄUSCHEN

Die oben gestellte Frage zeigt, wie sehr gewisse Formulierungen und Vorurteile uns zu irrtümlichen Schlüssen führen können. Speziell das Corporate Banking ist oft Opfer falscher Vorstellungen.

Nur wenigen ist zum Beispiel bekannt, daß Paribas mit Aktiva von über 290 Milliarden US-Dollar eine der drei größten Banken in der internationalen Rohstoff- und Handelsfinanzierung ist und in Europa den zweiten Platz in der Finanzierung des Medienselctors einnimmt.

Über zwei Drittel unserer Geschäftsaktivitäten konzentrieren sich auf strukturierte Finanzierungen und betreffen Wachstumssektoren wie internationale Großprojekte und Export, Luft- und Raumfahrt sowie Gesundheitswesen.

Doch lassen sich viele, die mit dem Bankwesen nicht allzu vertraut sind, leicht von bekannteren Namen beeindrucken. Oft zu Unrecht.

Womit wir wieder bei unserem Baumstamm mit den beiden Elefanten angelangt wären. Die Moral von der Geschichte: "Hüte Dich vor falschen Annahmen".

Wir gehen nämlich davon aus, daß der große Elefant männlich sein muß. Ist er aber gar nicht. Ganz einfach: Der große Elefant ist die Mutter des kleinen. http://www.paribas.com

PARIBAS Thinking beyond banking

Abbildung 1: „Größe kann täuschen“

Beispiel 2: Ausschöpfungsquote schön gerechnet!

Lehrbuchgerechte Zufallsstichproben hat man in der Praxis bekanntlich selten oder nie. Für landesweite Befragungen liegen die Ausschöpfungsquoten etwa im Bereich von 50 bis 70 Prozent. Je niedriger die Ausschöpfungsquote, desto größer ist – abhängig vom untersuchten Merkmal – die Gefahr eines Stichprobenselektionsfehlers und damit die Gefahr mehr oder minder verzerrter Schätzungen. Deshalb möchte man bei Umfragen natürlich eine möglichst hohe Ausschöpfungsquote erreichen.

Tabelle 1: Die Berechnung der Ausschöpfungsquote in der Praxis

Bruttostichprobe	8.218	100 %
Kein Anschluss unter der Nummer	361	
Modem/Fax/Natel im Tunnel	28	
Telefonbeantworter	218	
Andere technische Probleme	89	
Kein Privathaushalt	294	
Gehört nicht zur Grundgesamtheit	1.410	
Sprachproblem	291	
Nicht verfügbar: Krankheit	265	
Nicht verfügbar: abwesend	358	
Zielperson nicht erreichbar	82	
Stichprobenneutrale Ausfälle insgesamt	3.396	41,32 %
Bereinigter Stichprobenansatz	4.822	100 %
kein Interesse	718	
keine Auskunft zum Thema	166	
keine Zeit	383	
keine Auskunft am Telefon	233	
Hörer aufgelegt	145	
Andere Verweigerung	146	
Kein Kontakt nach 99 Versuchen	12	
Systematische Ausfälle insgesamt	1.803	37,39 %
Durchgeführte Interviews und Ausschöpfungsquote	3.019	62,60 %

Das Beispiel zeigt die Berechnung des Meinungsforschungsinstituts zur Ausschöpfung bei einer telefonischen Befragung. Nach der Institutsrechnung beträgt die Ausschöpfungsquote 62,6 %. Die vier Kategorien Telefonbeantworter, Krankheit, abwesend, nicht erreichbar wird man aber kaum zu den stichprobenneutralen Ausfallgründen rechnen können. Zählt man diese Kategorien zu den systematischen Ausfällen, dann vermindert sich die Ausschöpfungsquote auf 52,6 %! Rechnet man auch noch „Sprachprobleme" hinzu, erhalten wir eine Ausschöpfungsquote von 50 %.

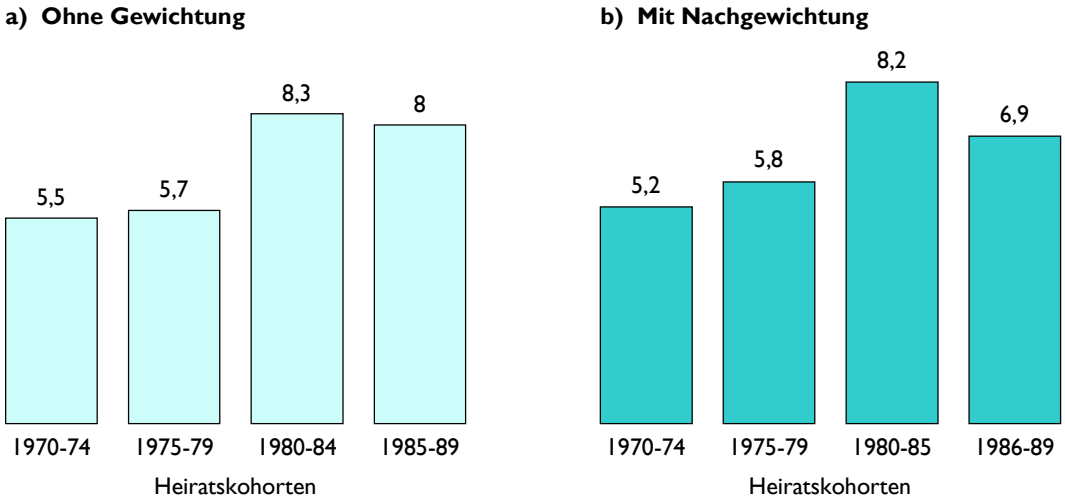
Die Ausschöpfungsquote entspricht der Anzahl auswertbarer Interviews dividiert durch den Umfang der Nettostichprobe. Letzterer errechnet sich aus dem Umfang der Bruttostichprobe abzüglich der stichprobenneutralen Ausfälle. Wann aber sind Ausfälle neutral und wann systematisch? Je mehr Ausfälle als neutral deklariert werden, desto höhere Werte lassen sich für die Ausschöpfungsquote errechnen. Da verbindliche Normen nicht existieren, gibt es hier einigen Gestaltungsspielraum. Man betrachte dazu das in der Tabelle aufgeführte Beispiel. Das beauftragte Institut hatte eine Ausschöpfungsquote von rund 63 Prozent angegeben. Wir haben die stichprobenneutralen Ausfälle geprüft und nachgerechnet. Bei Anlegung der üblichen Maßstäbe ergibt sich eine Ausschöpfungsquote von gerade 53 Prozent.

Die Ausschöpfungsquote ist ein Gütekriterium wie bei einem Auto z. B. die Angabe eines sparsamen Benzinverbrauchs. Nur wird der Verbrauch bei einem Auto nach einer einheitlichen europäischen Vorschrift ermittelt. Solche Normierungen existieren in der Umfragepraxis leider noch nicht. Das Beispiel der Ausschöpfungsquote zeigt, dass die Beurteilung der methodischen Qualität von Erhebungen unnötig erschwert wird, weil man es versäumt hat, gewisse Standards festzulegen. Ein wichtiger Schritt in die Richtung der Festlegung von Standards ist aber die Arbeit einer Kommission der „Deutschen Forschungsgemeinschaft", die eine Denkschrift zu den „Qualitätskriterien in der Umfrageforschung" vorgelegt hat (Kaase 1999).

Beispiel 3: Nachgewichtung ist kein Heilmittel!

In der Regel weichen die Randverteilungen einiger Merkmale in neu erhobenen Stichproben wie Alter, Zivilstand, Geschlecht, Erwerbsstatus, Bildung usw. von den Randverteilungen der amtlichen Statistik ab. Unter der durchaus diskussionswürdigen Annahme, dass die amtliche Statistik näher an der Wahrheit liegt, wird in der Praxis meist nachgewichtet. Man spricht auch von Redressment oder „Nachschichtung". Dies ist etwas ganz anderes als eine Designgewichtung gemäß Stichprobenplan. Eine Designgewichtung erfolgt auf der Grundlage der Stichprobentheorie und korrigiert für unterschiedliche Auswahlwahrscheinlichkeiten. Wenn z. B. Tessiner Befragte mit einer x-mal höheren Wahrscheinlichkeit in die Stichprobe aufgenommen werden, kann hierfür mit einem Gewicht korrigiert werden, das dem Kehrwert der Auswahlwahrscheinlichkeit entspricht. Nachgewichtung dagegen ist durch keine Theorie gedeckt.

Wenn man viele Merkmale erhebt, ist auch bei perfekten Zufallsstichproben zu erwarten, dass „rein zufällig" bei der einen oder anderen Randverteilung signifikante Abweichungen von der Verteilung in der Population auftreten. Aber selbst bei systematischen Fehlern gibt es keine Garantie, dass eine Nachgewichtung die Schätzungen verbessert. Natürlich wird es als „unschön" empfunden, wenn z. B. Frauen, bestimmte Altersklassen oder Bildungskategorien krass über- oder unterrepräsentiert sind. Anpassung an bekannte Randverteilungen stellt rein optisch die Proportionen wieder her. Die entscheidende Frage ist aber, ob durch die Nachgewichtung auch die Schätzungen anderer Verteilungen, d. h. Verteilungen von Merkmalen, die nicht in die Gewichtungsformel eingehen, im allgemeinen verbessert und nicht verschlechtert werden. Hierfür kenne ich aber keinen überzeugenden Nachweis.



Schweizer Mikrozensus Familie Teilstichprobe Frauen; N = 2143.
Nachgewichtung mit den Variablen Alter, Zivilstand und Nationalität.

Abbildung 2: Prozentualer Anteil geschiedener Erstehen nach fünf Jahren Ehedauer

Mit einem Kollegen analysiere ich derzeit die Daten des „Family and Fertility Surveys“ (FFS) bezüglich einer Hypothese, die einen Zusammenhang zwischen dem Geschlecht von Kindern und dem Scheidungsrisiko der Eltern behauptet. Demnach sollen Ehen mit Töchtern – ceteris paribus – instabiler sein als Ehen mit Söhnen. Nebenbei bemerkt, hat sich die in der Demographie viel zitierte Hypothese bei einer Prüfung anhand der FFS-Stichproben aus rund zwanzig Ländern eindeutig als falsch herausgestellt. Beim Schweizer Datensatz (ca. 6.000 Face-to-Face Interviews 1994/95, Stichprobenziehung aus dem Telefonregister) fiel uns eine Besonderheit auf. Im Gegensatz zu anderen Datenquellen nimmt das Scheidungsrisiko in der jüngsten Eheschließungskohorte (1985–89) im Vergleich zu den Vorgängerkohorten nicht zu, sondern der Tendenz nach sogar ab. Diese Entwicklung steht so sehr im Widerspruch zu anderen Statistiken, dass man wohl eher von Stichprobenfehlern oder anderen Erhebungsfehlern ausgehen kann. Wie üblich wurde auch eine Nachgewichtung mit den Variablen Alter, Nationalität (Schweizer bzw. Ausländer) und Zivilstand vorgenommen. Betrachten wir jetzt das Scheidungsrisiko nach Kohorten vor und nach der Gewichtung (Abbildung 2). Es zeigt sich, dass die Nachgewichtung die Abweichung noch vergrößert. Wenn man es nicht besser wüsste, würde man jetzt prognostizieren, dass die Entwicklung der Ehescheidungen rückläufig ist.

Stattdessen ist der Grund vielleicht, dass die jüngeren Geschiedenen einfach schlechter erreichbar waren. Die Nachgewichtung hat dabei den Fehler noch vergrößert.

Die Beispiele demonstrieren, dass eine Bewertung der Qualität von Umfragen sich nicht schematisch auf wenige oberflächliche Hinweise wie die angebliche „Repräsentativität“ und den Stichprobenumfang stützen kann. Zu viele Details spielen eine Rolle. Eine wichtige Forderung ist aber, dass erstens über die einzelnen Merkmale der Erhebungsmethodik informiert wird. Die obige, eventuell ergänzungsbedürftige Liste könnte dabei als Richtschnur oder „Checkliste“ dienen. Zweitens erscheint es wünschenswert, die Details von Berechnungen zu standardisieren. Die Ausschöpfungsquote z. B. sollte von jedem Institut nach einem Standardschema zur Kategorisierung stichprobenneutraler und systematischer Ausfälle berechnet werden. Abweichungen von den Standards stehen jedem frei, müssen aber begründet werden. Transparenz und Standardisierung vereinfachen die Evaluation der Erhebungsmethodik.

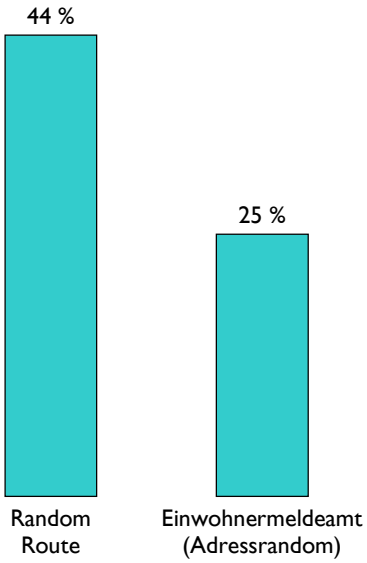
2 Datenfälschung: Umfang, Konsequenzen und Diagnose

Zur Datenqualität zählt sicher auch, dass die Daten nicht gefälscht sind. Datenfälschung ist vor allem bei Face-to-Face Interviews ein ernstzunehmendes Problem. Dabei handelt es sich meist um Teilfälschungen, die bei den üblichen Feldkontrollen nicht entdeckt werden. Bei Teilfälschungen stellt der Interviewer einige Fragen und ergänzt später die Kurzinterviews. Interviewern, die von ihrem Beruf leben müssen und im Akkord arbeiten, ist dies nicht einmal zu verdenken. Und die Institute beschwichtigen und verdrängen das Problem (Dorroch 1994).

Eine extreme These besagt, dass Interviewfälschungen die Datenqualität nicht beeinträchtigen. Sind hundert Prozent der Interviews gefälscht, dann hat man eben eine Befragung von Interviewern und die wissen meist besser Bescheid als die Befragten selbst. Gefälschte Interviews sind gewissermaßen Experteninterviews. Natürlich ist diese These unhaltbar, denn erstens sind Interviewer ein selektives Sample und zweitens können sie kaum stellvertretend für die Zielpersonen sprechen, wenn es um neue Themen und Stimmungen oder ihnen unbekannte Sachverhalte geht. Weniger ra-

dikal ist die Auffassung, dass bis zu fünf Prozent gefälschte Interviews die Ergebnisse nur wenig beeinflussen. Die Toleranzgrenze von fünf Prozent wurde mit Simulationsrechnungen ermittelt (Schnell 1991; siehe auch Reuband 1990). Ob solche Befunde, die dann häufig als legitimierende Daumenregeln herangezogen werden, wirklich verallgemeinerbar sind, sei dahingestellt. Einige Indizien deuten aber darauf hin, dass der Umfang gefälschter Interviews fünf Prozent erheblich übertreffen kann. Hier einige Indizien, die diese Vermutung unterstützen:

- Kommen wir zurück auf die Alternative Adressrandom oder Random Route. In der Theorie ist Random Route gut, sofern sich Interviewer daran halten. Die Einhaltung ist aber kaum kontrollierbar. Und die Interviewer haben einen Anreiz, von den Regeln abzuweichen, wenn der Zielhaushalt schlecht erreichbar oder im Haushalt keine Person anwesend ist. Dann wird eben beim Nachbarn geläutet. Beim „Deutschen Familiensurvey“ (ca. 10.000 Face-to-Face Interviews, 1988) wurde eine Teilstichprobe aus Gemeinderegistern gezogen (Adressrandom), die Befragten der zweiten Teilstichprobe wurden per Random Route ermittelt. Dieser einmalige Methoden-Split erlaubt u. a. einen Vergleich der beim ersten Kontakt realisierten Interviews. Es sollte bei Einhaltung der Regeln zur Auswahl der Befragten eigentlich keinen Unterschied geben. Tatsächlich findet man eine sehr starke Differenz. Random-Route Interviewer realisieren beim ersten Kontakt weitaus mehr Interviews als ihre Kollegen mit den vorgegebenen Adressen der Zielpersonen aus der Gemeindestichprobe (Abbildung 3).



Familiensurvey des Deutschen Jugendinstituts.
Einwohnermeldeamt: N = 3011
Random Route: N = 6931

Angaben nach Alt (1991)

Abbildung 3: Interviewhäufigkeit beim ersten Kontaktversuch

- In einer Rostocker Verkehrsstudie im Auftrag der Bundesanstalt für Straßenverkehr, an der ich mitbeteiligt war, wurden von dem beauftragten Meinungsforschungsinstitut ca. 600 Face-to-Face Interviews durchgeführt. Eine Teilgruppe von 80 Autofahrern wurde wenig später zur Vorbereitung einer Interventionsstudie erneut befragt. Dabei stellte sich heraus, dass manche „Autofahrer“ über kein Auto verfügten oder sogar keinen Führerschein hatten. Eine intensive Kontrolle aller 80 Interviews ergab 16 vollständige oder Teilfälschungen. Hochgerechnet also 20 Prozent Fälschungsquote. Das Institut hatte eine Feldkontrolle durchgeführt und dabei keines der gefälschten Interviews entdeckt. (Unsere Reaktion: Wir haben die Daten aller 600 Interviews weggeworfen und eine vollständig neue Erhebung mit ausgeklügelten Kontrollen durchgeführt.)

- Teilnehmer an einer medizinischen Studie sollten nach vorgegebenen Regeln täglich ein Inhalationsgerät benutzen. Die Teilnehmer wussten, dass die verbrauchte Menge gemessen wurde. Wem die Einnahme lästig war, der hatte allerdings die Möglichkeit – so konnte man denken – kurz vor dem Kontrollbesuch unbemerkt die gesamte Ladung zu versprühen. Tatsächlich wurde aber die Regelverletzung mit einer den Versuchspersonen unbekannten elektronischen Vorrichtung registriert. Innerhalb von zwölf Monaten hatten 30 Prozent der Versuchspersonen (30 von 101) mindestens einmal vorgetäuscht, dass sie das Medikament eingenommen hätten, obwohl dies nicht der Fall war (Simmons et al. 2000).

Fälschungen kommen häufiger vor, als man denkt. Nicht nur bei Interviewern, auch bei Wissenschaftlern.

Gibt es Diagnosemöglichkeiten? Man kann eine wesentlich genauere Feldkontrolle praktizieren (mit der Abfrage einer Kombination objektivierbarer Merkmale), um auch Teilfälschungen auf die Schliche zu kommen. Diese Methode prüfen wir derzeit bei der neuen Erhebung des Deutschen Familiensurveys. Ich möchte abschließend aber noch über eine unkonventionelle Methode sprechen, die ich derzeit ausprobiere. Diese Methode basiert auf statistischen Regelmäßigkeiten von Ziffernhäufigkeiten.

Statistische Daten sind meistens Zahlen und Ziffern. Unter gewissen Bedingungen folgen die Ziffern von echten Daten Gesetzmäßigkeiten, die gefälschte Daten nicht gleichermaßen erfüllen.

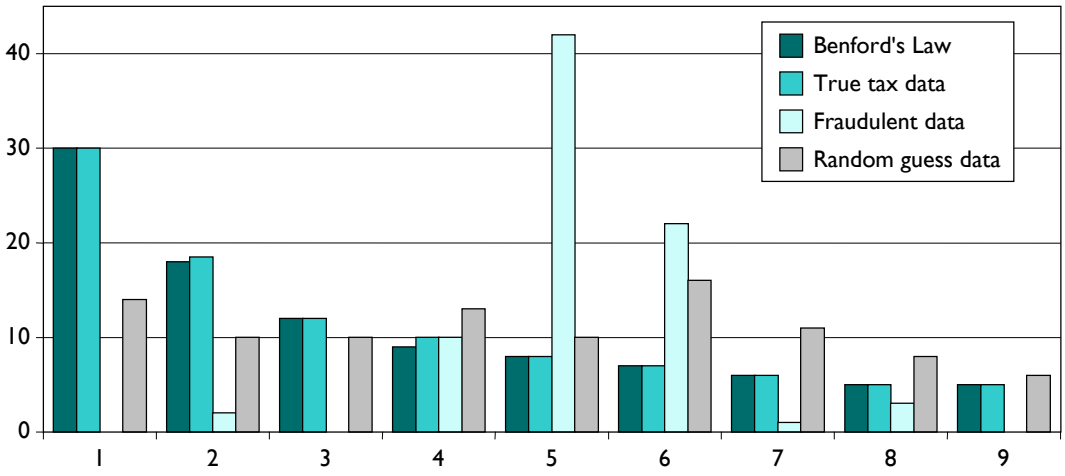
Ein einfaches Beispiel von Hill (1998) zur Erläuterung des Prinzips. Stellen Sie sich vor, Ihnen wird die folgende Aufgabe gestellt. Sie sollen die Daten von 200 Münzwürfen als Abfolge von „Kopf“ und „Wappen“ notieren. Sie können eine Münze werfen und echte Daten berichten; Sie können sich aber auch die Daten ausdenken. Ich behaupte nun, dass ich über magische Kräfte verfüge und die echten von den gefälschten Versuchsreihen unterscheiden kann. Das ist ganz einfach. Mit sehr hoher Wahrscheinlichkeit findet man bei einem echten Zufallsexperiment von 200 Münzwürfen eine Abfolge von mindestens sechsmal Kopf oder Wappen in ununterbrochener Folge. Wer dagegen solche Daten fälscht, schreibt höchst selten sechsmal hintereinander das gleiche Symbol.

Ähnlich verhält es sich mit Ziffern, die aus natürlichen oder sozialen Prozessen resultieren. Auch hier ein Beispiel. Ich biete einer Person folgende Wette an: Ich setze darauf, dass die erste Ziffer in dem Artikel rechts unten auf der Wirtschaftsseite der Ausgabe der Neuen Zürcher Zeitung, die am nächsten Tag erscheinen wird, im Bereich eins bis vier liegt. Ich verliere die Wette, wenn sich die Ziffer im Bereich fünf bis neun befindet. Bei Gewinn erhalte ich zehn Franken, bei Verlust zahle ich die gleiche Summe an meinen Wettpartner. Ist das ein gutes Angebot?

Ist es nicht. Denn die ersten Ziffern von Zahlen verschiedenster Dinge wie Hausnummern, Börsenkurse, die Fläche von Gewässern, die Flügelspannweite von Vögeln u.s.w. sind nicht gleichverteilt. Die „1“ z. B. kommt wesentlich häufiger vor als die „9“. Unter bestimmten Voraussetzungen folgt die erste Ziffer vieler Daten einer logarithmischen Verteilung, der so genannten Benford-Verteilung.

„Benfords Gesetz“ hat man sich z. B. zunutze gemacht, um Steuerbetrug und Bilanzfälschungen aufzudecken. Die Logik ist ganz einfach. Echte Zahlen folgen der Benford-Verteilung, gefälschte Daten weichen davon ab. Weitere Tests kann man entwickeln, indem man die zweite und dritte Ziffer heranzieht und darüber hinaus die bedingten Verteilungen (z. B. die Verteilung der zweiten Ziffer unter der Voraussetzung, dass die erste Ziffer n ist) inspiziert.

Die Untersuchung betrügerischer Manipulationen in der Buchhaltung von Unternehmen, von Steuererklärungen usw. durch Nigrini (1997, 1999) zeigt, dass der Benford-Test zumindest in diesem Bereich Anhaltspunkte liefert, um gefälschte Daten herauszufiltern. Bei den gefälschten Daten kommen bestimmte Ziffern viel häufiger vor als bei den echten Angaben (Abbildung 4).



Aus: Hill (1998)

Abbildung 4: Aufdeckung von Fälschungen mit Benfords Gesetz

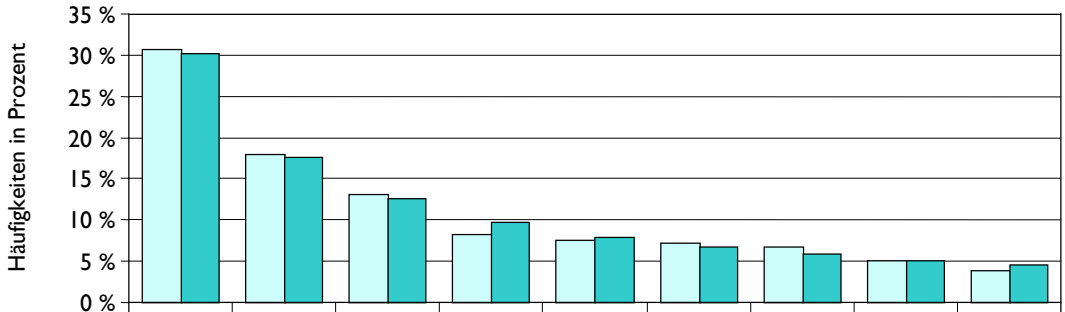
Die ersten Ziffern der Angaben aus 169.662 Steuererklärungen, die von M. Nigrini ausgewertet wurden, stimmen sehr gut mit Benfords Gesetz überein. Dies war bei betrügerischen Geschäftsdaten nicht der Fall. Ebenso stimmen die ersten Ziffern der Angaben von 743 Studenten, die gebeten wurden, sechsstellige Zufallszahlen aufzuschreiben, nicht mit der Benford-Verteilung überein (nach Hill 1998).

Nach diesem ermutigenden Schritt hat man mit Benfords Gesetz auch die Steuererklärung von U.S. Präsident Bill Clinton inspiziert. Der Test war negativ. Kein Fall für Kenneth Starr. Es gab keine Hinweise auf auffällige Ziffernhäufigkeiten.

Nun wird man natürlich einwenden, dass Fälscher sich schnell auf den Benford-Test einstellen werden. Das ist aber kaum möglich. Es ist ungeheuer schwierig, Daten konsistent zu fälschen und dann noch Benford-konform zu frisieren, insbesondere, wenn auch noch die zweite und dritte Ziffer berücksichtigt werden muss.

Was bei der Entdeckung gefälschter Steuererklärungen und Bilanzfälschungen hilfreich ist, könnte sich auch bei gefälschten Interviews und generell zur Diagnose gefälschter wissenschaftlicher Daten als nützlich erweisen.

In vielen empirischen Studien werden z. B. tabellenweise Regressionskoeffizienten berichtet. Um eine Art Referenzstichprobe zu gewinnen, habe ich eine Auszählung der ersten Ziffer von tausend geschätzten Regressionskoeffizienten aus Artikeln des American Journal of Sociology veranlasst. Es ist auch statistisch ganz interessant, dass die Häufigkeitsverteilung recht gut mit der Benford-Verteilung übereinstimmt (Abbildung 5 auf der nächsten Seite). Gerät eine Veröffentlichung unter Fälschungsverdacht, könnte man den Benford-Test auf die Ziffern geeigneter Statistiken anwenden. Bevor das gemacht wird, sind aber weitere Tests erforderlich.



	1	2	3	4	5	6	7	8	9
AJS	30,65 %	18,00 %	12,99 %	8,15 %	7,56 %	7,13 %	6,71 %	5,01 %	3,82 %
Benford	30,10 %	17,60 %	12,50 %	9,70 %	7,90 %	6,70 %	5,80 %	5,10 %	4,60 %

Stichprobe aus Tabellen von Artikeln, die im Zeitraum von Januar 1996 (Vol 101) bis Mai 1997 (Vol 102) im American Journal of Sociology publiziert wurden (N = 1.000).

Abbildung 5: Verteilung der ersten Ziffer von unstandardisierten Regressionskoeffizienten

Drei Experimente wurden dazu am Institut für Soziologie in Bern mit Studierenden durchgeführt. In einem ersten Experiment sollten die Versuchspersonen Zufallszahlen generieren. In zwei weiteren Experimenten wurden die Teilnehmer und Teilnehmerinnen an einem Statistik-Seminar in Soziologie sowie Studierende der Ökonometrie gebeten, zu tun, was sie sonst nie tun dürfen, nämlich Tabellen mit Regressionskoeffizienten „passend“ zu einer vorgegebenen Hypothese zu erfinden. In allen drei Experimenten waren bei der ersten Ziffer keine auffälligen Abweichungen zu erkennen. In den zwei Experimenten mit gefälschten Tabellen entsprachen auch die gefälschten ersten Ziffern der Benford-Verteilung, obwohl die Versuchspersonen mit dieser nicht vertraut gemacht wurden. Interessanterweise zeigten sich aber in den drei Experimenten übereinstimmend signifikante Abweichungen gegenüber der statistischen Erwartung (Zufallszahlen) bzw. Benford-Verteilung (Regressionskoeffizienten) bei der zweiten Ziffer, wobei allerdings nicht immer dieselben Ziffern als Ausreißer in Erscheinung traten. Ein ähnliches Resultat bezüglich der Fabrikation von Zufallszahlen berichten Mosimann et al. (1995). Wenn diese Ergebnisse robust sein sollten, müsste demnach bei Fälschungsverdacht insbesondere die Verteilung zweiter Ziffern unter die Lupe genommen werden.

Um keine ins Kraut schießenden Erwartungen bezüglich eines einfachen „Lackmustests“ für Fälschungen zu wecken, sei deutlich betont: Die Prüfung des Verfahrens befindet sich in einem ersten Versuchsstadium. Ob etwas herauskommt, wird sich zeigen. Ich denke aber, es lohnt sich, unkonventionelle Methoden zu testen, um Fälschungen auf die Spur zu kommen.

3 Literatur

Alt, C. (1991): Stichprobe und Repräsentativität, in: H. Bertram (Hg.); Die Familie in Westdeutschland. Stabilität und Wandel familialer Lebensformen, DJI: Familien-Survey 1, Opladen: Leske & Budrich, S. 497–531.

Buhmann, B.; Achermann, Y. und Martinovits, A. (1994): Comparability of Labour Force Data from Different Sources. SAKE-News 3/94, Neuenburg: Bundesamt für Statistik.

Diekmann, A. (2001): Empirische Sozialforschung, 7. Aufl. Reinbek: Rowohlt.

Diekmann, A.; Engelhardt, H.; Jann, B.; Armingeon, K.; und Geissbühler, S. (1999): Der Schweizer Arbeitsmarktsurvey 1998. Codebuch. Universität Bern: Mimeo.

Hill, T. P. (1998): The First Digit Phenomenon, in: American Scientist, 86: 358–363.

Kaase, M. (1999): Qualitätskriterien der Umfrageforschung. Memorandum, Berlin: Akademie Verlag.

Mosimann, J. E.; Wiseman, C. V. und Edelman, R. E. (1995): Data Fabrication: Can People Generate Random Digits? In: Accountability in Research, 4: 31–55.

Nigrini, M. (1999): I've Got Your Number. How a Mathematical Phenomenon Can Help CPAs Uncover Fraud and Other Irregularities, in: Journal of Accountancy: 79–83.

Nigrini, M. (2000): Digital Analysis Using Benford's Law, Vancouver: Global Audit Publications.

Reuband, K.-H. (1990): Interviews, die keine sind. „Erfolge" und „Misserfolge" beim Fälschen von Interviews, in: Kölner Zeitschrift für Soziologie und Sozialpsychologie, 42: 706–733.

Schnell, R. (1991): Der Einfluss gefälschter Interviews auf Survey-Ergebnisse, in: Zeitschrift für Soziologie, 20: 25–35.

Simmons, M. S., Nides, M. A., Rand, C. S., Wise, R. A. und Tashkin, D. P. (2000): Unpredictability of Deception in Compliance With Physician-Prescribed Bronchodilator Inhaler Use in a Clinical Trial, in: Chest, 118: 290–295.

Bisher erschienene manu:scripte

ITA-01-01 Gunther Tichy, Walter Peissl (12/2001): Beeinträchtigung der Privatsphäre in der Informationsgesellschaft. <http://www.oeaw.ac.at/ita/pdf/ita_01_01.pdf>

ITA-01-02 Georg Aichholzer(12/2001): Delphi Austria: An Example of Tailoring Foresight to the Needs of a Small Country. <http://www.oeaw.ac.at/ita/pdf/ita_01_02.pdf>

ITA-01-03 Helge Torgersen, Jürgen Hampel (12/2001): The Gate-Resonance Model: The Interface of Policy, Media and the Public in Technology Conflicts. <http://www.oeaw.ac.at/ita/pdf/ita_01_03.pdf>

ITA-02-01 Georg Aichholzer (01/2002): Das ExpertInnen-Delphi: Methodische Grundlagen und Anwendungsfeld „Technology Foresight“. <http://www.oeaw.ac.at/ita/pdf/ita_02_01.pdf>

ITA-02-02 Walter Peissl (01/2002): Surveillance and Security – A Dodgy Relationship. <http://www.oeaw.ac.at/ita/pdf/ita_02_02.pdf>

ITA-02-03 Gunther Tichy (02/2002): Informationsgesellschaft und flexiblere Arbeitsmärkte. <http://www.oeaw.ac.at/ita/pdf/ita_02_03.pdf>

ITA-02-04 Andreas Diekmann (06/2002): Diagnose von Fehlerquellen und methodische Qualität in der sozialwissenschaftlichen Forschung. <http://www.oeaw.ac.at/ita/pdf/ita_02_04.pdf>