Prof. Dr. Ulrich Rendtel, FB Wirtschaftswissenschaft,
Freie Universität Berlin, Garystr. 21, D-14195 Berlin

Prof. Dr. Ulrich Rendtel
Inst. für Statistik und Ökonometrie
FB Wirtschaftswissenschaft
Freie Universität Berlin
Garystraße. 21
14195 Berlin

Telefon +49 (30) 838-54202
Sekretariat +49 (30) 838-55791
E-Mail ulrich.rendtel@fu-berlin.de
Internet www.wiwiss.fu-berlin.de

Berlin, 15.11.2018

Gutachten zur Repräsentativität von Online-Umfragen

Online-Umfragen, die ihre Resultate über die Rekrutierung aus dem Internet gewinnen, sehen sich häufig dem Vorwurf ausgesetzt, dass die daraus abgeleiteten Ergebnisse nicht aussagekräftig für die Bevölkerung sind. Dieses Gutachten setzt sich mit dem Vorwurf auseinander, dass die Ergebnisse von Online-Umfragen generell nicht „repräsentativ" sind.

Genau dies ist der Gegenstand einer Beschwerde aus dem Mai 2018 vor dem Presserat, in der drei Meinungsforschungsinstitute darüber Klage führen, dass die Firma Civey die Ergebnisse ihrer Online Umfragen als repräsentativ bezeichnet. Konkret ging es um die Veröffentlichung einer Umfrage in FOCUS Online über die Zugehörigkeit von Mesut Özil und Ilkay Gündogan zur deutschen Nationalmannschaft, nachdem sie einen PR-Auftritt mit dem türkischen Präsidenten Erdogan hatten.

1) Grundsätzliches zum Begriff der Repräsentativität

Die Beschwerdeführer verweisen darauf, dass gemäß eines Grundsatzbeschlusses des Plenums des Deutschen Presserates „nicht-repräsentative Online-Umfragen als solche gekennzeichnet werden müssen". Weiter heißt es: „Als repräsentative Umfragen werden gemeinhin solche Umfragen bezeichnet, die … nach anerkannten und allgemein überprüfbaren wissenschaftlichen Methoden durchgeführt werden und aufgrund ihrer Methodik geeignet sind, bezogen auf die zugrunde liegende Fragestellung eine statistisch gültige Aussage über das Stimmungsbild der Gesamtheit zu treffen". Der Umfrage, auf der die Veröffentlichung beruht, wird unterstellt, dass sie „ein Befragungsdesign verwendet, das nach den allgemein anerkannten wissenschaftlichen Kriterien der empirischen Sozialforschung ganz grundsätzlich nicht geeignet ist, „repräsentative" Ergebnisse zu liefern". Die Beschwerde richtet sich also prinzipiell gegen die Bezeichnung von Online-Umfragen als repräsentative Erhebung. Es ist also nicht nur diese eine spezifische Erhebung gemeint, die von FOCUS Online veröffentlicht wurde. Der Vorwurf unterstellt also allen Online-Umfragen „Nicht-Repräsentativität". Immerhin hätte man den Grundsatzbeschluss ja auch noch so interpretieren können, dass nur nicht-repräsentative

Online-Umfragen im Gegensatz zu repräsentativen Online-Umfragen als solche zu kennzeichnen sind. Die Kategorie „repräsentative Online-Umfrage“ scheint es aber nach Meinung der Beschwerdeführer nicht zu geben.

Der Streit, welche Umfrage sich „repräsentativ“ nennen darf, stößt auf das Problem, dass sich dieser Begriff einer exakten Definition entzieht. In methodisch orientierten Lehrbüchern über Stichprobenverfahren findet man zu diesem Begriff keinen substanziellen Eintrag (Beleg: die international weit verbreiteten Lehrbücher von Särndal/Swensson/Wretman (1992) (Model Assisted Survey Sampling, Springer Verlag, Heidelberg) sowie Lohr (1999) (Sampling, Design and Analysis, Duxbury Press, Pacific Grove). Das Problem, dass sich der Begriff der repräsentativen Stichprobe einer exakten Definition entzieht, wurde in einer 4-teiligen Artikelserie von William Kruskal und Frederick Mosteller 1979/80 diskutiert (Representative Sampling I: Nonscientific literature, International Statistical Review, 47, 13-24; Representative Sampling II: Scientific literature excluding statistics, International Statistical Review, 47, 111-127; Representative Sampling III: Current statistical literature, International Statistical Review, 47, 245-265; Representative Sampling IV: The history of the concept in statistics, International Statistical Review, 48, 169-195). Die Autoren nennen mehrere unterschiedliche Bedeutungen, unter denen der Begriff der repräsentativen Stichprobe oder der repräsentativen Stichprobenziehung benutzt wird:

1. General acclaim for data
2. Absence of selective forces
3. Miniature of the population
4. Typical or ideal cases
5. Coverage of the population
6. Vague term for some specific kind of sampling procedure that is later on made precise
7. Some specific sampling methods
8. Permitting good estimation

Diese Begriffe sind zum Teil widersprüchlich. Beispielsweise liefert ein geschichtetes Ziehungsverfahren mit einer einfachen zufälligen Stichprobe innerhalb der Schichten ein nicht-proportionales Abbild der Population, wenn die Stichprobenumfänge nicht proportional zu der Populationsgröße der Schichten gewählt werden. Dieses kann bei der Neymann-Allokation der Stichproben durchaus passieren. Wie schon erwähnt, gibt es keine feste Zuordnung des Begriffs „representative sampling“ zu einem genau spezifizierten Verfahren. Die Autoren belegen auch, dass selbst eine weite Definition des Begriffs einer repräsentativen Stichprobenziehung als zufällige Stichprobe nicht immer geteilt wird, z.B. weil der Stichprobenumfang nicht groß genug ist. „Some writers adopt arbitrary definitions, for example, that „representative sampling” is synonymous with „random sampling”. Then why introduce a new term? Or that representative sampling is stratified sampling with stratum subsample sizes proportional to stratum size”. (Aufsatz I, S. 15).

Die von den Beschwerdeführern gemachte Aussage, dass repräsentative Umfragen statistisch gültige Aussagen über die Grundgesamtheit zulassen, deckt sich in etwa mit der Konnotation unter Punkt 8. „Permitting good estimation“. Damit stellt sich die Frage: Ist es möglich, mit dem vom Erhebungsinstitut Civey benutzten Verfahren statistisch valide Schlüsse über die Grundgesamtheit zu treffen? Die Beschwerdeführer verneinen dies mit dem Hinweis auf die Selbstselektion von Online-Panels. Diese Selbstselektion basiert nicht auf einem „zufälligen oder systematischen Stichprobenverfahren.“ „Dieser

sogenannte self-selection bias führt dazu, dass auf diesem Weg durchgeführte Umfragen keineswegs repräsentativ sind und eine Repräsentativität der Ergebnisse auch dann nicht hergestellt werden kann, wenn soziodemographische Merkmale der untersuchten Gruppen (z.B. durch Gewichtungsverfahren) nachträglich mit denen der Grundgesamtheit in Übereinstimmung gebracht werden“, so die Einschätzung der Beschwerdeführer.

2) Zur Repräsentativität von Non-Probability Samples

Die Frage, wie und unter welchen Voraussetzungen man aus Umfragen, die nicht über ein zufälliges Auswahlverfahren gewonnen wurden, Rückschlüsse auf die Verteilung in der Grundgesamtheit schließen kann, beschäftigt Statistiker seit dem Aufkommen von Online-Umfragen verstärkt. In einem Überblicksartikel (Elliott/Valliant (2017): Inference for Nonprobability Samples, Statistical Science, 32, 249-264) werden insgesamt drei Ansätze vorgestellt. Beide Autoren sind Mitglieder des renommierten Institute for Social Research der Universität Michigan, so dass der Terminus der „anerkannten Forschung“ für diese Ergebnisse in Anspruch genommen werden kann. Weiterhin haben diese Ergebnisse mittlerweile Eingang in Lehrbücher gefunden. So z.B. in Kapitel 16 des Lehrbuchs von Valliant/Dever/Kreuter (2013) (Practical Tools for Designing and Weighting Survey Samples, Springer Verlag, Heidelberg).

Die drei in dem Überblicksartikel dargestellten Methoden basieren a) auf einer indirekten Schätzung von Teilnahmewahrscheinlichkeiten von Nonprobability Samples auf Basis eines Vergleichs von Merkmalen mit einer Zufallsstichprobe oder b) auf einer Prädiktion der nicht beobachteten Werte über einen Regressionsansatz oder c) auf einem Bayesianischen Ansatz, der auf Vorwissen über die Verteilung in der Population beruht. Die indirekte Schätzung von Teilnahmewahrscheinlichkeiten, auch als „Quasi-Randomisierungs“-Ansatz bezeichnet, liefert Gewichtungsfaktoren, die unabhängig von dem betrachteten Merkmal bei der Schätzung verwendet werden können. Sie entsprechen daher quasi den klassischen Gewichten bei der zufälligen Stichprobenziehung.

Der Quasi-Randomisierungs Ansatz basiert auf dem Vergleich der Verteilung von Merkmalen, die die Auswahl für eine randomisierte Stichprobe erklären. Die randomisierte Stichprobe kann beispielsweise der Mikrozensus sein. Der Ansatz läuft darauf hinaus, die Verteilungen dieser Merkmale für die randomisierte Stichprobe mit der Verteilung in der nicht randomisierten Stichprobe zu vergleichen und die Designgewichte der randomisierten Verteilung multiplikativ mit einem Korrekturfaktor zu verknüpfen. Dieser Korrekturfaktor ergibt sich aus dem Verhältnis der Häufigkeiten in der zufälligen und nicht-zufälligen Stichprobe für die jeweiligen Werte der erklärenden Merkmale. Falls die Werte der zufälligen Stichprobe durch diejenigen der Grundgesamtheit ersetzt werden, ergibt sich ein Designgewicht von konstant 1 und der Korrekturfaktor reduziert sich auf das Zahlenverhältnis in Grundgesamtheit und nicht-zufälliger Stichprobe. Damit ist man fast bei der Standard Soll/Ist-Gewichtung der klassischen Stichprobenverfahren angelangt.

Die Autoren benennen auch die Voraussetzung für die Gültigkeit des Verfahrens. Dies ist die Annahme, dass die Auswahl der Einheiten in der nicht-zufälligen Stichprobe nicht mehr von dem interessierenden Merkmal sondern nur noch von den gemeinsamen bekannten Merkmalen aus dem Probability und Non-Probability Sample abhängt (Missing at Random – Annahme im Sinne von Rubin). Diese Voraussetzung ist intrinsisch nicht überprüfbar, da die zugehörigen Werte für die Non-Responder nicht beobachtet werden.

3) Zur Repräsentativität von Telefonumfragen

Allerdings steht die klassische Umfrageforschung mit ihren Telefoninterviews und Nonresponse-Raten von 90 % vor demselben Problem. Auch hier erhält man nur dann unverfälschte Populationsschätzer, wenn die Responsewahrscheinlichkeiten exakt durch die Kalibrationsmerkmale beschrieben werden, vgl. Särndal/Lundström (2005) (Estimation in Surveys with Nonresponse, Wiley, New York). Wenn das zu untersuchende Merkmal einen direkten Einfluss auf die Responsewahrscheinlichkeit hat, so sind die resultierenden Schätzungen für die Population verfälscht. Wenn also eine Befragung zum Thema Sozialhilfebezug ein besonderes Interesse bei Sozialhilfebeziehern hervorruft, so kann man davon ausgehen, dass die Antwortwahrscheinlichkeit nicht nur von den demographischen Anpassungsmerkmalen abhängt sondern zusätzlich von dem Merkmal Sozialhilfebezug. Dieser Effekt wird in der Survey Statistik als „Saliency Effect“ bezeichnet.

Es sei an dieser Stelle auch angemerkt, dass Telefonumfragen vor dem methodischen Problem stehen, dass in Deutschland kein vollständiges Verzeichnis aller Telefonnummern existiert, aus dem man zufällig mit gleicher Wahrscheinlichkeit auswählen kann. Bis zur Umstellung des Auswahlverfahrens auf den von Häder/Gabler vorgeschlagenen Ansatz wurde mit zufälligen Endziffern (Methode „Last Digit Random“ oder „Last 2 Digit Random“) gearbeitet, vgl Gabler/Häder (1998) (Probleme mit der Anwendung von RLD Verfahren. In Gabler et al. Hrsg. Telefonumfragen in Deutschland,Springer , Heidelberg, S. 58- 68). Bei diesen Verfahren schwankte die Auswahlwahrscheinlichkeit unkontrolliert um den Faktor 100.

Diese Probleme hindern die Sozialforschungsinstitute der Beschwerdeseite nicht daran, für ihre Ergebnisse „Repräsentativität“ zu beanspruchen, während sie ihrer Konkurrenz dieselbe Schlussweise (Annahme von „Missing at Random“ bei Kontrolle demographischer Merkmale) prinzipiell absprechen. Auch die Sozialforschungsinstitute können ihre MAR-Annahme nicht beweisen! Dies wäre nur möglich, wenn zumindest über einige Nonrespondenten das interessierende Merkmal beispielsweise im Rahmen einer Nachbefragung bekannt wäre. Diese Ansätze werden aber nicht praktiziert.

4) Die Stichprobenauswahl bei Civey

Die Argumentation der Beschwerdeführer richtet sich ganz auf eine ungeregelte Selbst-Selektion der Teilnehmer eines Online-Panels. Diese würde dann vorliegen, wenn jeder Internetnutzer, der eine Fragestellung von Civey interessant findet, mit Abgabe seiner Meinung direkt und ungeprüft in die Befragung eingeht. Dies entspricht jedoch nicht der Vorgehensweise von Civey. Hinweise hierauf hätten die Beschwerdeführer der Darstellung entnehmen können, die im Anschluss an die Antwortgewährung angeboten wird (siehe https://civey.com/blog/civey-methodik-so-funktionierts/).

Diese Maßnahmen umfassen:

a. Die Kontrolle von maschinellen Antworten (Bots etc.)
b. Mehrfachantworten (Kontrolle über IP-Adresse, Zeitstempel etc.)
c. Unplausible Antworten (über Abgleich mit früheren Antworten)

Es werden nur Personen aufgenommen, die zusätzlich Angaben zu Geschlecht, Alter, Bildung, Postleitzahl, Einkommen sowie Parteipräferenz machen. Diese Merkmale werden

für eine Quotierung benutzt, wobei Randverteilungen der Amtlichen Statistik übernommen werden.

Erwähnenswert erscheint mir in diesem Zusammenhang die Auswahl der Plätze (URLs), wo die Umfrage geschaltet wird. Hier verfügt Civey über URLs mit einer breiten Streuung sowohl in regionaler Sicht als auch hinsichtlich des Charakters der Plattform. Die Auswahl dieser URLs wird als „River Sampling“ bezeichnet. Es ist damit ein Pendant zu dem Random Sampling in der klassischen Umfrageforschung.

Das Verfahren von Civey ist damit keine einfache Quotenstichprobe, bei der der Auswahlprozess lediglich über die Festlegung der Quotenmerkmale gesteuert wird. Man beachte, dass die Einkommens- und Verbrauchsstichprobe (EVS) des Statistischen Bundesamts, auf deren Basis der Verbraucherpreisindex festgelegt wird, eine reine Quotenstichprobe ist, für die teilweise in Zeitungsannoncen inseriert wird. Dieses Vorgehen hängt damit zusammen, dass die Führung von Verbrauchstagebüchern eine hohe Befragungsbelastung bedeutet und die Befragung über eine zufällige Stichprobe als kaum zumutbar erachtet wird. Trotzdem werden die Ergebnisse der EVS allgemein akzeptiert.

5) Alternativen zum Bias als Bewertungskriterium einer Umfrage

Die Argumentation der Beschwerdeführer stellt ganz auf einen möglichen „Self-Selection Bias“ ab. Dies ist der Abstand des Erwartungswerts eines Schätzers von seinem zu schätzenden Wert, in diesem Fall also dem Wert in der Grundgesamtheit. In der Statistik untersucht man jedoch eher den quadratischen Abstand eines einzelnen Schätzwerts vom Populationswert. Der erwartete Wert dieses quadratischen Abstands wird als „Mean Squared Error“ (MSE) bezeichnet. Er ergibt sich als Summe aus quadriertem Bias plus der Varianz des Schätzers. Hier zeigt sich, dass viele Schätzer, die in freier Hochrechnung über eine zufällige Stichprobe erhoben werden, zwar unverfälscht sind, jedoch über eine hohe Varianz verfügen. Hinsichtlich des MSE sind häufig Schätzer zu bevorzugen, die zwar einen gewissen Bias aufweisen, jedoch deutlich stabiler sind, also eine geringe Varianz haben. In der Regel sind diese Schätzer modellbasiert, d.h. sie arbeiten ohne die Berücksichtigung von Auswahlwahrscheinlichkeiten. Hier ergibt sich eine Brücke zu den Non-Probability-Stichproben und dem Prediction-Ansatz in Elliott/Valliant (2017). Derartige Schätzer werden heute schon im Bereich der Small-Area-Verfahren (SAE) eingesetzt, vgl. den Überblick von Münnich et al (2013) (Small Area-Statistik: Methoden und Anwendungen, AStA Wirtschafts- und Sozialstatistisches Archiv, 6, 149-191). Die Betrachtung über den MSE eröffnet neue Schätzmethoden und erweitert die verengte Sichtweise auf einen möglichen Selektionsbias von Umfrageergebnissen.

6) Die FOCUS Veröffentlichung

Diese Stellungnahme hat bis jetzt nicht auf den konkreten Fall der Veröffentlichung im FOCUS Online Bezug genommen. Dies vor allem, weil hier prinzipielle Argumente im Vordergrund stehen.

Im konkreten Fall werden die geschätzten Anteilswerte für die Zustimmung zu drei Statements miteinander verglichen. Hierbei ergeben sich nach Ansicht der Beschwerdeführer gravierende Unterschiede:

a) Welt/Emnid-Umfrage: Özil und Gündogan aus WM-Kader streichen

36 % Zustimmung, 57 % nicht nötig

b) RTL/Forsa: Finden Sie es grundsätzlich in Ordnung, dass die deutschen Fußballnationalspieler Mesut Özil und Ilkay Gündogan mit dem türkischen Staatspräsidenten Erdogan einen PR-Termin in London wahrgenommen haben?

   Nein, hätten sie nicht tun sollen: 61 %, Ja: 29 %

   Mesut Özil und Ilkay Gündogan sollten aufgrund des Treffens mit Erdogan nicht für die kommende WM nominiert werden.

   Nein, das ist kein Grund dafür: 71 %, Ja: 25 %

   Basis: Alle Befragten, die das Verhalten von Özil und Gündogan nicht in Ordnung finden.

c) Civey: Sollten Gündogan und Özil nach ihrem Treffen mit Präsident Erdogan weiter für die deutsche Fußballnationalmannschaft spielen?

   Nein, auf keinen Fall: 56,1 %, Eher nein: 22,6 %

Es fällt auf, dass die Frageformulierungen zwischen den einzelnen Umfragen variieren (Aus WM-Kader streichen / Nicht für die kommende WM nominieren / Nicht für die deutsche Fußballnationalmannschaft spielen). Zudem werden bei Forsa nur die Personen befragt, die Özil und Gündogan bei der vorherigen Frage abgestraft hatten. Schließlich verwendet Civey eine Ordinal-Skala während die anderen Umfragen nur ja/nein-Antworten zuließen.

Grundsätzlich sind derartige Vergleiche problematisch, da es immer gewisse Unterschiede in Stimulus, Kontext und Skala gibt, die ein anderes Antwortverhalten hervorrufen. Und es gibt beim Vergleich von Telefon- und Internet-Befragung auch Mode-Effekte. Hier äußern sich Respondenten am Telefon zurückhaltender als im Internet, vergleiche Couper (2008, Seite 28) (Designing Effective Web Surveys. Cambridge University Press, New York)

Über die Größe dieser Effekte kann nur spekuliert werden. Insgesamt sind derartige numerische Vergleiche nur schwache Belege für einen Self-Selection-Bias.

7) Resümee

Der Begriff der repräsentativen Stichprobe ist nicht an bestimmtes spezifisches Erhebungsverfahren gekoppelt. Es sind ganz unterschiedliche, teilweise widersprüchliche Auslegungen dieses Begriffs möglich. Im vorliegenden Fall scheint der Begriff eher im Sinne einer allgemeinen Qualitätsaussage über Daten im Sinne von Punkt 1 der Liste von Kruskal/Mosteller benutzt zu werden, wobei den „repräsentativen Daten“ die höchste Qualitätsstufe zukommt. Eine Einstufung der Daten von Civey als repräsentativ wird wegen grundsätzlicher methodischer Vorbehalte abgelehnt. Diese methodischen Vorbehalte werden jedoch nicht genau ausgeführt. Es konnte belegt werden, dass die Umfrageergebnisse von Civey mit anerkannten statistischen Methoden Rückschlüsse auf die Grundgesamtheit zulassen. Hierbei basieren die Voraussetzungen für die Methodik auf den gleichen Annahmen, die die Umfrageinstitute für ihre Telefonumfragen in Anspruch nehmen. Andernfalls könnten selbst die Beschwerde führenden Umfrageinstitute für ihre Daten keine Repräsentativität beanspruchen.

Gleichwohl könnte Civey die Dokumentation seiner Vorgehensweise verbessern, um so die Nachvollziehbarkeit seiner Sampling-Strategie und Hochrechnungsverfahren zu erhöhen. Dieses ist jedoch kein prinzipieller Mangel der eingesetzten Methodik, die in der

wissenschaftlichen Literatur gut belegt ist und mittlerweile auch Eingang in Lehrbücher zur Stichprobentheorie gefunden hat.

Insgesamt ist das Internet eine interessante, schnelle und auch preiswerte Ergänzung für die Umfrageforschung geworden. Dieses Argument gilt umso mehr, je größer der Anteil der älteren Menschen ist, die das Internet aktiv nutzen. Die Möglichkeit, diese Informationsquelle zu nutzen, steht auch den anderen Meinungsforschungsinstituten offen.

Berlin, der 22. Nov. 2018