Research Methodology Group
Department of Social Sciences
University of Duisburg-Essen

Warum ausschließlich 'online' durchgeführte Bevölkerungsumfragen nicht 'repräsentativ' sind

Working Paper 11/2018

Rainer Schnell

16. November 2018

Einleitung

Laien erkennt man in der Regel daran, dass sie entweder keine Fachbegriffe oder diese falsch verwenden. Bei Journalisten findet man z.B. häufig den Begriff 'Demoskopie', der außerhalb des Instituts für Demoskopie kaum und in wissenschaftlichen Kontexten nur in den Niederungen der Kommunikationsforschung verwendet wird.
Noch weiter verbreitet bei Laien ist der Begriff 'Repräsentativität'. In der Geschichte der Statistik dauerte es von 1897 (Kiaer) bis 1934 (Neyman), bis geklärt war, dass man für verallgemeinerbare Stichproben zur Schätzung eines Populationsparameters Stichproben benötigt, bei denen für jedes Element der Grundgesamtheit die Auswahlwahrscheinlichkeit vor der Ziehung bekannt ist. Kennt man diese, sind unklare Begriffe wie 'Repräsentativität' unnötig. Daher gibt es diesen Begriff in der mathematischen Statistik nicht.
In Anwendungsgebieten tauchte der Begriff fast nur in zwei Kontexten auf: Entweder, weil man glaubte, einem Publikum den Begriff der bekannten Auswahlwahrscheinlichkeit nicht erklären zu können oder bei den Rückzugsgefechten der Vertreter der Quotenstichproben. In wissenschaftlichen Studien finden sich Quotenstichproben kaum noch, allerdings sowohl in der Marktforschungspraxis als auch in den dunkleren Bereichen der amtlichen Statistik in Deutschland. Da es keine mathematische Rechtfertigung für Quotenstichproben gibt, werden diese – mit Ausnahme von Vertretern des Instituts für Demoskopie – in Deutschland kaum aus Überzeugung, sondern eher als Notbehelf eingesetzt.
In der Marktforschung liegt der Grund für die Verwendung von Quotenstichproben zur Schätzung eines Populationsparameters fast immer in der fehlenden Bereitschaft (meist des Kunden), die Kosten einer Zufallsstichprobe zu tragen. Der Sündenfall der Verwendung von Quotenstichproben in der amtlichen Statistik (wie z.B. bei der EVS) basiert auf dem Verstoß des Gesetzgebers gegen den Rechtsgrundsatz, dass das Gesetz nichts Unmögliches fordern darf: An die amtliche Statistik in Deutschland werden vom Gesetzgeber Anforderungen gestellt, die wissenschaftlich nicht zu verantworten sind und trotzdem von der Behörde (Destatis) umgesetzt werden müssen.

Das Nonresponse-Problem

In der Statistik bestand an der Notwendigkeit der Verwendung von echten Zufallsstichproben (mit bekannten Auswahlwahrscheinlichkeiten) zur Schätzung von Populationsparametern sehr lange kein Zweifel. Dies änderte sich mit den wachsenden Problemen der zurückgehenden Bereitschaft der Bevölkerung an der stetig steigenden Zahl an Umfragen teilzunehmen. Der Erfolg der Verwendung von Befragungen für wissenschaftliche, administrative und kommerzielle Zwecke untergrub durch die wahrgenommene Beliebigkeit der Themen und der Konsequenzenlosigkeit der Teilnahme oder Nichtteilnahme an Befragungen eine der sozialen Voraussetzung für Umfragen. In der Folge entstand seit Anfang der 80er Jahre eine umfangreiche Forschungsliteratur zu fehlenden Daten und Nonresponse allgemein. Trotz ihrer erheblichen Relevanz für wissenschaftliche Forschung und praktische Anwendungen sind die Ergebnisse dieser Forschung nach mehr als 35 Jahren immer noch nicht allgemein bekannt.
Bemerkenswerterweise stammt die beste allgemeinverständliche Zusammenfassung des Problems

aus einer Rede von Donald Rumsfeld im Jahre 2003 (diesen Hinweis verdanke ich Gerald van Belle):

> „There are known knowns. There are things we know that we know. There are known unknowns. That is to say, there are things that we now know we don't know. But there are also unknown unknowns. There are things we do not know we don't know."

In der Sprache der Statistik entspricht dies der heutigen Standardklassifikation von Mechanismen, die zu unvollständigen Daten führen. Hierbei sind Rumsfelds 'known unknows' in der Statistik zwei verschiedene Mechanismen: Vollständig zufälliges Fehlen einer Information und Fehlen einer Information, wobei die Tatsache des Fehlens bekannt ist und der fehlende Wert vorhersagt werden kann (die beiden Mechanismen werden als MCAR und MAR bezeichnet). Fehlen Informationen durch MCAR sind die fehlenden Werte für eine Analyse unwichtig. Fehlen Informationen durch MAR, muss das bei einer Analyse berücksichtigt werden. Dramatischer sind die Konsequenzen beim dritten Mechanismus. Dies sind Rumsfelds 'unknown unknowns'. In diesem Fall wissen wir nicht, welche Informationen uns fehlen: Wir können weder die Tatsache des Fehlens noch den fehlenden Wert selbst vorhersagen. Diesen dritten Mechanismus nennt man MNAR: 'Missing not at random'.
MCAR muss man bei einer Analyse daher nicht berücksichtigen, MAR schon. Liegt MAR vor, können Standardverfahren der Statistik verwendet werden, um Ergebnisse zu korrigieren. Solche Standardmethoden sind z.B. Gewichtungsverfahren (z.B. IPF, propensity weights oder calibration). Bei MAR funktionieren diese Verfahren problemlos, falls die Annahmen erfüllt sind. Die intensive Forschung zu Nonresponse erbringt bei Standardbefragungen der allgemeinen Bevölkerung nahezu immer das gleiche Ergebnis: Nonresponse ist fast immer entweder ein MCAR oder ein MAR-Mechanismus. Teilnahme oder Nicht-Teilnahme an einer Befragung sind hochgradig situativ bedingt und daher entweder fast nicht erklärbar (zufällig) oder durch sehr wenige Variablen sehr gut erklärbar: Erreichbarkeit durch Bildung, Einkommen und Stellung im Lebenszyklus (Anzahl abhängiger Kinder), Verweigerung durch Bildung, Alter und vor allem Arbeitslosigkeit. Es gibt bei traditionellen Befragungen der allgemeinen Bevölkerung trotz mehr als 50 Jahre Forschung mit allen nur denkbaren Methoden keinen Hinweis darauf, dass Nonresponse durch einen MNAR-Mechanismus entsteht. Daher sind diese Art der Ausfälle durch Verweigerung oder Nichterreichbarkeit gut statistisch kompensierbar. Dies gilt auch bei bei sehr großen Anteilen von Ausfällen.
Das gilt nicht für Stichproben, bei denen es Hinweise auf MNAR gibt. Bei MNAR kommt Statistik an ihre Grenzen. Zwar gibt es Vorschläge für statistische Verfahren zur Korrektur von MNAR, aber alle diese Verfahren basieren auf sehr starken inhaltlichen Annahmen über den Ausfallprozess. In der Regel sind viele dieser Annahmen zumindest mit den vorhandenen Daten auch nicht testbar. Wichtig hierbei ist, dass dies inhaltliche (nicht statistische) Annahmen über den genierenden Prozess sind. Diese können nicht einfach als erfüllt angesehen werden.

Wie kann man eine realisierte Stichprobe beurteilen?

Auf diesem statistischen Hintergrund lässt sich die unsinnige Frage nach der 'Repräsentativität' einer Befragung durch die Beantwortung zweier präziser Fragen klären:

1. Besitzt jede Person der Grundgesamtheit eine berechenbare Wahrscheinlichkeit (größer Null) in die Stichprobe zu gelangen?

2. Ist der Mechanismus, der nach der ursprünglichen Auswahl dazu führt das eine ausgewählte Person tatsächlich befragt wird, MCAR, MAR oder MNAR?

Die Beantwortung der ersten Frage setzt voraus, dass die Grundgesamtheit überhaupt eindeutig definiert wurde. Beide Fragen lassen sich weder mit einer Beschreibung des Stichprobenverfahrens, der statistischen Korrekturen noch mit den vorhandenen Daten allein klären. Es sind zusätzliche Vergleichsdaten aus anderen Quellen erforderlich.
Für Standardbefragungen liegen solche Daten vor: Daten der amtlichen Statistik, prozessproduzierte Daten, andere Befragungen. Dazu kommen die Forschungsergebnisse von mehr als einem halben Jahrhundert internationaler Forschung zu Coverage-Problemen und dem Nonresponse-Problem. Ebenso gibt es – zunehmend – Validierungsstudien, die die Ergebnisse von Befragungen mit objektiv gemessenen Daten (GPS, Sensoren) oder administrativen Daten vergleichen. Die Ergebnisse sind natürlich heterogen und zeigen den erheblichen Einfluss der zahlreichen Quellen von Non-Sampling-Errors auf die Ergebnisse. Dass das Ausmaß des resultierenden Total-Survey-Errors weit größer ist als die naiven Schätzformeln einführender Lehrbücher (und z.B. das Politikbarometer) suggerieren, ist lange bekannt.

Online-Erhebungen

Nach diesen etwas trockenen Vorbemerkungen kommen wir nun zu dem Problem von Online-Befragungen. Diese sind in hohem Maße bei Laien populär. Die Popularität ist leicht erklärbar: Online-Befragungen sind einfach implementierbar, erfordern vergleichsweise geringe Investitionskosten und erbringen scheinbar keine offensichtlich falschen Ergebnisse. In der wissenschaftlichen Literatur werden Online-Befragungen in der Regel etwas differenzierter beurteilt. Dabei spielt vor allem die Art der Rekrutierung der Befragten eine Rolle. Hier muss zunächst klar sein, ob für die Stichprobenziehung eine vollständige Liste der Population zur Verfügung steht oder nicht. Nur falls so eine Liste zur Verfügung steht, kann es sich überhaupt um eine Zufallsstichprobe handeln: Sonst sind die Auswahlwahrscheinlichkeiten nicht berechenbar.
Aus diesem Grund basieren z.B. Websurveys der amtlichen Statistik in den Niederlanden (dem CBS) auf Stichproben aus dem Bevölkerungsregister. Personen, die sich nicht an einer solchen Websurvey beteiligen (weil sie z.B. dazu körperlich nicht in der Lage sind oder das Internet nicht nutzen) werden vom CBS dann erneut und zunächst postalisch kontaktiert.
Steht eine solche Personenliste nicht für die Ziehung zur Verfügung, kann die Berechnung von Auswahlwahrscheinlichkeiten nur dann erfolgen, wenn vollständige Listen zusammengefasster Einheiten existieren (z.B. bei Flächenstichproben wie dem Mikrozensus).

Selbstrekrutierte Online-Erhebungen

Basieren Web-Surveys hingegen auf Rekrutierungen im Internet oder 'öffentlichen Aufrufen' handelt es sich nicht um Zufallsstichproben. Die Befragten rekrutieren sich selbst. Das wird euphemistisch gelegentlich 'opt-in'-Auswahl genannt, aber das ist nicht der technische Begriff. Stichproben durch Selbstrekrutierung werden zumeist als 'non-probability sample' oder 'convenience sample' bezeichnet, aber die traditionelle Bezeichnung in der deutschsprachigen Bezeichnung ist etwas präziser: Es handelt sich um 'willkürliche Stichproben'. Willkür ist aber kein Zufall. Man kann zwar eventuell eine Regel angeben (z.B.: 'Jeder der, will' oder 'Jeder, den ich kenne'), diese Regel führt aber nicht zu berechenbaren Wahrscheinlichkeiten.
Damit lässt sich in Hinsicht auf selbstrekrutierte Online-Panels festhalten, dass die erste der oben dargestellten Fragen zur Beurteilung von realisierten Stichproben negativ beantwortet werden muss: Es werden systematisch Bevölkerungsanteile ausgeschlossen. Die zweite Frage ist bestenfalls nicht geklärt, es gibt aber keine Belege dafür, dass die Annahme ignorierbarer Ausfälle erfüllt ist.

Ein Beispiel für nicht ignorierbare Ausfälle

Das Statistische Bundesamt gibt für 2013 insgesamt 8 Millionen Deutsche an, die mindestens 75 Jahre alt waren. Da in Regel nur volljährige Personen befragt werden, entspricht dies ca. 13.1% der Grundgesamtheit. Im Allbus 2016 gaben 82% der Befragten an, das Internet privat zu nutzen. Bei den 60-75-Jährigen lag der Anteil nur noch bei 71% und bei den über 75% nur noch bei 27%. Bei den über 75-Jährigen nutzen daher vermutlich mindestens 5.8 Millionen Personen nicht das Internet. Das sind mehr Personen als Schleswig-Holstein, Hamburg und Bremen zusammen Einwohner haben.
Daher sind die Gewichtungsfaktoren für diese Personengruppe in Opt-In-Surveys entsprechend hoch, vor allem bei zusätzlicher Kontrolle der Bildung der Befragten. Als Konsequenz sollen durch die Gewichtung sehr wenige Personen (die zu dem interessiert und gesund genug sind, sich selbst für politische Befragungen zu registrieren) die Mehrheit dieser Altersgruppe, die nicht im Netz vertreten ist, kompensieren. Genau dies kann ein Gewichtungsverfahren nicht leisten.

Kann Statistik nicht doch zaubern?

Laien verstehen in der Regel leider auch nicht die prinzipielle Funktionsweise wissenschaftlichen Vorgehens: Man stellt zunächst eine inhaltliche Hypothese auf, die dann empirisch getestet wird. Diese Tests erfolgen nicht vereinzelt, sondern wiederholt und durch unabhängige Gruppen, wobei das Ziel darin besteht, die Hypothese als falsch nachzuweisen. Erst wenn dies trotz ernsthafter und wiederholter Bemühungen nicht gelingt, wird die Hypothese – vorläufig – als bewährt betrachtet. Die Beweislast für eine Hypothese liegt bei demjenigen, der sie formuliert. Entsprechendes gilt auch für jedes neue statistische Verfahren und jede Forschungsmethode: Ob diese funktionieren, muss erst gezeigt werden. Auch hier liegt die Beweislast bei denjenigen,

die ein neues Verfahren vorschlagen. Für den Test eines Verfahrens müssen unabhängige Forschungsgruppen jedes Detail eines Datenerhebungsprozesses und der Analyseschritte kennen. Bei kommerziellen Online-Surveys ist dies aber nicht der Fall.
Ein Verfahren, das im Detail unklar ist, ist per Definition kein wissenschaftliches Verfahren. Natürlich kann man ein Verfahren trotzdem anwenden. Laien machen dies jeden Tag: Homöopathie ist ein exzellentes Beispiel. Hier gibt es keinen bekannten Mechanismus, die Hypothese widerspricht wissenschaftlichen Grundprinzipien, alle vorhandenen systematischen Studien zeigen kein positives Resultat. Dies hält weder die Hersteller solcher Produkte vom Verkauf ab noch viele Laien vom Kauf.
Die Situation bei selbst-rekrutierten Online-Panels ist ähnlich: Die Konstruktion widerspricht wissenschaftlichen Grundprinzipien, die Gewichtungsverfahren sind in vielen Fällen nicht im Detail dokumentiert, unabhängige Wiederholungen und systematische Vergleiche mit anderen Studien fehlen. Das hält selbst öffentliche Auftraggeber oder vermeintlich kritische Medien nicht davon ab, solche Methoden unhinterfragt anzuwenden. Bei oberflächlicher Betrachtung scheinen diese Verfahren zu funktionieren – ähnlich wie Homöopathie für ihre Adepten.
Der starke Wunsch, etwas umsonst zu bekommen, hat Menschen zu erstaunlichen Handlungen getrieben, so zu endlosen Versuchen, ein Perpetuum mobile zu konstruieren. Das statistische Äquivalent sind 'non-probability samples': Einfach, schnell, fast kostenlos. Leider funktionieren willkürliche Stichproben genauso wenig wie ein Perpetuum mobile. Das wissenschaftliche Grundprinzipien einer fixen Idee widersprechen, hat Menschen selten gestört, wenn es – kurzfristig – lohnend erscheint.

Schlussbemerkungen

Ergebnisse durch willkürliche Stichproben können zu falschen politischen Entscheidungen führen. Daher stellt sich die Frage, wie sich die zunehmende Verbreitung von Zahlen auf der Basis ausschließlich online erhobener Befragungen einschränken lässt.
Man kann versuchen, aus Laien aufgeklärte Laien zu machen, in dem man z.B. Journalisten schult oder bessere Lehrbücher verwendet. So notwendig dies ist, so wenig erfolgversprechend ist dies.
Weder die Verbreitung der Psychoanalyse noch der Homöopathie haben sich durch Bildung oder Appelle einschränken lassen. In beiden Fällen hat sich aber weitgehend verhindern lassen, dass Steuergelder dafür ausgegeben werden oder es an (ernsthaften) Hochschulen unterrichtet wird: Psychoanalyse wird von Literaturwissenschaftlern ernster genommen als von wissenschaftlich arbeitenden Psychologen oder Psychiatern. Möglich war das nur durch den Nachweis, dass weder Psychoanalyse noch Homöopathie empirisch Erfolgsquoten aufweisen, die über den Selbstheilungs- bzw. Placeboquoten liegen.
Für Befragungen sind externe Validierungen schwierig, vor allem, wenn der Gegenstand der Forschung subjektive Zustände sind. Daher bietet sich der Vergleich verschiedener Vorgehensweisen bei beobachtbaren Zuständen an: Besitz besonderer Qualifikationen bzw. amtlicher Erlaubnisse, Besitz von Kraftfahrzeugen oder Konsumgütern usw. sowie z.B. objektive Krankheitszustände, jeweils aufgeschlüsselt nach den zahlreichen Zellen demographischer Variablen.

Sind solche Daten nicht öffentlich zugänglich, dann ist ein Vergleich der Ergebnisse unterschiedlicher Vorgehensweisen bei der Datenerhebung aufschlussreich. Bei solchen Vergleichen sollten Zufallsstichproben deutlich näher an den bekannten Parametern liegen als willkürliche Stichproben.
Wenn man die unsinnige Debatte um die 'Repräsentativität' ausschließlich online durchgeführter Befragungen beenden will, benötigt man Vergleichstests, wie sie oben beschrieben wurden. Die Demonstration des wiederholten Versagens eines Verfahrens ist das einzige Mittel seine öffentlich finanzierte Anwendung zu verhindern (zumindest in modernen, demokratischen Gesellschaften). Dazu muss aber die Meinungs- und Marktforschung zu solchen Tests selbst bereit sein.

Literatur

Bellhouse, D.R. (1988): A Brief History of Random Sampling Methods; in: Krishnaiah, P.R./Rao, C.R. (Hrsg.): Handbook of Statistics, Amsterdam, Vol. 6, S.1-14.

Schnell, R. (1993): Die Homogenität sozialer Kategorien als Voraussetzung für 'Repräsentativität' und Gewichtungsverfahren. In: Zeitschrift für Soziologie 22 (1): 16-32.

Schnell, R. (1997): Nonresponse in Bevölkerungsumfragen: Ausmaß, Entwicklung und Ursachen. Opladen.

Schnell, R. (2012): Survey-Interviews: Methoden standardisierter Befragungen. Wiesbaden.

Schnell, R./Noack, M. (2014): The Accuracy of Pre-election Polling of German General Elections. In: MDA - Methods, Data, Analysis 8 (1): 5-24.

Schnell, R./Noack, M./Torregroza, S. (2017): Differences in General Health of Internet Users and Non-users and Implications for the Use of Web Surveys. In: Survey Research Methods 11 (2): 105–123.

Wentland, E.J./Smith, K. (1993): Survey Responses: an Evaluation of their Validity, London.

IMPRINT

Publisher

Research Methodology Group
Department of Social Sciences
University of Duisburg-Essen
Lotharstr. 65
47057 Duisburg
Germany

Editor

Prof. Dr. Rainer Schnell

Homepage

www.uni-due.de/methods