Ludwigs-Maximilian Universität
München

Spezielle Themen der Wirtschafts- und Sozialstatistik

Total Survey Error

Autor:
Severin
Czerny

Betreuung
Prof. Dr. Thomas
Augustin

Abstract

Der totale Umfragefehler ist ein Konzept, das versucht die bei Umfragen entstehenden Fehler, möglichst genau aufzugliedern und zu definieren. Dies mit dem Ziel, die einzelnen Fehler separat betrachtet besser bekämpfen zu können und generell die Aufmerksamkeit auf andere Fehlerquellen zu lenken als das Stichprobenziehen. Neben den Stichprobenfehlern, liegt in dieser Arbeit der Fokus auf den Nicht-Stichprobenfehlern, die hier in drei Unterkategorien zerlegt werden: Die Nichtbeobachtungsfehler, hervorgerufen durch Under- bzw. Overcoverage und Nonresponse, die Beobachtungsfehler, die hier eingeteilt werden in Messfehler, Bearbeitungsfehler und technische Fehler, sowie die Spezifikationsfehler. Außerdem soll ein kurzer Überblick über die Entstehung des Konzepts gegeben werden und einige Gedanken zur Messung und Minimierung des totalen Umfragefehlers zusammengetragen werden.

Inhaltsverzeichnis

1 Einleitung

Die Befragung als Umfragemethode kann wohl als das klassische Konzept der Datenerhebung beschrieben werden. Insbesondere in den empirisch ausgerichteten Disziplinen der Wirtschafts- und Sozialwissenschaften, zählt die Befragung zu den wichtigsten Instrumenten der Datengewinnung. Obwohl sich die Befragungsformen durch technische Neuerungen (Onlinebefragung, etc.) sowie durch die Weiterentwicklung der Theorie zu Befragung (Interviewereffekt, etc.) seit 1973 gewandelt haben, ist die damalige Definition einer Befragung als Interview auch heute noch anwendbar. Scheuch (1973, S.70) definierte damals ein Interview als „ein planmäßiges Vorgehen mit wissenschaftlicher Zielsetzung, bei dem die Versuchsperson durch eine Reihe gezielter Fragen oder mitgeteilter Stimuli zu verbalen Informationen veranlasst werden soll.“

Bei Anwendung von standardisierten Befragungen, bzw. Interviews, auf umfangreiche Populationen, also bei der Durchführung von Umfragen, ist ein wichtiger Aspekt sowohl für die die Umfrage durchführenden Personen als auch die Auftraggeber der Umfrage die Umfragequalität. Diese steht auch im Mittelpunkt des Konzepts des totalen Umfragefehlers (TSE, Total Survey Error).

Der TSE ist ein Konzept, welches darauf abzielt, die statistischen Eigenschaften von Schätzern, die durch Umfragen gewonnen wurden, unter Einbeziehung vieler Fehlerquellen zu beschreiben. Der Fokus liegt hierbei auf den verschiedenen Fehlerquellen, die von dem Operationalisieren, dem Studiendesign, der Stichprobenauswahl, der Datenerhebung bis zu der Auswertung der Daten auftreten können und die Schätzung der verschiedenen Parameter verzerren können. Dabei geht es darum, die Abweichung der Parameterschätzer von dem wahren Wert einer gegebene Population, auf die verschiedenen Fehlerquellen, die auf den unterschiedlichen Stufen der Umfrage auftreten können, zurückzuführen. Das hängt eng mit dem verwandten Term der Umfragegenauigkeit (survey accuracy) zusammen. Dieser beschreibt genau das: Die Abweichung des Parameterschätzers von dem zugrunde liegendem wahren Wert, bzw. die Abweichung des vorhandenem Schätzers von dem „idealem“ Schätzer, also jenem den eine Umfrage unter idealen Bedingungen hervorgebracht hätte (Biemer; (2010); S. 817). Der TSE konzentriert sich im Grunde darauf, die Einflüsse und verschiedenen Ebenen einer Umfrage voneinander zu trennen, um Fehler oder Abweichungen genauer auf bestimmte Aspekte zurückführen zu können. Dabei liegt der Fokus auf jenen Faktoren die tendenziell eine verzerrende Wirkung auf die Schätzer haben und messbar sind.

Es ist jedoch ein Irrglaube, dass der TSE ein feststehendes Konzept ist.

Viel mehr ist der Term nicht einheitlich definiert und verschiedene Forscher fassen verschiedenen Fehler unter diesem Term zusammen. Auf der einen Seite sind das Bestandteile, welche sich ohne große Abänderung von bestehenden Umfragedesigns messen lassen, so ist es inzwischen weitgehend akzeptiert, dass die Stichprobenvarianz sich in den meisten Zufallsstichproben berechnen lässt (Groves, Lyberg; (2010); S. 850), aber andere Faktoren, auf der anderen Seite, lassen sich nicht ohne weiteres berechnen, oder benötigen zur Berechnung Annahmen, die oft nicht erfüllt sind. Der TSE lässt sich deswegen eher als ein theoretisches Rahmenwerk beschreiben. Dieses kann für verschiedene Sachen benutzt werden. Auf einer praktischen Ebene kann er beispielsweise Umfragedesignern als Planungskriterium dienen, in dem Sinne, dass bei der Entscheidung zwischen verschiedenen Umfragedesigns jenes mit dem geringsten erwartetem TSE gewählt werden sollte. Auf einer theoretischen Ebene hilft das Konzept dabei, die Aufmerksamkeit auf lange vernachlässigte Punkte der Umfragequalität zu lenken. In der vorliegenden Arbeit soll zuerst ein kurzer Überblick über die Entwicklung des Konzepts gegeben werden, danach werden die einzelnen Faktoren und die Zusammensetzung des TSE untersucht, um sich dann kurz Gedanken über die Messung bzw. die Minimierung des Fehlers zu machen und danach die Stärken und Schwächen des Konzepts betrachtet.

2 Geschichte des Totalen Umfragefehlers

Die Geschichte des Konzepts des TSE beginnt mit einem Artikel in der soziologischen Zeitschrift „American Sociological Review“ aus dem Jahr 1944. Dieser Artikel behandelt mögliche Fehler in Umfragen und identifiziert dreizehn verschiedene Faktoren die die Umfragequalität beeinträchtigen können. Außerdem lenkt er die Aufmerksamkeit darauf, zu versuchen alle diese Fehler zu minimieren und nicht bloß die offensichtlichsten (Deming; (1944); S. 359).

1. Variability in response;
2. Differences between different kinds and degrees of canvass;
   (a) Mail, telephone, telegraph, direct interview;
   (b) Intensive vs. extensive interviews;
   (c) Long vs. short schedules;
   (d) Check block plan vs. response;
   (e) Correspondence panel and key reporters;
3. Bias and variation arising from the interviewer;
4. Bias of the auspices;
5. Imperfections in the design of the questionnaire and tabulation plans;
   (a) Lack of clarity in definitions; ambiguity; varying meanings of same word to different groups of people; eliciting an answer liable to misinterpretation;
   (b) Omitting questions that would be illuminating to the interpretation of other questions;
   (c) Emotionally toned words; leading questions; limiting response to a pattern;
   (d) Failing to perceive what tabulations would be most significant;
   (e) Encouraging nonresponse through formidable appearance;
6. Changes that take place in the universe before tabulations are available;
7. Bias arising from nonresponse (including omissions);
8. Bias arising from late reports;
9. Bias arising from an unrepresentative selection of date for the survey, or of the period covered;
10. Bias arising from an unrepresentative selection of respondents;
11. Sampling errors and biases;
12. Processing errors (coding, editing, calculating, tabulating, tallying, posting and consolidating);
13. Errors in interpretation;
    (a) Bias arising from bad curve fitting; wrong weighting; incorrect adjusting;
    (b) Misunderstanding the questionnaire; failure to take account of the respondents' difficulties (often through inadequate presentation of data); misunderstanding the method of collection and the nature of the data;
    (c) Personal bias in interpretation.

Abbildung 1: Demings Faktoren der Umfragequalität

Auch wenn sich diese Aufzählung von späteren Definitionen des TSE unterscheidet, gibt es einige Überschneidungen. So schließt die Liste Demings als Fehlerquellen unter anderem Nonresponse, Inteviewereffekte, Stichprobenziehung und verschiedene Datenverarbeitungs- sowie Interpretationsfehler mit ein. Die Aufzählung der verschiedenen Fehlerquellen und der Aufruf keine davon zu vernachlässigen, ist ein Schritt weg davon nur die Stichpro-

benziehung als Fehlerquelle zu betrachten. Diese damalige Fixierung auf die Stichprobenziehung, kann allerdings auch daran liegen, dass zu der Zeit die Aussagekraft von Stichproben noch kein Gemeinplatz war und deswegen von ihren Befürwortern versucht wurde voranzutreiben (Groves, Lyberg; (2010); S. 853). Alles in allem lässt sich trotzdem sagen, dass es einige Jahre gedauert hat, bis andere Faktoren ein ähnlich hoher Einfluss auf die Qualität von Umfragen zuerkannt wurde. Als unmittelbarer Vorgänger des Begriffs des TSE, kann der Ausdruck „total survey design“ verstanden werden. Dieser wurde 1974 von Dalenius als Teil eines umfangreiches Forschungsprojekt mit dem Name „Errors in Surveys“ eingeführt und bezieht sich auf drei Blickwinkel auf Umfragen: Die Anforderungen, die Spezifizierung und die Durchführung der Umfrage (Groves, Lyberg; (2010); S. 853). Das Buch „Total Survey Error“ welches 1979 von Anderson, Kasper und Frankel herausgegeben wurde, gab dem Konzept schließlich den bis heute benutzten Namen. Darin bemühen sich die Autoren um eine Aufschlüsselung des TSE, unter anderem beschreiben sie die Unterschiede zwischen Varianz und Bias, zwischen Fehlern bei der Stichprobengenerierung und den restlichen Fehlern („sampling errors“ und „non-sampling errors“) und dem Unterschied zwischen Beobachtungsfehlern wie Mess- und Verarbeitungsfehlern und Nichtbeobachtungsfehlern wie Undercoverage oder Nonresponse. Diese Unterscheidung zwischen Beobachtungs- und Nichtbeobachtungsfehlern, kann als ein Verdienst des Konzepts des TSE gesehen werden. Zu Beobachtungsfehlern ist inzwischen eine große Anzahl an Literatur vorhanden, welche die verschiedenen Aspekte wie den Einfluss des Interviewers, des Fragebogendesigns, der Art der Datenerhebung und deren gegenseitigen Zusammenhänge untersucht. Verschiedene andere Aspekte wie die Auswertung von Paradaten und der Idee des „fitness for use“ von Umfragen, welche sich auf die Anwenderfreundlichkeit von Daten bzw. Ergebnissen bezieht haben das Konzept noch verfeinert (Groves, Lyberg; (2010); S. 856). Zusammenfassend kann gesagt werden, dass das Konzept des TSE aus der Erkenntnis hervorging, dass nicht nur Stichprobenfehler die Qualität von Umfragen beeinträchtigen und der daraus folgenden Aufschlüsselung und Untersuchung der Nichtstichprobenfehler.

3 Der totale Umfragefehler als Teil der totalen Umfragequalität

Der TSE ist Teil des weitergehenden Konzepts der totalen Umfragequalität. Dieses dient der Einschätzung inwiefern eine Umfrage „fit for use/purpose“ ist. Während der TSE einen eher operativen Fokus hat, zielt die totale Umfragequalität auf andere Dimensionen einer Umfrage, wie die Relevanz, die Reliabilität oder die Zugänglichkeit von Daten. Das „fitness of use“ Konzept setzt sich mit dem Problem auseinander, dass Produzenten und Nutzer von Umfragedaten oft unterschiedliche Interessen haben und die Umfragequalität anhand von verschiedenen Faktoren beurteilen. Personen, die in die Erhebung von Daten involviert sind, legen häufig einen stärkeren Fokus auf die Datenqualität, beispielsweise auf eine möglichst große Stichprobe, eine hohe Antwortrate und eine gute Abdeckung der Zielpopulation und würden dadurch einen Großteil der Anstrengungen und des Budgets darauf verwenden, statistisch einwandfreie Daten zu erzeugen, um bestimmte Schätzer möglichst akkurat zu schätzen (Biemer; (2010); S. 818). Die Auftraggeber bzw. Nutzer von Umfragedaten hingegen, halten akkurate Schätzer oft für ohnehin gegeben und legen einen stärkeren Fokus auf die Zugänglichkeit und Benutzerfreundlichkeit von Daten. Außerdem ist die Aktualität der Daten oft wichtig und das die Fragen das messen was gemessen werden soll (Biemer; (2010); S. 818). Anhand dieser Gegenüberstellung wird deutlich, dass totale Umfragequalität über die Dimension des TSE hinausgeht und dass es zwei, sich teilweise gegenüberstehende Aspekte des Konzepts von Qualität gibt: Auf der Einen Seite die Abwesenheit von Fehlern und auf der Anderen das Eingehen auf die Bedürfnisse des Datennutzers (Biemer; (2010); S. 818). Die Abwesenheit von Fehlern ist gleichartig zu dem Konzept des TSE, während das Eingehen auf die Bedürfnisse des Endnutzers der Daten, nur in der Zeit des Umfragedesigns umgesetzt werden kann. Wenn auf die Bedürfnisse des Datennutzers nicht eingegangen wird, besteht die Gefahr, dass die Daten „unfit for use“ (Biemer; (2010); S. 818) sind, zum Beispiel weil sie zu spät veröffentlicht werden oder nicht leicht zugänglich sind. Es kann die Situation entstehen, dass die Daten statistisch korrekt erhoben und ausgewertet wurden, aber nicht nutzbar sind, ihnen fehlt totale Umfragequalität, also von beiden Standpunkten als qualitativ angesehenen Daten: Von den Produzenten als auch den Nutzern. Ebenso wie für den TSE gibt es für Umfragequalität keine allgemeingültige Definition. Es gibt jedoch einige Faktoren, die meistens als Teil der Umfragequalität erachtet werden. Diese sind in Tabelle 1 abgebildet:

Genauigkeit	Der TSE ist so klein wie möglich
Glaubwürdigkeit	Die Daten werden als vertrauenswürdig angesehenen
Vergleichbarkeit	Vergleiche mit anderen Daten sind zulässig und möglich
Benutzerfreundlichkeit	Eine gute Dokumentation sowie Metadaten liegen vor
Relevanz	Die Daten befriedigen die Bedürfnisse des Nutzers
Zugänglichkeit	Der Zugang zu den Daten ist benutzerfreundlich
Aktualität/Pünktlichkeit	Datenlieferung an ausgemachten Terminen
Vollständigkeit	Daten entsprechen den Anforderungen der Analyse-Methode
Kohärenz	Schätzer aus verschiedenen Quellen können kombiniert werden

Tabelle 1: Faktoren, die die Umfragequalität beeinflußen (vgl.: Biemer; (2010); S. 819)

Einige dieser Faktoren sind jedoch qualitativer Natur und daher schwer zu quantifizieren, so zum Beispiel die Zugänglichkeit oder die Glaubwürdigkeit. Daher gestaltet es sich schwierig, eine einzige Messgröße zu konstruieren, mit welcher die totale Umfragequalität quantifiziert werden könnte. Eine mögliche, bereits vorhandene, Alternative stellen die Qualitäts-Berichte („quality reports“) da, welche für die verschiedenen Dimensionen der Umfragequalität eine Beschreibung der Stärken und Schwächen der jeweiligen Umfrage liefern (Biemer; (2010); S. 820). Das Konzept der totalen Umfragequalität ist auch bei dem Umfragedesign nützlich, so können sich Datennutzer und -produzenten auf verbindliche Zusagen für alle Dimensionen der Umfragequalität einigen und beschließen, auf welche Dimensionen ein besonderer Fokus gelegt werden soll, um Fehler in diesem Bereich möglichst zu vermindern. Die beste Umfrage ist also jene, welche die Bedürfnisse des Nutzers sowie des Produzenten am besten trifft und somit die höchst totale Umfragequalität besitzt.

4 Der totale Umfragefehler

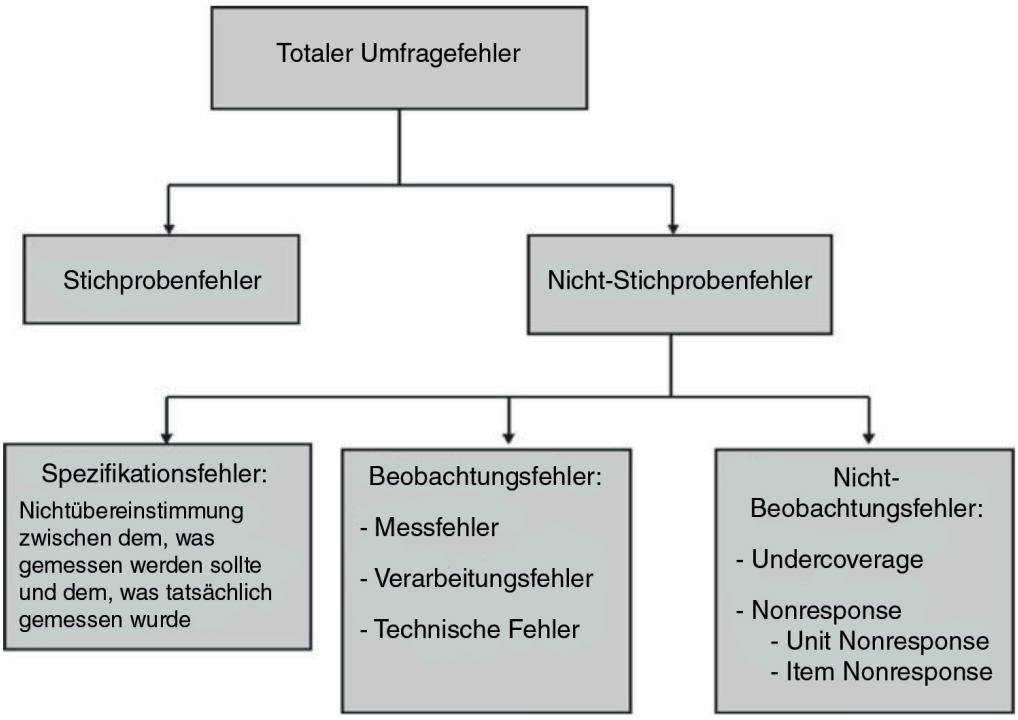

Abbildung 2: Mögliche Zerlegung des Totalen Umfragefehlers nach Faulbaum (2014; S. 440)

Die verschiedenen Fehler, die während einer Umfrage entstehen können, ergeben aufsummiert den TSE. Dabei werden alle möglichen Fehler in Betracht gezogen, von dem Umfragedesign, über die Datengewinnung bis hin zur Aufbereitung und der Auswertung der Daten. Die mit dem TSE verbundene Absicht ist, „die Abweichung der Parameterschätzung (...) auf verschiedene Fehlerquellen zurückzuführen, die auf unterschiedlichen Stufen des Erhebungsprozesses in unterschiedlichem Ausmaß zur Gesamtabweichung der Parameterschätzung vom wahren Populationswert beitragen“ (Faulbaum; (2014); S. 439). Eine ideale Umfrage kann somit über die Minimierung des TSE, unter Berücksichtigung von Beschränkungen durch Anforderungen des Datennutzers (z.B. Pünktlichkeit oder Vergleichbarkeit; siehe auch „3. Der totale Umfragefehler als Teil der totalen Umfragequalität“) beschrieben werden. Unter dem Gesichtspunkt der totalen Umfragequalität ist hierbei das Ziel nicht alle Stufen des Umfrageprozesses komplett fehlerfrei zu gestalten – mit zeitlichen sowie monetären Beschränkungen (und selbst ohne diese)

ein unrealistisches Unterfangen – sondern die „schlimmsten“ Fehler zu vermeiden und die restlichen in dem Maße zu kontrollieren, dass sie tolerierbar werden. Der erste Schritt um Fehler so weit zu kontrollieren, dass sie tolerierbar oder sogar unbedeutend werden, ist sie genaustens zu identifizieren. Daher ist es nötig den TSE soweit als möglich zu zerlegen und genaustens zu definieren. Eine mögliche Zerlegung der Fehlerquellen, aus denen sich der TSE zusammensetzt, ist in Abbildung 2 dargestellt.

Zunächst lassen sich die Fehler, die bei der Durchführung einer Umfrage auftreten können, in Stichproben sowie Nichtstichprobenfehler aufteilen. Stichprobenfehler entstehen immer und zwar einfach dadurch, dass statt der vollständigen Zielpopulation nur eine Stichprobe dieser betrachtet wird, dieser Stichprobenfehler kann noch einmal in einen Schätzfehler sowie einen Auswahlfehler zerlegt werden (Faulbaum; (2014); S. 440). Die Nicht-Stichprobenfehler können in drei weiter Unterklassen zerlegt werden. Zum Einen die Klasse der Beobachtungsfehler, diese schließt Messfehler, Verarbeitungsfehler sowie technische Fehler mit ein, dann Nichtbeobachtungsfehler, diese sind entweder Fehler aufgrund von Under- oder Overcoverage, oder Fehler wegen fehlenden Antworten („Nonresponse“) und schließlich noch Spezifikationsfehler. Diese Fehlerquellen sollen im Folgendem beschrieben und so gut wie möglich definiert werden.

4.1 Stichprobenfehler und Schätzqualität

Eine Entscheidung die immer am Anfang des Umfragedesigns steht ist, ob nur ein Teil der Elemente untersucht werden soll oder ob alle Elemente der Grundgesamtheit in die Umfrage miteinzubeziehen sind. Gegen eine Vollerhebung spricht vor allem der finanzielle sowie zeitliche Aufwand, der mit einer solchen Erhebungsprozess verbunden ist. So verursachte der Zensus im Jahr 2011 finanzielle Kosten von mehr als 700 Mio. Euro und das, obwohl es keine Vollerhebung war. Die Kosten für eine traditionelle Volkszählung würden noch einmal erheblich über denen für den registergestützten Zensus von 2011 liegen (www.zensus2011.de). Aber es gibt auch forschungspraktische Gründe, die gegen eine Vollerhebung sprechen. Die Ressourcen und Mittel die für eine Untersuchung zur Verfügung stehen sind in der Regel begrenzt und können effektiver genutzt werden, wenn nur eine relativ kleine Anzahl an zu untersuchenden Elementen vorliegt. So können durch verschiedene Methoden bei Stichprobenerhebungen teilweise eine bessere Ausschöpfung und Antwortraten erreicht werden als bei Vollerhebungen (Häder; (2014); S. 283). Bei einer Stichprobe handelt es sich also um eine Auswahl von Elementen aus einer Grundgesamtheit. Diese Auswahl kann auf verschiedene Arten getroffen werden. Idealerweise erfolgt die Auswahl auf Grund bestimmter statistischer

Regeln, sodass aus den Resultaten Schlüsse über die zu Grunde liegende Gesamtheit gezogen werden können. Eine wichtige Voraussetzung dafür, dass auf die Grundgesamtheit geschlossen werden kann ist, dass die Stichprobe eine Zufallsauswahl ist. Solche zeichnen sich dadurch aus, dass „die Wahrscheinlichkeit für jedes Element der Grundgesamtheit, ein Element der Stichprobe zu werden, berechnet werden kann und größer als null ist“ (Häder; (2014); S. 284). Die bei diesem Prozess auftretenden Fehler lassen sich noch einmal in Schätzfehler sowie Auswahlfehler unterteilen. Schätzfehler entstehen dadurch, dass infolge der Zufallsauswahl bei jeder erneuten Ziehung eine von der ursprünglichen Stichprobe verschiedene Stichprobe aufkommt, mit ihrerseits verschiedenen Schätzern, die sodann mehr oder weniger stark von dem wahren Populationswert abweichen. Auswahlfehler entstehen dadurch, dass die Wahrscheinlichkeit eines Elements Teil oder nicht Teil der Stichprobe zu werden unbekannt ist bzw. falsch angegeben wird und dadurch zu verzerrten Schätzern führt. So sind beispielsweise bei Internetumfragen die Auswahlwahrscheinlichkeiten meistens unbekannt und es kann somit nicht mehr von einer Zufallsstichprobe gesprochen werden, da eine der Voraussetzungen die Kenntnis bzw. Berechenbarkeit dieser Auswahlwahrscheinlichkeiten ist. Wichtige Mittel um die Schätzqualität beurteilen zu können sind der Bias, die Varianz, der Designeffekt, sowie der mittlere quadratische Fehler (MSE) des Schätzers. Auf diese Aspekte soll im Folgenden kurz eingegangenen werden, der MSE wird an einer anderen Stelle jedoch noch einmal einer tieferen Betrachtung unterzogen (siehe 5. Messung und Minimierung des Totalen Umfragefehlers).

4.1.1 Bias

Bei Betrachtung der Stichprobenfehler, beschreibt der Bias die Verzerrung des Schätzers, mit der aufgrund des gewählten Auswahlverfahrens der Stichprobe gerechnet werden muss. Bei einer Zufallsauswahl, muss so beispielsweise von jeder möglichen Stichprobe die Wahrscheinlichkeit, mit der diese Stichprobe ausgewählt wird, bekannt sein. Der Bias ist hierbei eine Größe, die unabhängig von der tatsächlichen Erhebung auf der Basis statistischer Ableitungen berechnet werden kann und gibt an, „wie genau ein Populationsparameter θ im Durchschnitt alle möglichen Stichproben des gleichen Umfangs aus der Zielpopulation U von einem Schätzer (...) geschätzt wird“ (Faulbaum; (2014); S. 441).

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta \tag{1}$$

Formal ist der Bias als der Erwartungswert des Schätzers minus den Populationsparameter definiert. Falls der Bias gleich Null ist, so ist der Schätzer

erwartungstreu. Sein Erwartungswert ist also gleich dem wahren Wert des zu Schätzenden Parameters.

4.1.2 Varianz

Ein erwartungstreuer Schätzer, bzw. ein Schätzer mit möglichst kleinem Bias ist anzustreben, ist aber nicht das einzige Kriterium um einen Schätzer zu beurteilen, auch eine große Rolle spielt die Varianz. Eine kleine Varianz ist ein ebenso wichtiges Kriterium für die Qualität des Schätzers, der Schätzer sollte also möglichst wenig um den zu schätzenden Wert schwanken. Die Varianz ist also ein Maß für die Präzision eines Schätzers und gibt an, wie stark die Schätzung von Stichprobe zu Stichprobe variiert (Faulbaum; (2014); S. 442).

4.1.3 MSE

Das Ziel ist also ein erwartungstreuer Schätzer mit möglichst geringer Varianz. Dieses Verhältnis lässt sich über den mittleren quadratischen Fehler ausdrücken, dieser beschreibt die mittlere quadratische Abweichung des Schätzers vom Populationsparameter und kann als die Summe aus dem quadrierten Bias und der Varianz des Schätzers berechnet werden.

4.1.4 Designeffekt

Der Designeffekt ist ein Mittel den Präzisionsgewinn oder -verlust zu messen der entsteht, wenn anstatt einer einfachen Zufallsauswahl ein anderes Stichprobendesign verwendet wird. Dabei wird die Varianz des Schätzers unter einem gegebenem Stichprobendesign mit der Varianz des Schätzers bei einer einfache Zufallsstichprobe verglichen. Falls die Varianz des anderen Schätzers größer als die Varianz des Schätzers der einfachen Zufallsauswahl ist, ist bei einer Entscheidung für das andere Stichprobendesign von einem Präzisionsverlust auszugehen und umgekehrt (Faulbaum; (2014); S. 443-444).

4.2 Nicht-Stichprobenfehler

Neben den Fehlern die bei der Generierung der Stichprobe und bei dem Stichprobendesign auftreten können, gibt es weitere Faktoren die eine Verzerrung bewirken und die nicht auf das Stichprobendesign zurückzuführen sind. Dies lassen sich in drei Klassen einteilen:

1. Nichtbeobachtungsfehler sind Fehler, die weder auf die Stichprobenauswahl noch auf die Beobachtung der Daten zurückzuführen sind.
2. Beobachtungsfehler sind Fehler, die durch die Erhebung oder Verarbeitung der Daten entstehen.
3. Spezifikationsfehler entstehen wenn das, was gemessen wird sich von dem unterscheidet, was eigentlich gemessen werden soll.

4.2.1 Coverage-Fehler

Um eine Zufallsauswahl einer Gesamtpopulation durchführen zu können, wird eine Liste (auch: Auswahlgrundlage) aller in ihr enthaltenen Elemente benötigt. Das an sich stellt häufig schon ein Problem dar: Wie ist es zum Beispiel möglich, ein Verzeichnis aller Raucher und Raucherinnen in Deutschland zu erstellen? Falls in der zur Verfügung stehendem Auswahlgrundlage nicht alle Elemente enthalten sind, spricht man von Undercoverage. Dies stellt in vielen heute gebräuchlichen Umfragetypen, wie in Internetumfragen, aber auch in den viel genutzten random digit dialed Telefonumfragen, ein Problem dar. In den USA beispielsweise verlassen sich viele offizielle Statistiken auf Telefonumfragen, die wichtigsten Schätzungen über Kriminalität basieren auf Telefonumfragen (Peytchev; (2010); S. 287). Diese Telefonumfragen erfassen aber nur einen zunehmend kleineren Teil der Bevölkerung, da immer mehr Menschen kein Festnetzanschluss mehr besitzen und nur noch über den Mobilfunk zu erreichen sind. Diese Menschen sind also kein Teil der Auswahlgesamtheit. Umgekehrt kann es zu Overcoverage kommen, falls die Auswahlgrundlage Elemente aus der Grundgesamtheit doppelt enthält oder Elemente enthält die überhaupt nicht Teil der Grundgesamtheit sind. So sind zum Beispiel in Karteien der Meldebehörden alle Personen jedes Alters enthalten, die Zielpopulation besteht aber nur aus Personen im Alter von über 18 Jahren, alle Personen im Alter unter 18 Jahren wären also nicht auswählbar. Zu einem ernsthaften Problem werden Coverage-Fehler, wenn sie nicht zufällig auftreten, also wenn bestimmte Personengruppen durch Coverage-Fehler in der Auswahlgrundlage unter- bzw. überrepräsentiert sind und sich diese Personengruppen systematisch von den restlichen Personen unterscheiden. Um

auf das Beispiel von Telefonbefragungen zurückzukommen, konnte Peytchev (2010; S. 295) zeigen, dass es in den USA signifikante Unterschiede zwischen Festnetz und nur Mobilfunk benutzenden Personen gab. So waren alleinige Mobilfunknutzer signifikant öfter jünger und nicht weiß als Personen mit Festnetzanschluss. Das kann im Rahmen der Stichprobenziehung zu ernsthaften Problemen führen, da der systematische Ausfall von Elementen für die Ziehung zu verzerrten Schätzern führen kann (Häder; (2014); S. 284).

Der Coverage-Fehler lässt sich wie folgt definieren:

$$\bar{\mathbf{Y}}_c - \bar{\mathbf{Y}} = \frac{\mathbf{U}}{\mathbf{N}}(\bar{\mathbf{Y}}_c - \bar{\mathbf{Y}}_u) \tag{2}$$

wobei:

$\bar{\mathbf{Y}}$: Mittelwert einer Variablen Y in der gesamten Zielpopulation

$\bar{\mathbf{Y}}_c$: Mittelwert der Population, die in der Auswahlgrundlage enthalten ist

$\bar{\mathbf{Y}}_u$: Mittelwert der Zielpopulation, die nicht in der Auswahlgrundlage enthalten ist

$\mathbf{N}$: Gesamtanzahl der Elemente in der Zielpopulation

$\mathbf{U}$: Gesamtanzahl der auswählbaren Elemente, die nicht in der Auswahlgrundlage enthalten sind

Der Coverage-Fehler lässt sich hierbei über die Differenz zwischen dem Mittelwert des durch die Auswahlgrundlage abgedeckten Teil der Zielpopulation und dem Mittelwert der gesamten Zielpopulation beschreiben (Peytchev; (2010); S. 289).

4.2.2 Nonresponse-Fehler

Verzerrungen durch Nonresponse lassen sich in zwei Kategorien einteilen: Den vollständigem Ausfall aller Messungen eines Stichprobenelements (Unit-Nonresponse) sowie den Ausfall einzelner Messungen auf bestimmten Variablen (Item-Nonresponse). Von Unit-Nonresponse spricht man, wenn die in einer Stichprobe ausgewählten Personen die Teilnahme an der Befragung verweigern, die Personen nicht erreicht werden können oder auf Grund von z.B. Sprachproblemen nicht an der Umfrage teilnehmen können. Hierbei beschreibt die Ausschöpfungsquote, wie viele Personen der ursprünglich gezogenen Stichprobe, im Endeffekt an einer Umfrage teilgenommen haben. Ob diese Rate über die Zeit zu oder abnimmt, ist Gegenstand einer wissenschaftlichen Debatte. So gibt es verschiedene Studien, die zu dem Schluss kommen,

dass Antwortraten generell über die letzten Jahre gesunken sind und Studien die dieser Auffassung widersprechen und diese Senkung eher auf verschiedene Umfragemodi zurückführen (Engel; (2014); S. 331-332).

Der größte Teil der Totalausfälle ist auf Teilnahmeverweigerung und Nichterreichbarkeit zurückzuführen. Von Verweigerung spricht man, wenn die zu befragende Person zwar erreicht werden konnte, die Teilnahme aber ablehnt wird. Nicht-Erreichbarkeit bezeichnet die Situation, dass Personen aus unterschiedlichen Gründen während der Durchführung der Umfrage nicht erreicht werden konnten. Es gibt verschiedene Ursachen die zu Unit-Nonresponse führen können. Eine davon ist der Befragungsmodus. Die Antwortraten unterscheiden sich in Abhängigkeit von der Art der Befragung. Hohe Antwortraten von ca. 50% können bei persönlicher, mündlicher Befragungen erwartet werden, wohingegen bei telefonischen Befragungen mit Antwortraten von nur 20% gerechnet werden kann (Engel; (2014); S. 332-333). Ein anderer wichtiger Faktor, unabhängig von der Art der Befragung, ist die individuelle Entscheidung für oder gegen die Teilnahme, diese wird oft als „mehrstufige, rationale Kosten-Nutzen-Abwägung“ versucht zu beschreiben, wobei die Antwortraten steigen, falls der Befragte den Eindruck hat, „dass er mit seiner Teilnahme einen wichtigen Beitrag zur öffentlichen Meinungsbildung sowie Wissenschaft und Forschung leistet“ (Engel; (2014); S. 333). Ein praktischer Nutzen kann aber auch durch eine monetäre „Teilnahme-Entschädigung“ dargestellt werden. Auch ein guter, in einem Pre-Test getesteter, Fragebogen trägt zu einer höheren Ausschöpfungsquote bei.

Das bestimmte Personen heutzutage immer schlechter zu erreichen sind, liegt auch an gesellschaftlichen Veränderungen, so gibt es beispielsweise immer mehr Singelhaushalte, wodurch bei „berufsbedingter Abwesenheit“ (Engel; (2014); S. 335) keine Person des Haushalts mehr erreichbar ist, bzw. besitzen immer mehr, vor allem jüngere Menschen, keinen Festnetzanschluss mehr, oder viele Personen besitzen falsche Postanschriften oder haben E-Mail-Spamfilter, die dazu führen, dass Personen nicht erreicht werden können (Engel; (2014); S. 335). Falls es zu einer Kontaktaufnahme kommt gibt es verschiedene Faktoren die beeinflussen ob eine Befragung zustande kommt. Diese Faktoren können in vier Blöcke eingeteilt werden: Die Soziale Umwelt (Umfrageklima, Urbanisierungsgrad) des Befragten, die Eigenschaften des Haushaltes bzw. des Befragten selbst (Haushaltsstruktur, soziodemografische Charakteristika, psychologische Dispositionen), das Survey Design (Thema, Auswahl der Befragten, Länge des Fragebogens etc.) und den Interviewer (soziodemografische Charakteristika, Erfahrung, Erwartungen) (vgl. Engel; (2014); S.336).

Nonresponse wird nie völlig zu vermeiden sein und wird auch erst dann für die Datenqualität problematisch, wenn sich die teilnehmenden systematisch

von den nicht-teilnehmenden Personen unterscheiden. Es gibt verschiedene Maßnahmen die zu einer Erhöhung der Antwortrate führen können, z.B. die schriftliche Ankündigung der Befragung, mehrmalige und verschiedenartige Kontaktaufnahme, oder die Verwendung von Befragungsanreizen. Außerdem kann systematischen Unterschieden zwischen jenen, die an einer Umfrage teilnehmen und jenen, die das nicht tun durch den Einsatz von sogenannten Gewichtungsfaktoren begegnet werden.

Item-Nonresponse hingegen liegt vor, falls aus unterschiedlichen Gründen nur ein Teil des Fragebogens beantwortet wird. Eine Person nimmt also an der Umfrage teil, beantwortet aber bestimmte Fragen nicht oder bricht das Interview vorzeitig ab, dadurch gibt es nur für einen Teil der Fragen gültige Antworten und es entstehen fehlende Daten. Es gibt verschieden Ursachen, die zu so einer Situation führen können. Beispielsweise, dass es einer Person schwer fällt eine Frage zu beantworten oder sie überhaupt zu verstehen, die mangelnde Motivation der zu befragenden Person, oder falls es sich um eine persönliche Befragung handelt, die soziale Interaktion mit dem Interviewer. Item-Nonresponse hängt auch stark mit der Sensitivität der Frage zusammen.So konnte gezeigt werden, dass mit zunehmender Sensitivität der Frage auch der Nonresponse zunimmt (Engel; (2014); S.342). Der Umgang mit den so entstanden fehlenden Werten verdient eine eigene Betrachtung, es soll hier nur kurz auf verschiedene Möglichkeiten zum Umgang mit diesen hingewiesen werden. Generell liegt den meisten statistischen Methoden, die mit fehlenden Werten umgehen, die Annahme zugrunde, dass diese Werte zufällig fehlen (missing at random). Mit den fehlenden Werten kann dann auf verschiedene Weisen umgegangen werden. Einerseits können Fälle mit fehlenden Werten bei statistischen Berechnungen nicht berücksichtigt werden, oder die fehlenden Werte werden ersetzt. Bei dem Ausschluss von Fällen, kann man grob zwei Herangehensweisen unterscheiden: Einmal die Fälle mit fehlenden Werten von der kompletten Analyse auszuschließen, also auch von Analysen wo in den betrachteten Variablen gültige Werte vorliegen, und zum Anderen die Fälle nur von den Analysen auszuschließen, wo die interessierenden Variablen fehlende Werte aufweisen. Die Alternative zu diesen Ausschlussverfahren besteht darin, fehlende Werte durch inhaltlich plausible Daten zu ersetzen, das wird als Imputation bezeichnet. Hierzu werden z.B. statistische Modelle wie die Regressionsanalyse benutzt, so können anhand der restlichen Variablen bedingte Mittelwerte für die fehlende Variable geschätzt werden.

Der Nonresponse-Fehler kann also über die „Differenz zwischen dem Mittelwert einer Zielvariablen in der Ausgangsstichprobe und dem Mittelwert dieser Variablen in der Teilstichprobe der Respondenten“ (Faulbaum; (2014); S. 446) dargestellt werden.

$$\bar{\mathbf{y}}_r - \bar{\mathbf{y}}_s = \frac{\mathbf{m}_s}{\mathbf{n}_s}(\bar{\mathbf{y}}_r - \bar{\mathbf{y}_m}) \tag{3}$$

wobei:

$\bar{\mathbf{y}}_s$: Mittelwert von y in der vollständigen Stichprobe s
$\bar{\mathbf{y}}_r$: Mittelwert der Respondenten in der Stichprobe s
$\bar{\mathbf{y}_m}$: Mittelwert der Nichtrespondenten in der Stichprobe s
$\mathbf{n}_s$: Gesamtanzahl der Elemente in der Stichprobe s
$\mathbf{m}_s$: Gesamtanzahl der Nichtrespondenten in der Stichprobe s

Der Nonrespone-Fehler nimmt den Wert Null an, wenn der Mittelwert der Teilnehmer und Nicht-Teilnehmer gleich ist und wird umso größer, je größer diese Differenz ist. Außerdem gilt, dass falls die Teilmenge der Umfrageteilnehmer eine zufällige Teilstichprobe der Ausgangsstichprobe ist, keine Verzerrung des Schätzers zu erwarten ist (Faulbaum; (2014); S. 447), aufgrund der geringeren Stichprobengröße allerdings eine Beeinflussung der Varianz.

4.3 Beobachtungsfehler

Beobachtungsfehler sind die Fehler, die bei der Beobachtung der Daten entstehen und lassen sich zur genaueren Betrachtung noch einmal in Messfehler, Verarbeitungsfehler und technische Fehler unterteilen.

4.3.1 Messfehler

Das Ziel jeder Umfrage oder Studie sind exakte, fehlerfreie Messergebnisse. In der Praxis jedoch sind Messfehler nicht vermeidbar, da es nicht möglich ist, etwas ohne Fehler zu messen. Seinen Ursprung hat das Konzept des Messfehlers unter anderem in der klassischen Testtheorie. Diese geht davon aus, dass sich der Wert einer jeden Variable in einen wahren Wert und einen Messfehler additiv zerlegen lässt, sie nimmt somit einen linearen Zusammenhang zwischen der beobachteten Variable $\mathbf{x}$ und der wahren Variable τ an. Die beobachtete Variable wird somit als Summe der wahren Variable und einer Fehlervariable ϵ betrachtet (Faulbaum; (2014); S. 448).

$$\mathbf{x} = \tau + \epsilon \tag{4}$$

Den Fehler kann man somit als Differenz zwischen beobachteten und wahren Wert definieren.

$$\epsilon = \mathbf{x} - \tau \tag{5}$$

Laut der klassischen Testtheorie, entspricht der wahre Wert einer Messung, dem Erwartungswert einer gegen unendlich gehenden Anzahl an Wiederholungsmessungen unter gleichen Bedingungen, der Messfehler wird somit als eine normalverteilte Zufallsvariable mit dem Erwartungswert Null angenommen. Für die praktische Anwendung ist es bedeutend, dass Messfehler durch verschiedene Faktoren entstehen können. Interviewsituationen sind oft anfällig für Messfehler, zum Einen durch Fehler die vom Interviewer abhängen und zum Anderen durch Fehler die vom Befragten abhängen. Fehler die vom Interviewer abhängen, sind beispielsweise absichtliches Fehlverhalten, bestimmte demographische oder andere Merkmale des Interviewers, die das Verhalten der Befragten beeinflussen, Auslassen, Übersehen oder Umformulierung von Fragen und verschiedene Hilfeleistungen, z.B. bei Nicht-Verstehen einer Frage. Aber auch die Befragten können der Grund für Messfehler sein, so z.B. durch Anpassung der Antworten an die vermeintliche Meinung des Interviewers, durch unwahre Antworten aufgrund der sozialen Erwünschtheit von bestimmten Antworten, oder durch den Wunsch nach positiver Selbstdarstellung. Weitere mögliche Quellen für Messfehler sind eine unpassende Befragungssituation, z. B. durch die Anwesenheit nicht an der Befragung beteiligter Dritter, die Formulierung der Fragen und die Gestaltung des Fragebogens, sowie die Art der Befragung (Faulbaum; (2014); S. 449). So gibt es Hinweise darauf, dass es weniger akkurat ist, Daten per Telefonumfrage als per persönlichen Interview zu erheben (Biemer; (2010); S. 823-824).

4.3.2 Verarbeitungsfehler und technische Fehler

Unter dem Term Verarbeitungsfehler werden viele verschiedene Fehler zusammengefasst, wie Editierfehler, Eingabefehler, Tabellierungsfehler und Kodierungsfehler. Die Gewichtung, die ungleiche Auswahlwahrscheinlichkeiten, Coverage- und Nonresponse-Fehler ausgleichen soll, kann falsch berechnet werden, oder es können technische Fehler wie Programmierungsfehler der Software vorliegen. Andere technische Fehler werden aufgrund der zunehmenden Bedeutung von technologischen Datenerhebungsmethoden in naher Zukunft bestimmt an Bedeutung gewinnen. Vorstellbar ist beispielsweise eine neue Art von Item-Nonresponse, wenn das Befragungsgerät (z.B. Smartphone) abstürzt und der Proband keine Motivation hat, die Befragung erneut von vorne zu beginnen.

4.3.3 Spezifikationsfehler

Spezifikationsfehler betreffen „den Grad der Übereinstimmung zwischen dem, was gemessen werden soll und dem, was tatsächlich gemessen wird“ und entstehen, wenn das Konzept, das durch die Fragebogen gemessen wird, sich von dem Konzept, das in der Umfrage gemessen werden soll, unterscheidet. In dem Fall wird ein falsches Konstrukt gemessen und somit die falschen Parameter geschätzt. Umso größer die Abweichung zwischen dem gemessenem und dem interessierendem Konzept, desto größer der Spezifikationsfehler. Dadurch wird die inhaltliche Interpretation der Daten immer schwieriger. Theoretisch ist eine Umfrage vorstellbar, in der alle Messungen und somit auch die Parameterschätzung komplett fehlerfrei abgelaufen sind, diese sich aber auf eine falsch spezifizierte Variable beziehen. Das würde bedeuten, dass die Interpretation aller statistischen Ergebnisse, die mit dieser Variable zusammenhängen anzuzweifeln wären (Faulbaum; (2014); S. 450). Spezifikationsfehler gehen oft auf eine schlechte Kommunikation zwischen Forscher und Fragebogendesigner zurück und lassen sich im Rahmen spezifischer Umfragedesigns, beispielsweise durch die Korrelation eines empirischen Indikators mit dem durch ihn gemessenen Konstrukt, quantifizieren.

4.4 Zusammenfassung

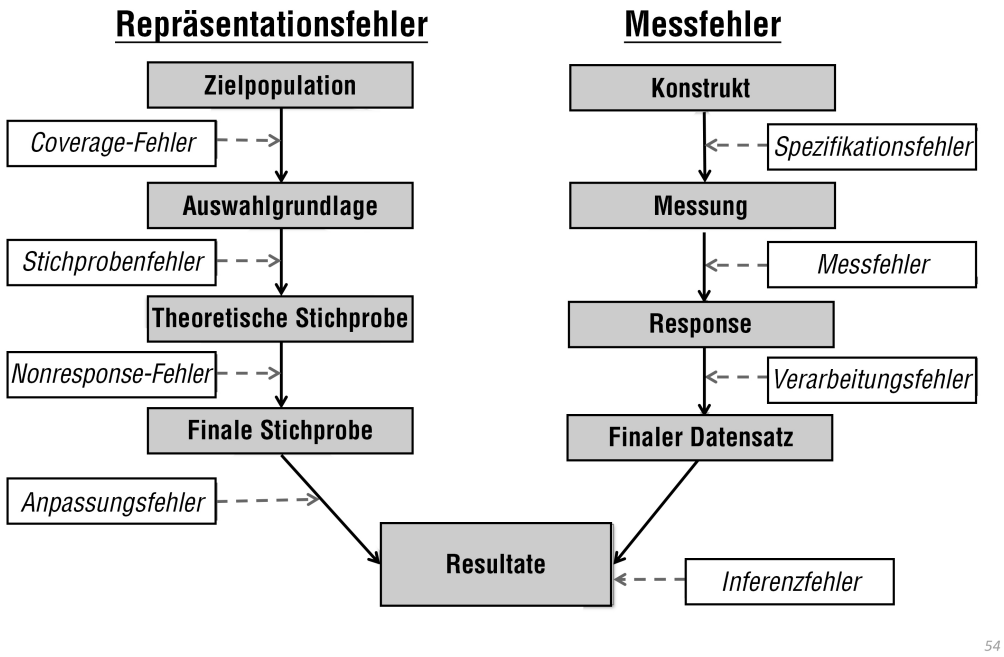

Abbildung 3: Der totale Umfragefehler aus einer Studiendesign-Perspektive

Der TSE lässt sich auch aus anderer Perspektive darstellen. Im Gegensatz zu Abbildung 2 auf Seite 9, können die verschiedenen Fehlerquellen die bei einer Umfrage die Resultate verzerren können, auch aus einer Perspektive des Studiendesigns betrachtet werden. Dabei gibt es zwei Felder, die beim Studiendesign berücksichtigt werden müssen. Auf der einen Seite der praktische, auf der anderen Seite der theoretische Aspekt. Hier werden diese beiden Felder unter den Begriffen Repräsentationsfehler und Messfehler zusammengefasst. Repräsentationsfehler beschreiben die Stichproben- sowie Nichtbeobachtungsfehler und Messfehler betreffen die Beobachtungs- sowie Spezifikationsfehler.

Die erste Fehlerquelle liegt in der Bestimmung der Auswahlgrundlage aus der Zielpopulation, hierbei kann es durch Under- oder Overcoverage zu Coverage-Fehlern kommen. Bei der Bestimmung der zu betrachtenden Stichprobe aus der Auswahlgrundlage, kommt es zu Stichprobenfehlern und die Diskrepanz zwischen der ausgewählten Stichprobe und den tatsächlichen Antworten oder Rückmeldungen wird durch den Nonresponse-Fehler beschrieben. Bei dem Versuch die fehlenden Werte durch Gewichtung oder andere Methoden auszugleichen, kann es noch zu Anpassungsfehlern kommen. Auf

der theoretischen Ebene liegt die erste Schwierigkeit darin, das zu messende Konstrukt richtig zu spezifizieren, die nächste Fehlerquelle liegt in der Auswertung der Messung und dann in der richtigen Verarbeitung der Daten. Schließlich können bei dem Schluss von den vorliegenden Daten auf die interessierende Grundgesamtheit noch inferenzstatistische Fehler auftreten.

5 Messung und Minimierung des Totalen Umfragefehlers

Der TSE ist ein Werkzeug, das dabei helfen kann eine Umfrage gut zu planen. Um die richtigen Entscheidungen in Bezug auf das Umfragedesign zu treffen, müssen viele Qualitäts- und Kostenfaktoren in Betracht gezogen werden und parallel dazu, die Kombination an Umfragedesignfaktoren gewählt werden, die den TSE innerhalb dieser Beschränkungen minimiert. Um diesen Prozess leichter zu gestalten, ist es wichtig Mittel zur Hand zu haben, anhand welcher es möglich ist den TSE zu quantifizieren. So können verschiedene Umfragedesigns miteinander verglichen werden, um das Bestmögliche zu wählen. Auch hilft eine Methode den TSE zu quantifizieren dabei, die oft beschränkten Umfrageressourcen so aufzuteilen, dass der Umfragefehler möglichst minimal ist. Auch wenn es einige verschiedene Möglichkeiten gibt den TSE zu messen, ist die üblichste Methode der mittlere quadratische Fehler (MSE, mean squared error). Jeder Schätzer, der aus den Umfragedaten berechnet wird, hat einen zugehörigen MSE, welcher die Auswirkungen aller Fehler auf diesen Schätzer zusammenfasst. Während es theoretisch möglich ist den TSE über den MSE zu berechnen, kommt es in der Praxis jedoch oft zu Schwierigkeiten, weil dazu normalerweise ein komplett fehlerfreier Schätzer benötigt wird (Biemer, (2010); S. 826). Trotzdem ist das Konzept nützlich um die Natur und die Zusammensetzung von Fehlern zu begreifen. Statistisch gesehen ist der MSE die erwartete, quadrierte Differenz zwischen dem Schätzer und dem zu schätzendem Parameter.

$$\mathbf{MSE}(\hat{\theta}) = \mathbf{E}(\hat{\theta} - \theta)^2 \tag{6}$$

wobei:

$\hat{\theta}$: Schätzer
θ : Parameter

Oder anders ausgedrückt, der quadrierte Bias des Schätzers addiert mit seiner Varianz.

$$\mathbf{MSE}(\hat{\theta}) = \mathbf{Bias}^2(\hat{\theta}) + \mathbf{Var}(\hat{\theta}) \tag{7}$$

Der MSE repräsentiert also den zusammengefassten Effekt aller Stichproben- und Nichtstichprobenfehler. Jede Fehlerquelle kann somit zu dem zufälligen Fehler, dem systematischen Fehler oder zu beiden beitragen. Zufällige Fehler spiegeln sich in der Varianz des Schätzers wieder, systematische in dem Bias.

Zur Minimierung des TSE ist es zunächst nötig die Fehlerquelle so genau wie möglich zu lokalisieren, um dann zu versuchen diese zu beheben. Eine möglich Aufschlüsselung des TSE findet sich unter Punkt 4 in dieser Arbeit, im Folgenden sollen für einige dieser Fehler mögliche Schritte zur Reduzierung des Fehlers aufgezeigt werden. Ausgehend von Abbildung 3 werden die Fehler in der Reihenfolge ihres Auftretens betrachtet. Um Coverage- sowie Stichprobenfehler zu minimieren, ist eine gute Auswahlgrundlage notwendig, das bedeutet Skepsis bei beispielsweise Telefon- und Internetbefragungen, aber auch bei Daten aus Melderegistern ist Vorsicht geboten. Um Nonresponse-Fehler zu minimieren, ist es wichtig gut vorbereitet und trainierte Interviewer zu haben, sowie einen nicht zu langen und mit Filterfragen ausgestatteten Fragebogen zu benutzen. Außerdem sollten die Probanden einen Anreiz oder Sinn in ihrer Teilnahme sehen. Damit Messfehler möglichst gering bleiben, ist es notwendig einen gut getesteten Fragebogen zu besitzen, der das misst, was er zu messen vorgibt.

6 Zusammenfassung

Nach einer ausführlichen Beschreibung des TSE und kurzen Gedanken über seine Reduzierung, sollen jetzt zusammenfasend die Schwächen und Stärken des Konzepts betrachtet werden. Ein Problem ist, dass wichtige Qualitätsmerkmale innerhalb des TSE vernachlässigt werden. So gibt es verschiedene Ebenen von Informationsqualität, die U.S. Key National Indicator Initiative zum Beispiel nutzt bei der Beschreibung der Qualität von Indikatoren, ein weites Feld an Kriterien. Die Qualität von Indikatoren fußt ihrer Meinung nach, auf vier Grundlegenden Kriterien: der Relevanz, der Kredibilität, der Qualität des Schätzers, sowie der Qualität der Daten. Das Konzept des TSE bezieht sich aber hauptsächlich auf die Qualität des Schätzers, während es andere Diensionen von Qualität vernachlässigt. Die Relevanz einer Umfrage oder Daten bezieht sich darauf, inwiefern sie den Anforderungen der Endnutzer gerecht werden, ob das betrachtet wird, was den Nutzer der Daten interessiert. Die Kredibilität oder Glaubwürdigkeit von Daten, wird von der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD, 2011) als das Vertrauen das Datennutzer in Daten haben, nur aufgrund deren Bild der Datenproduzenten, definiert. Diese Dimensionen von Qualität fehlen, neben verwandten Dimensionen wie der Zugänglichkeit, der Transparenz und der Pünktlichkeit der Daten, dem TSE-Konzept (Groves; (2010); S. 863-864). Es legt keinen großen Wert auf die schlussendliche Nutzbarkeit von Daten. Ein weiteres Problem ist der hohe Aufwand für die Messung von bestimmten Fehlern, bzw. das Problem, dass viele Fehler nicht mathematisch messbar sind, außerdem sind häufig notwendige Informationen zur Messung von bestimmten Komponenten nicht vorhanden, oder nur unter hohem Kostenaufwand zu messen. Als tatsächliche Messgröße, ausgedrückt zum Beispiel über den MSE, ist der TSE also nicht geeignet. Außerdem kann gesagt werden, dass obwohl das Konzept schon einige Jahre alt ist, in den Grundformen schon seit 1944, es nicht zu dem Hauptwerkzeug von Umfragedesignern geworden ist und die meisten Umfragedesigner sich nach wie vor auf die Betrachtung von Stichprobenfehlern beschränken.

Eins der größten Verdienste des Konzepts des TSE ist sicherlich, dass es den Fokus auf andere Fehlerquellen außer die Stichprobenfehler gelenkt hat. Das war in der Tat auch eines der Ziele welches Deming (1944) mit seiner Arbeit hatte (Groves; (2010); S. 868). Die möglichst genaue Aufschlüsselung der Fehlerquellen einer Umfrage hat auf jeden Fall zu einer verbesserten Qualität von Umfragen geführt. So hat die isolierte Betrachtung von Messfehlern dabei geholfen bessere Fragebögen zu konzipieren, die Aufteilung von Nonresponse-Fehler in Unterkategorien, hat dazu beigetragen Nicht-Erreichbarkeit und Teilnahmeverweigerung separat anzugehen und die

alleinige Betrachtung von Coverage-Fehlern hat dabei geholfen, auf zu kritisierende Umfragemethoden wie die Telefonumfrage aufmerksam zu machen. Diese genaue Aufschlüsselung der Umfragefehler, welches ein Kernelement des TSE ist, hat dazu beigetragen, die Fehlerquellen besser zu verstehen und sie dadurch voneinander getrennt angehen zu können. Ein weiterer Verdienst des Konzepts ist die Unterscheidung zwischen Beobachtungs- und Nicht-Beobachtungsfehlern.

Obwohl das Konzept des TSE nicht so verbreitet ist, wie es es verdient hätte und trotz der praktischen Unzulänglichkeiten darin, den TSE konkret zu messen, stellt das Konzept allen Umfragedesignern ein nützliches Rahmenwerk zur Verfügung um eine Umfrage zu planen und hilft dabei, Umfrageressourcen besser einteilen zu können. Es bleibt abzuwarten, ob sich der Trend fortsetzt, nicht nur Stichprobenfehler zu betrachten (Biemer; (2010); S. 845) und weitere Forschung dazu beiträgt verschiedene vernachlässigte Fehlerquellen genauer zu untersuchen und sie so mehr in das Bewusstsein von Umfragedesignern zu rücken.

7 Quellen

Biemer, Paul P.; (2010): „Total Survey Error: Design, Implementation, and Evaluation“; In: Public Opinion Quarterly, Vol. 74, No. 5; S.817-848

Deming, Edwards; (1944): „On Errors in Surveys“; In: American Sociological Review, Vol. 9, Issue 4; S. 359-69

Engel, Uwe; Schmidt, Björn Oliver; (2014): „Unit- und Item-Nonresponse“; In: Handbuch Methoden der empirischen Sozialforschung; S. 331-348

Faulbaum, Frank; (2014): „Total Survey Error“; In: Handbuch Methoden der empirischen Sozialforschung; S. 439 – 453

Groves, Robert M., Lyberg, Lars; (2010): „Total Survey Error: Pasr, Present, and Future“; In: Public Opinion Quarterly, Vol. 74, No. 5; S.849-879

Häder, Michael; Häder, Sabine; (2014): „Stichprobenziehung in der quantitativen Sozialforschung“; In: Handbuch Methoden der empirischen Sozialforschung; S. 283-297

OECD; (2011): „Quality Framework and Guidelines for OECD Statistical Activities“; Von: http://www.oecd.org/officialdocuments/publicdisplay documentpdf/?cote=std/qfs(2011)1doclanguage=en, aufgerufen am 08.06.2016

Peytchev, Andy; Carley-Baxter, Lisa R.; Black, Michele C.; (2010): „Coverage Bias in Variances, Associations, and Total Error From Exclusion of the Cell Phone-Only Population in the United States“; In: Social Science Computer Review, Vol. 28, No. 3; S. 287-302

Scheuch, Erwin K.; (1973): „Das Interview in der Sozialforschung“; In: König (Hg.); S. 66-190

Zensus; Von: https://www.zensus2011.de/SharedDocs/Aktuelles/Welche_ Kosten_verursacht_der_Zensus.html, aufgerufen am 03.06.2016