CIVEY

Whitepaper

Der statistische Fehler als Qualitätsindikator bei Civey

Bayesianische Kredibilitätsintervalle in der Berechnung des statistischen Fehlers bei Onlineumfragen

Prof. Dr. Oliver Serfling, Jacob Kastl M.Sc., Denise Sengül M.Sc.

Der klassische Stichprobenfehler

Der statistische Fehler (“margin of error”) ist eine der gängigsten Kennzahlen in der quantitativen Meinungsforschung. Der klassische statistische Fehler gibt dabei das Maß an Variabilität einer reinen Zufallsstichprobe an, also wie stark die Stichprobe und daraus resultierende Ergebnisse bei wiederholter zufälliger Ziehung schwanken würden. In seiner Reinform hängt der statistische Fehler nur von der Größe der Stichprobe ab und wird als wichtigstes Robustheitsmaß genutzt. Diese Berechnung verlangt, dass alle Individuen der Grundgesamtheit mit positiver und identischer Wahrscheinlichkeit ausgewählt werden.

Der statistische Fehler errechnet sich als die Hälfte der Breite des Konfidenzintervalls einer Schätzung und lässt sich somit aus Standardabweichung und Stichprobengröße herleiten. Als Konfidenzniveau wird in der Regel 95% unterstellt. Genau wie die Schätzung an sich variiert auch das Konfidenzintervall bei wiederholter Stichprobenziehung. In der klassischen Statisik, auch frequentistische Statistik genannt, lautet die Interpretation dieses Konfidenzintervalls daher wie folgt:
“Bei einer sehr hohen Anzahl von Wiederholungen der Stichprobenziehung ist zu erwarten, dass 95% der errechneten Konfidenzintervalle den wahren Wert aus der Grundgesamtheit beinhalten.[1]”

Die Grenzen des 95%-Konfidenzintervalls werden unter den beschriebenen Bedingungen als 1,96 Standardabweichungen in beide Richtungen um die Punktschätzung berechnet. Der Wert 1,96 ergibt sich aus dem entsprechenden Quantil der Standardnormalverteilung, die nach dem Gesetz der Großen Zahlen die ungefähre Verteilung aller Schätzungen darstellt. Voraussetzung ist eine große Stichprobe.
Der statistische Fehler lässt sich dann durch folgende Formel approximieren:

$$MoE \simeq \frac{0{,}98}{\sqrt{n}} .$$

Wie bereits erwähnt hängt der klassische statistische Fehler lediglich von der Stichprobengröße n ab.

Die Realität der Stichprobenziehung

In der Praxis der Stichprobenziehung gibt es in der Regel keine perfekte Zufallsstichprobe. Konkret ist oft nicht für alle Personen der Zielpopulation die Auswahlwahrscheinlichkeit bekannt. In einer solchen Situation spricht man vom non-probability sampling. Auch die Erreichbarkeit aller Personen der Grundgesamtheit kann meist nicht sichergestellt werden. Damit ist die Grundannahme für die Berechnung des klassischen statistischen Fehlers verletzt und streng genommen sind der frequentistischen Theorie nach gar keine Rückschlüsse auf die Bevölkerung möglich.[2] In Online-Panels wie bei Civey wird in der Regel in großem Maße auf non-probability Sampling zurückgegriffen. Die Rekrutierung des Panels kann zum einen nicht alle Personen in der Wohnbevölkerung erreichen (siehe Coverage-Bias). Zum anderen gibt es keine “Liste aller Internetnutzer” aus der man eine klassische Zufallsstichprobe ziehen könnte. Teilnehmer wählen sich vielmehr selbst für die Stichprobe aus.

1 Das heißt also spezifisch nicht, dass der wahre Wert “mit 95 prozentiger Sicherheit innerhalb des Intervalls” liegt.

2 Dies gilt für den Industriestandard der Telefonumfragen mit seinen niedrigen Antwortraten und abnehmender Erreichbarkeit über das Festnetz genauso wie für Online-Umfragen.

Folglich wird die Annahme der einfachen Zufallsstichprobe gleich mehrfach verletzt. Im Rahmen der klassischen Schule können hier also keine Konfidenzintervalle und Standardfehler berechnet werden und folglich auch kein frequentistischer statistischer Fehler.

Ein neues Robustheitsmaß für Non-Probability Stichproben

Neben der klassischen, frequentistischen Statistik gibt es allerdings noch einen anderen Zweig, die sogenannte Bayesianische oder Bayessche Statistik, die sich im letzten Jahrzehnt immer größerer Beliebtheit in verschiedensten Anwendungsfeldern erfreut. In der Bayesianischen Statistik beruht eine Schätzung nicht fundamental auf der Wiederholbarkeit des zugrunde liegenden Zufallsexperimentes.
Vielmehr wird eine Schätzung auf Basis aller zum gegebenen Zeitpunkt vorhandenen Informationen durchgeführt. Dazu können, zusätzlich zur erhobenen Stichprobe, zum Beispiel Vorwissen über die Eigenschaften der Grundgesamtheit oder Erfahrungswerte aus anderen Datenquellen zählen.[3]
Mit Hilfe von Modellen, die beispielsweise Abweichungen der Merkmale zwischen Stichprobe und Grundgesamtheit, Design-Effekte sowie Verzerrungen durch Erhebungsmodus und Messfehler nach der Stichprobenerhebung korrigieren, liefert die Bayesianische Statistik eine theoretische Fundierung, die Rückschlüsse auf die Grundgesamtheit erlaubt.[4] In der Survey Statistik bezeichnet man den klassischen, frequentistischen Ansatz oft als designbasierten Ansatz und den Bayesianischen als modellbasierten Ansatz.
Die Robustheit der Schätzungen lässt sich im Bayesschen Kontext über “Kredibilitätsintervall” messen. Diese treffen eine Aussage darüber, wie sicher die Schätzung den wahren Parameter beinhaltet, basierend auf allen Informationen, die zum Zeitpunkt der Schätzung vorliegen. Insofern lassen Kredibilitätsintervalle eine intuitivere und realitätsnähere Deutung der Unsicherheit zu. Ein 95%-Kredibilitätsintervall für eine Schätzung kann dann folgendermaßen verstanden werden: “Unter Berücksichtigung aller derzeit vorliegender Informationen und der erhobenen Stichprobe, liegt der wahre Wert aus der Grundgesamtheit mit einer Wahrscheinlichkeit von 95% innerhalb des Kredibilitätsintervalls.”

Bei Civey interessiert uns vorrangig der Anteil der Zustimmung zu einer Aussage in der Gesamtbevölkerung. Diese Größe stellt den wahren Wert dar, also einen unbekannten Parameter, den es zu schätzen gilt. In einem solchen Fall könnte Vorwissen über diesen Parameter, etwa hinsichtlich der Eigenschaften der Grundgesamtheit oder aus Erfahrungswerten aus anderen Datenquellen, durch eine Beta-Verteilung mit entsprechenden Hyperparametern abgebildet werden. Die gemessenen Werte[5] folgen einer Binomialverteilung. Der Parameter dieser Verteilung ist der wahre Anteil der Zustimmung, den wir schätzen möchten. Aus der Kombination dieser beiden Verteilungen ergibt sich die Beta-Binomialverteilung. Sie stellt die Verteilung des zu schätzenden Parameters auf Basis des Vorwissens dar, aktualisiert mit den vorliegenden Daten aus einer Stichprobenziehung.

3 Das Vorwissen schlägt sich in der prior-Verteilung nieder.

4 Die Voraussetzung hierfür ist, dass bedingte Ignorierbarkeit vorliegt. In non-probability Online-Stichproben muss dafür eine adäquate Modellierung von Verzerrungen, Design-Effekten, Messfehlern etc. vorliegen. So wird sichergestellt, dass der zu messende Wert keinen weiteren Einfluss auf die Teilnahmebereitschaft hat (IPSOS 2016, Little & Rubin 2002).

5 Eine binäre Variable die entweder den Wert 1 (stimme der Aussage zu) oder 0 (stimme der Aussage nicht zu) annimmt

Ähnlich wie im frequentistischen Fall kann über die Quantile dieser Verteilung ein Intervall und daraus wiederum ein Qualitätsmaß hergeleitet werden. Die Bayessche Entsprechnung des Konfidenzintervalls nennt sich hierbei Kredibilitätsintervall. Dieses verwenden wir in der Berechnung eines Bayesschen statistischen Fehlers auch bei Civey.

Bei Civey gehen wir zunächst davon aus, dass wir keinerlei Vorwissen über die Verteilung der zu schätzenden Parameter haben. Alle Ausprägungen zwischen 0% und 100% gelten dann als gleich wahrscheinlich.[6] Für ein 95% Kredibilitätsintervall folgt unter Anwendung der Formel (32) aus der Beschreibung des Beta-binomial Modells von Navarro & Perfors ein Bayesianischer statistischer Fehler von:

$$Bayes - MoE = [Beta(0{,}975 \mid \beta 1 + k,\ \beta 2 + n - k) - Beta(0{,}025 \mid \beta 1 + k,\ \beta 2 + n - k)] / 2$$
$$Bayes - MoE = [Beta(0{,}975 \mid 1 + \tfrac{n}{2},\ 1 + \tfrac{n}{2}) - Beta(0{,}025 \mid 1 + \tfrac{n}{2},\ 1 + \tfrac{n}{2})] / 2,$$

wobei $\beta 1 = \beta 2 = 1$, $k = \frac{n}{2}$ und damit $n - k = \frac{n}{2}$ gelten. Zudem ist $Beta(q \mid \beta 1 + k,\ \beta 2 + n - k)$ das q-Quantil der Beta-Binomialverteilung.

Diese Formel stellt die Hälfte der maximalen Breite des 95% Kredibilitätsintervalls dar, was die konzeptionelle Nähe zum klassischen statistischen Fehler hervorhebt.[7]
Dieser berücksichtigt allerdings nicht die zusätzliche Unsicherheit, die sich ergibt, wenn Stichproben nach der Erhebung gewichtet werden. Nachträgliche Gewichtungen der Stichprobe können die adäquate Schätzung relevanter Parameter der Grundgesamtheit in der Stichprobe sicherstellen. Je stärker die nachträgliche Gewichtung ausfällt, desto höher die zusätzliche Variabilität und desto geringer die Effizienz der Stichprobe (vgl. Kish 1992). Für eine gleichwertige Präzision der Schätzung wird dann also eine größere Stichprobe benötigt. In Online Panels ist aufgrund der genannten Probleme mit dem Erhebungsmodus eine solche nachträgliche Gewichtung der Stichprobe hinsichtlich demographischer, politischer und Verhaltensvariablen unabdingbar.
Um die Erhöhung des Bayesschen statistischen Fehlers adäquat abzubilden, wird er, ähnlich wie bei anderen Online-Panels, mit dem Faktor eins plus dem quadrierten Variationskoeffizienten der Gewichte berechnet. Dieser Faktor wird oft als Design Effekt bezeichnet.
Damit resultiert für die Berechnung des bayesianischen statistischen Fehlers, unter Berücksichtigung der Quadratwurzel des Design-Effekts (vgl. Pew Research Center 2010, YouGov 2015), folgende Formel:

$$Bayes - MoE = ([Beta(0{,}975 \mid 1 + \tfrac{n}{2},\ 1 + \tfrac{n}{2}) - Beta(0{,}025 \mid 1 + \tfrac{n}{2},\ 1 + \tfrac{n}{2})] / 2) * \sqrt{1 + CV^2}\ .$$

6 Die beschriebene Situation stellt eine “uninformative prior”-Verteilung dar.

7 Die rechnerische Nähe von bayesianischer Inferenz (bei “uninformative prior”-Verteilungen) und klassischer Inferenz wird z. B. in Little & Rubin (2002) dargelegt. In der Regel können die resultierenden bayesianischen Credible Intervals durch die klassischen Konfidenzintervalle approximiert werden.

Diskussion

Die Entscheidung, den statistischen Fehler mit Bayesianischen Methoden zu berechnen, akzeptiert ganz explizit ein subjektives Element in der Schätzung der Umfrageergebnisse. Dieser Einfluss schlägt sich in der Modellierung von demographischen Diskrepanzen und des Einflusses von politischen und Verhaltensvariablen nieder. Ähnliche Methoden werden allerdings in den allermeisten Umfragen angewandt,[8] sodass wir es für konsequent halten, offen mit diesem Umstand umzugehen und hoffen, damit einen Beitrag zur Debatte über die Zukunft der Meinungsforschung zu leisten. In bestem Bayesianischem Sinne entwickeln wir unserer Modelle natürlich dauerhaft weiter und beziehen neue Informationen ein, um die bestmögliche Genauigkeit in unseren Umfragen zu erreichen.

IPSOS (2016) “Towards the Use of Bayesian Credibility Intervals in Online Survey Results”:
http://www.ipsos-na.com/dl/pdf/knowledge-ideas/public-affairs/IpsosPA_POV_BayesianCredibilityIntervals.pdf

Kish (1992) “Weighting for Unequal Pi”:
http://www.jos.nu/articles/abstract.asp?article=82183

Little & Rubin (2002):
“Statistical Analysis with Missing Data”

Pew Research Center (2010) “The internet gives citizens new paths to government services and information”:
http://www.pewinternet.org/files/old-media//Files/Reports/2010/PIP_Government_Online_2010_with_topline.pdf

YouGov (2015) “The Methodology of the 2016 YouGov/CBS News Battleground Tracker”:
https://today.yougov.com/news/2015/09/13/methodology-2016-cbs-news-battleground-tracker/

Navarro & Perfors “An introduction to the Beta-Binomial model”:
https://www.cs.cmu.edu/~10701/lecture/technote2_betabinomial.pdf

8 PEW Research berechnet den Design-Effekt wie oben beschrieben, siehe Pew Research Center (2010). YouGov verwendet zur Ermittlung des statistischen Fehlers ebenfalls einen Modell-basierten Ansatz, vgl.: YouGov (2015).