1.5 Datenmanagement und Gewichtung


Stephan Gerhard Huber


Isabella Lussi


Florian Keller

Im ch-x/YASS werden zwei Stichproben von Jugendlichen befragt. Zum einen eine quasi Vollerhebung der 19-jährigen Schweizer Männer. Zum andern eine Ergänzungsstichprobe von 19-jährigen Frauen. Bei beiden Stichproben wird der gleiche Paper-&-Pencil-Fragebogen eingesetzt. Die Durchführung der Befragung unterscheidet sich jedoch je nach Stichprobe. Die jungen Schweizer Männer bearbeiten den Fragebogen in einer Klassenzimmerbefragung im Rahmen des ordentlichen Rekrutierungsverfahrens der Schweizer Armee. Die jungen Frauen füllen den Fragebogen zu Hause aus.

Datenmanagement

Nach der Befragung werden die Fragebogen an das Bundesamt für Informatik und Telekommunikation (BIT) gesandt. Das BIT scannt die Fragebogen ein und erzeugt eine ANSI-CSV Datei mit den Rohdaten. Die ANSI-CSV Datei ist im Grunde eine einfache Textdatei, die nur aus 0 und 1 besteht, wobei eine Zeile einen Fall darstellt und jedes Zeichen durch ein Komma getrennt ist. Diese Datei mit den Rohdaten der Befragung wird zur Datenaufbereitung an die Leitung des Forschungskonsortiums ch-x/YASS geschickt.

Die Datenaufbereitung erfolgt als ein dreistufiger Prozess:

1. Einlesen der Daten
2. Bereinigen der Daten
3. Plausibilisieren der Daten

Einlesen der Daten

In einem ersten Schritt der Datenaufbereitung werden die Rohdaten des BIT zu einem Datensatz aufbereitet, der von einem gängigen Statistikprogramm wie SPSS, STATA oder R interpretiert werden kann. Dazu werden alle ANSI-CSV Dateien in einer txt-Datei zusammengestellt. Diese Datei enthält die Urdaten. Anschliessend werden die Urdaten in eine Statistiksoftware eingelesen (hier SPSS) und zu Ergebnisvariablen transformiert. Das heisst, die ursprünglich binären Variablen werden zu Variablen mit kategorialen Werten rekodiert. So wird beispielsweise aus den Variablen V1- V6 mit den Werten 0, 0, 0, 0, 1, 0 die Variable F001 mit dem Wert 5. Danach werden die Variablen einheitlich formatiert und nach einer vorgegebenen Nomenklatur «gelabelt».

Bereinigen der Daten

In einem zweiten Schritt der Datenaufbereitung werden technisch bedingte fehlende oder falsche Antworten, beispielsweise Einlesefehler beim Scannen der Fragebögen, identifiziert. Zudem wird die Datei daraufhin überprüft, ob gewisse Fragebögen doppelt oder mehrfach eingelesen worden sind.

Plausibilisieren der Daten

Neben dieser Bereinigung rein technischer Fehler werden die Daten auch auf unmögliche und widersprüchliche Angaben kontrolliert. So werden beispielsweise Fälle aufgedeckt, die ein Alter von über 100 Jahren angeben oder die bei der Frage nach dem Geschlecht sowohl «Mann» als auch «Frau» ankreuzen (univariate Plausibilisierung).

Bei der Datenplausibilisierung werden nicht nur unmögliche, sondern auch unglaubwürdige Werte identifiziert. Dieser Schritt ist deshalb so wichtig, weil die Durchführung der Befragungen einen hohen Anteil an unglaubwürdigen Daten erwarten lässt. Dies gilt insbesondere für die Befragung der jungen Männer. Die jungen Männer füllen den Fragebogen in den Rekrutierungszentren der Armee aus. Die Bearbeitung des Fragebogens zur verweigern, wäre zwar juristisch möglich. Angesichts der sozialen Kontrolle während der Klassenzimmerbefragung ist die Rücklaufquote nahezu 100 Prozent. Diese potenziell obstruktiven Fälle gilt es vor den Datenanalysen im Prozess der Datenaufbereitung

Datenmanagement und Gewichtung

Im ch-x/YASS werden zwei Stichproben von Jugendlichen befragt: zum einen eine quasi Vollerhebung der 19-jährigen Schweizer Männer, zum andern eine Ergänzungsstichprobe von 19-jährigen Frauen. Bei beiden Stichproben wird der gleiche Paper-&-Pencil-Fragebogen eingesetzt. Die Durchführung der Befragungen unterscheidet sich jedoch: Die jungen Schweizer Männer bearbeiten den Fragebogen in einer Klassenzimmerbefragung im Rahmen des ordentlichen Rekrutierungsverfahrens der Schweizer Armee. Die jungen Frauen füllen den Fragebogen zu Hause aus.

Huber, S.G., Lussi, I. & Keller, F. (2016). Datenmanagement und Gewichtung. In S.G. Huber (Hrsg.), Young Adult Survey Switzerland (S. 30–33). Bern: BBL / OFCL / UFCL.

herauszufiltern. Werden die Daten nicht auf das Genaueste plausibilisiert, so bleiben nicht nur die Datenwerte, sondern letztlich auch die Ergebnisse der folgenden wissenschaftlichen Analysen unglaubwürdig.

Die Datenplausibilisierung erfolgt univariat, bivariat sowie multivariat.

Bei der univariaten Datenplausibilisierung wird zunächst theoretisch ein plausibler Wertebereich für jede einzelne Variable definiert. Danach wird jede Variable überprüft, und Werte ausserhalb des definierten Wertebereichs werden identifiziert. So können beispielsweise Grössenangaben von über 2,5 Metern oder Altersangaben von über 30 Jahren erkannt und als «unplausibel» kodiert werden.

Bei der bivariaten Datenplausibilisierung werden Angaben über mehrere Variablen hinweg auf Konsistenz und «Sinnhaftigkeit» (logisch widerspruchsfrei) überprüft. So ist es beispielsweise höchst unplausibel, dass eine Person im Alter von 19 Jahren bereits ein Universitätsstudium erfolgreich abgeschlossen hat.

Mit der multivariaten Datenplausibilisierung können schliesslich Fälle von durchgehender Antwortverwei-

Gestion et pondération des données

Les enquêtes de ch-x/YASS s'adressent à deux échantillonnages de jeunes : d'une part, à la quasi-totalité des jeunes Suisses de 19 ans, d'autre part, à un échantillonnage complémentaire des jeunes femmes de 19 ans. Les deux groupes répondent au même questionnaire écrit. L'organisation des enquêtes est cependant différente : les jeunes hommes y répondent en classe dans le cadre du processus ordinaire de recrutement de l'armée suisse. Les jeunes femmes remplissent le questionnaire à la maison.

Gestione e ponderazione dei dati

I sondaggi ch-x/YASS si rivolgono a due campioni di giovani: alla quasi totalità dei maschi svizzeri di 19 anni e a un campione complementare di ragazze della stessa età. I due gruppi rispondono alle domande di uno stesso questionario scritto. L'organizzazione delle inchieste è tuttavia diversa: i giovani rispondono al questionario in aula durante i giorni di reclutamento dell'esercito; le giovani riempiono il questionario a casa propria.

gerung, durchgehender Obstruktion, ausgeschlossen werden. Zudem werden mit Datenscreening-Techniken inhaltsunabhängige Antwortmuster (z.B. das konsequente Ankreuzen der höchsten oder der tiefsten Kategorie) aufgedeckt.
Im Anschluss an die Plausibilisierung aller Daten werden die fehlenden bzw. unplausiblen Antworten summiert. Übersteigt die Anzahl fehlender Antworten eine bestimmte Grenze, wird der Fall ganz aus den weiteren Analysen ausgeschlossen, da davon ausgegangen werden muss, dass der gesamte Fragebogen nicht wahrheitsgetreu ausgefüllt wurde.

Gewichtung der Daten und Analyseperspektiven

Um korrekte Aussagen über die jungen Erwachsenen machen zu können, müssen die Daten gewichtet werden. Dadurch werden die unterschiedlichen Auswahlwahrscheinlichkeiten einerseits (Haltiner, 2011) und die unterschiedliche Stichprobengrösse der Frauen- und der Männerstichprobe andererseits ausgeglichen. Bei der Gewichtung wurde ein Verfahren eingesetzt, das bereits bei der ch-x-Studie 2006/07 angewendet wurde (Keller & Moser, 2013).

Dabei wird, einfach gesagt, an alle Frauen das Gewicht 1 vergeben, die Männer aber so stark untergewichtet, dass das Geschlechterverhältnis in der Stichprobe jenem in der Population entspricht. Mit diesem Vorgehen vermeidet man bei den Frauen zu hohe Gewichtungsfaktoren, behält aber trotzdem alle statistischen Informationen der Männer. Auch werden so die Standardfehler eher überschätzt und statistische Signifikanzen zwischen Gruppen eher konservativ beurteilt. Die Vorteile des grossen Stichprobenumfangs und der hohen Desaggregationstiefe bleiben aber grundsätzlich erhalten.

Da sich die Teilnahmebereitschaft bei den Frauen nach Bildungsgrad unterscheidet, wurde zudem noch nach der nachobligatorischen Ausbildung (Allgemeinbildung, Berufsbildung, keine Ausbildung auf der Sekundarstufe II) gewichtet, so dass die Verteilung der Bildungsabschlüsse in der Stichprobe jener in der Grundgesamtheit entspricht. Als Referenzwerte wurden die offiziellen Statistiken des Bundesamts für Statistik zur Anzahl 19-Jähriger in den Jahren 2010 und 2011 nach Geschlecht und Kanton und die Abschlussquoten der Sekundarstufe II in den Jahren 2010 und 2011 nach Geschlecht verwendet.
Tabelle 1 zeigt die Anzahl plausibler Fälle (n) im Datensatz der ch-x/YASS nach Geschlecht einmal ungewichtet und einmal gewichtet. Insgesamt sind im Datensatz der ch-x/YASS die Angaben von insgesamt

Tabelle 1: Stichprobenumfang ungewichtet und gewichtet nach Geschlecht

	Männer	Frauen	Total
n ungewichtet	26'444	1'420	27'864
Anteil ungewichtet	94.9%	5.1%	100%
n gewichtet	1'457	1'420	2'877
Anteil gewichtet	50.6%	49.4%	100%

27'864 jungen Erwachsenen. Davon sind 26'444 bzw. 94.9 Prozent Männer und 1'420 bzw. 5.1 Prozent Frauen. Durch die Gewichtung wird der Männeranteil auf 50.6 Prozent reduziert. Dies entspricht gemäss BFS dem Männeranteil in der Population.

Mit den gewichteten Daten können nun Aussagen gemacht werden über die jungen Erwachsenen mit Schweizer Nationalität und mit Wohnsitz in der Schweiz, die in den Jahren 2010 bzw. 2011 19 Jahre alt waren. Bei der Beschreibung von sehr kleinen Gruppen hingegen (z.B. von jungen Erwachsenen, die eine Sonderklasse besucht haben) müssen die Analysen auf die Daten der Männerstichprobe reduziert werden. Die Stichprobe der jungen Frauen ist für solche Analysen zu klein, um zuverlässige Aussagen machen zu können. Die Aussagen beschränken sich dann auf die jungen Schweizer Männer (vgl. z.B. Kapitel 3.2).

Literatur:

Haltiner, K. (2011). Wie und wen befragen die Eidgenössischen Jugendbefragungen? Organisation und Erhebungsverfahren der ch-x. In Eidgenössische Jugendbefragungen ch-x (Hrsg.), Laufende Jugendstudien. Werkstattbericht 2010/2011 (S. 28–31). Bern.

Keller, F. & Moser, U. (2013). Schullaufbahnen und Bildungserfolg. Auswirkungen von Schullaufbahn und Schulsystem auf den Übertritt ins Berufsleben (Wissenschaftliche Reihe der ch-x, Bd. 22). Zürich/Chur: Rüegger.