Datenerfassung und Datenaufbereitung

Rolf Porst – unter Mitarbeit von Ruth Holthof[1]

Nach „Feldende“, wenn also alle – oder zumindest die meisten – der geplanten Interviews realisiert worden sind und Ihre SchülerInnen den Job als InterviewerInnen – hoffentlich, aber auch sehr wahrscheinlich – wohlbehalten absolviert haben, liegt Ihnen jetzt ein mehr oder weniger großer Stapel mehr oder minder gut und sorgfältig ausgefüllter Fragebogen vor. Davon ausgehend, dass SchülerInnen im Normalfalle keine computergestützten Befragungen durchführen, müssen wir nun dafür sorgen, dass die Fragebogen „maschinenlesbar“ gemacht, also in eine Datendatei überführt werden. Schließlich wollen wir die Daten, die wir produziert haben, auch angemessen auswerten. Welche Möglichkeiten der Datenerfassung gibt es? Und wie gehen wir beim Erfassen der Daten mit Fehlern im Fragebogen um? Mit Fragen dieser Art werden wir uns im vorliegenden Beitrag der Reihe „Schüler forschen“ beschäftigen. Konkret geht es darum, drei Fragen zu beantworten:

1. Welche *Programme zur Datenerfassung* können wir empfehlen unter dem Gesichtspunkt, dass sie an Schulen üblicherweise verfügbar bzw. leicht und vor allem kostengünstig anzuschaffen sind?
2. Wie gehen wir bei der *Eingabe der Daten*, also beim Übertrag der Antworten aus dem Fragebogen in die Eingabedatei, vor?
3. Wie funktioniert die *Datenbereinigung*, also das Erkennen und das Korrigieren von Fehlern in dem von uns erstellten Datensatz?

Bevor wir uns der Beantwortung dieser Fragen zuwenden, sollten wir zunächst aber noch ein paar Regeln für die Vorbereitung der Datenerfassung definieren.

Rolf Porst
Markt- und Sozialforscher, Römerberg

1. Zur Vorbereitung der Datenerfassung

Zur Vorbereitung der Datenerfassung sollten Sie als Erstes jedem Fragebogen eine eindeutige Nummer zuordnen, die Sie später auch in den Datensatz aufnehmen. Ohne eine solche „Fallnummer" werden Sie große Probleme bei der Datenbereinigung haben. Die Fallnummer, auch „Paginiernummer" genannt, können Sie von Ihren SchülerInnen per Hand in den Fragebogen eintragen lassen, am besten gleich auf der ersten Seite oben, als erste Variable des Fragebogens. Sicherer als die manuelle Nummerierung ist – vor allem bei einer sehr großen Anzahl zu erfassender Fragebogen – die Verwendung eines Paginierstempels, der allerdings nicht ganz billig ist.

Weiterhin sollten Sie Regeln für die Dateneingabe aus dem Fragebogen heraus definieren, um eindeutig festzulegen, wie Ihre SchülerInnen mit bestimmten Problemen umgehen, die bei der Datenerfassung häufiger auftreten können (und werden):

Wie viele Fragen eines Fragebogens müssen überhaupt beantwortet bzw. ausgefüllt worden sein, damit der Fragebogen verwertet werden kann?

Sie können natürlich rein formal einen Prozentsatz festlegen (z.B. „mehr als die Hälfte der Fragen muss beantwortet sein") oder inhaltliche Regeln definieren (z.B. „alle demografischen Fragen müssen beantwortet sein"). Verzichten Sie aber besser auf solche Definitionen. Nehmen Sie alle Informationen aus allen Fragebogen auf, auch wenn manche Fragebogen nur unvollständig oder nur ganz wenig ausgefüllt sind.

Wie gehe ich damit um, wenn Fragen nicht beantwortet worden sind?

Wenn Fragen nicht beantwortet oder Angaben nicht gemacht wurden, verkoden Sie diese als „fehlende Werte" („missing data"); je nach Datenerfassungsprogramm ist ein Code für missing data vorgesehen, oder Sie definieren ihn selbst (z.B. „Fehlende Werte werden bei einspaltigen Variablen mit „9" verkodet, bei zweispaltigen Variablen mit „99", usw.").

Wie gehe ich mit Werten um, die es nicht (oder zumindest höchstwahrscheinlich nicht) geben darf?

Wenn Sie im Fragebogen Informationen finden, die nicht nachvollziehbar sind (z.B. hat jemand beim Alter „121" eingetragen, verkoden Sie diese Information als „missing data", nach der gerade beschriebenen Regel also z.B. mit „999". Wenn Sie Informationen vorfinden, die unwahrscheinlich, aber nicht ausgeschlossen sind (z.B. jemand gibt beim Alter „95" an), prüfen Sie, ob die Information zulässig sein kann oder nicht; wenn Sie z.B. nur Personen im Alter von 18 bis 65 Jahren befragt haben, ist die „95" nicht zulässig. Verkoden Sie solche Informationen dann ebenfalls als missing data; können Sie aufgrund Ihrer Befragungsgruppe nicht ausschließen, dass jemand wirklich 95 Jahre alt ist, nehmen Sie diesen Wert als bare Münze und tragen ihn in den Datensatz ein.

Wie gehe ich mit dem Problem um, dass jemand bei einer Frage zwei Antworten angekreuzt hat, obwohl nur eine Antwort zulässig sein sollte?

Klarer Fall von „missing data" für die gesamte Frage. Sie vergeben also den von Ihnen für „missing data" vorgesehenen Wert, sofern das nicht programmseitig automatisch geschieht.

Wie gehe ich mit Filterfehlern um, wenn jemand eine Frage beantwortet hat, die er aufgrund eines Filters gar nicht hätte beantworten sollen? Oder wenn jemand eine Frage nach einem Filter nicht beantwortet hat, obwohl er sie hätte beantworten müssen?

Im ersteren Fall ignorieren Sie bei der Dateneingabe den fälschlicherweise eingegebenen Wert, geben also nichts ein. Streichen Sie ihn am besten im Fragebogen durch. Im letzteren Fall liegt ein „missing data" vor, den Sie wie oben beschrieben behandeln.

Soweit der Überblick über die üblichen Fehler in ausgefüllten Fragebogen. Wenden wir uns nun Programmen zur Erfassung der Daten zu.

2. Programme zur Erfassung der Daten

Es ist nicht unbedingt erforderlich, dass Sie die Daten in dem Programm erfassen, mit dem Sie die erfassten Daten später auswerten wollen – aber zu empfehlen ist das allemal. Erspart es Ihnen doch die Transformation der Daten aus dem einen in das andere Programm. SozialwissenschaftlerInnen arbeiten meistens mit speziellen, auf ihre spezifischen Auswertungsstrategien ausgerichteten Programmen wie *IBM SPSS Statistics*[2], *SAS*[3] oder *Stata*[4]. Für Ihre Zwecke sind diese Programme aber zu vielfältig, zu mächtig, zu teuer – und nicht zuletzt: überhaupt nicht erforderlich. Sie anzuschaffen würde sich für Ihre Schule allenfalls dann anbieten, wenn dort für die nächsten Jahrzehnte ein Schwerpunkt in der Durchführung von Projekten der empirischen Sozialforschung installiert werden sollte. Was vermutlich eher nicht der Fall sein wird. Konzentrieren wir uns also lieber auf Programme, die Ihnen allen vorliegen oder leicht und ohne große Kosten verfügbar gemacht werden können und vor allem: für Ihre schulischen Projekte vollkommen ausreichend sind. Da wäre zunächst das Programm *Microsoft Excel*.

Microsoft Excel dürfte Ihnen allen bekannt sein als zentraler Baustein von Microsoft Office. Bescheiden ausgedrückt handelt es sich dabei um ein Programm zur Tabellenkalkulation. Wenn Sie in dem Programm auf „Formeln" – „Mehr Funktionen" klicken und dort „Statistisch" auswählen, können Sie auf eine Vielzahl statistischer Auswertungsverfahren zugreifen; Sie finden dort alles, was Sie in Ihrem schulischen Projekt je brauchen können, aber auch hier viel mehr, als Sie dabei je brauchen werden. Warum also nicht nach einem Programm fragen, das sehr viel spezifischer auf Ihre Aufgabenstellung im Sozialkundeunterricht abzielt?

Speziell auf den Einsatz in schulischen Unterrichtsprojekten mit Umfragen ausgerichtet ist das Programm GrafStat[5], das von Uwe Diener 1985 entwickelt und seither – gefördert durch die Bundeszentrale für politische Bildung – gepflegt wird. Aktuell liegt das Programm in der Ausgabe von 2016 vor. Neben seiner Konzentration auf den Einsatz in schulischen Forschungsprojekten hat GrafStat eine Reihe weiterer entscheidender Vorteile für den Einsatz dort. Zunächst einmal gibt es bei der Bundeszentrale für politische Bildung mit dem Konzept „Forschen mit GrafStat" (http://www.bpb.de/lernen/grafstat/) eine an Beispielprojekten orientierte Darstellung der Möglichkeiten des Programms und seines Einsatzes in schulischen Forschungsprojekten. Sie finden dort u.a. Projekte mit Titeln wie „Rechtsextremistische Einstellungen im Alltag", „Mobbing – bei uns nicht" oder den Klassiker „Bundestagswahl 2013".

Weiterhin: Zu jedem Projektbeispiel werden Musterfragebogen und Beispieldatensätze zur Verfügung gestellt, die mit Hilfe von GrafStat sekundäranalytisch[6] ausgewertet werden können. Und schließlich: Für Angehörige öffentlicher Bildungseinrich-

tungen, damit auch für LehrerInnen und SchülerInnen, ist das Programm GrafStat kostenlos per Download zu beziehen: http://www.grafstat.de/service/anmeldung.htm.

Bei GrafStat handelt es sich um ein Programm, das Ihnen alle Bausteine der Durchführung einer Befragung – von der Fragebogenkonstruktion über die Datenerfassung bis zur Datenauswertung – zur Verfügung stellt und Ihnen darüber hinaus konkretes Forschungsmaterial zu unterschiedlichen gesellschaftlichen Themen und Datensätzen zur sekundäranalytischen Bearbeitung solcher Fragestellungen anbietet. Damit geht es substanziell über das Angebot von Excel hinaus. Deshalb habe ich bei allen schulischen Forschungsprojekten, die ich beraten und betreut habe, spätestens im Stadium der Datenerfassung mit GrafStat arbeiten lassen bzw. den Einsatz von GrafStat empfohlen. Was ich hiermit auch für Ihre Projekte tue. Von daher konzentrieren wir uns im Folgenden auf die Dateneingabe und Datenbereinigung unter Verwendung des Programms GrafStat.

3. Dateneingabe

GrafStat bietet zwei Varianten zur Dateneingabe an, entweder in Form eines Bildschirm-Interviews oder als Listeneingabe. Bei der *Listeneingabe* finden Sie alle Fragen, Antwortmöglichkeiten und die Plätze für die Dateneingabe pro Fall auf *einer* Bildschirmseite, und Sie können auf diese Seite alle Antworten aus Ihrem Papierfragebogen übertragen. Bei der Eingabe als *Bildschirminterview* simulieren wir quasi eine Befragung am PC; jede Frage hat eine eigene Seite auf dem Bildschirm, und Sie müssen Seite um Seite weiterblättern, um die Antworten aus Ihrem Fragebogen in den Computerfragebogen einzugeben.

Die Listeneingabe hat den Vorteil, dass Sie bei der Eingabe der Daten nicht durch den Fragebogen „blättern" müssen, sondern dass der einzelne Fall komplett auf einer Bildschirmseite abgebildet wird. Das geht schneller, ist aber erfahrungsgemäß ein klein wenig fehleranfälliger.

Beide Varianten machen es zunächst erforderlich, dass Sie über das Menü „Fragebogen" – „Neu", Ihren Fragebogen in GrafStat erstellen, auch wenn Sie die „echte" Befragung mit Papier-Fragebogen durchgeführt haben. Dazu geben Sie Ihren eigenen Fragebogen eins zu eins in das Programm ein. Sie haben die Möglichkeit, für jede Frage den kompletten Fragetext eintragen zu lassen oder eine „Kurzform", das wären z.B. der Name der jeweiligen Variablen, die Sie mit der Frage operationalisieren wollen, oder auch die ersten drei, vier Worte einer Frage. Wenn Sie GrafStat „nur" für die Dateneingabe verwenden wollen, empfehle ich, mit der Kurzform zu arbeiten, um unnötigen Aufwand durch das Eintragen der Fragentexte zu vermeiden. Wenn Sie die Fragen im Fragebogen – was ich Ihnen unbedingt empfehle – mit Ziffern gekennzeichnet haben, tragen Sie diese Fragennummern mit ein. Spätestens bei der Datenbereinigung und auch bei der Datenanalyse werden Sie feststellen, wie hilfreich die Nummerierung der Fragen im Fragebogen, aber auch im Datensatz ist.

Zur Erstellung des Fragebogens stehen Ihnen in GrafStat fünf mögliche Fragenarten zur Verfügung: geschlossene Frage mit *Einfachauswahl*, geschlossene Frage mit *Mehrfachauswahl*, *Skala*, *Maßzahl* (also Eingabe von Ziffern) und *frei* (also Texteingabe für offene Fragen).[7] Mit diesen vorgegebenen Optionen übertragen Sie Ihren Fragebogen komplett über die Fragebogen-Eingabemaske in das Programm.

Geschlossene Fragen mit Einfachnennung
Bei *geschlossenen Fragen mit Einfachnennung* ordnen Sie den Kategorien des Fragebogens ganze Ziffern zu, am besten immer von 1 bis n (Anzahl der Antwortmöglichkeiten zu einer Frage). Lautet die Frage z.B.

„Wie stark interessieren Sie sich für Politik, sehr stark, stark, mittel, wenig oder überhaupt nicht?" vergeben Sie für „überhaupt nicht" die 1, für „wenig" die 2, für „mittel" die 3, für „stark" die 4 und für „sehr stark" die 5. Achten Sie darauf, dass sich die „Richtung" der Antwortmöglichkeiten in der Ziffernfolge widerspiegelt: Der verbal niedrigste Wert erhält die niedrigste Ziffer, in der Regel die 1, der verbal höchste Wert bekommt die höchste Ziffer.

Geschlossene Fragen mit Mehrfachnennungen
Bei *geschlossenen Fragen mit Mehrfachnennungen* müssen Sie jede Antwortmöglichkeit als eigene Variable in den Datensatz aufnehmen und 0/1-verkoden. Lautet die Frage z.B.

„Welche der folgenden Geräte befinden sich in Ihrem Haushalt? Sie können hier mehrere Kreuze machen" mit den sechs Antwortmöglichkeiten Waschmaschine – Spülmaschine – Küchenmaschine – Espressomaschine – Staubsauger – Bügeleisen, dann müssen Sie sechs „Plätze" im Datensatz reservieren, für jedes Gerät einen eigenen. Jedes Gerät, das genannt wird, wird im Datensatz (üblicherweise) mit einer 1 für „im Haushalt vorhanden" erfasst, jedes Gerät, das nicht genannt wird, mit (üblicherweise) einer 0 für „im Haushalt nicht vorhanden".

Offene Fragen
Wie Sie mit *offenen Fragen* im Fragebogen umgehen, hängt davon ab, wie viel Zeit und Arbeit Sie investieren wollen. Die einfachste und schnellste Variante besteht darin, dass Sie pro offener (und nebenbei: auch für den offenen Teil einer halboffenen) Frage eine Textdatei einrichten, in der Sie die Antworten aller Befragungspersonen auf diese offene Frage als Text erfassen und die erfassten Texte danach überprüfen, wie häufig bestimmte Antworten (z.B. die Namen von Politikern) oder Antworttendenzen (z.B. alle Antworten, die Zustimmung zu einem Gesetz signalisieren) zu erkennen sind. Oder welche Tendenzen sich aus den Antworten ablesen lassen. Die Auswertung kann durch simples Auszählen erfolgen, wenn es nur darum geht, wie häufig ein Begriff oder eine Person genannt wird. Sollten komplexere Zusammenhänge in den Texten erkannt und bewertet werden, ist es erforderlich, die Analyse von mehr als einer Person vornehmen zu lassen, um subjektive Voreinstellungen der auswertenden Person(en) vermeiden oder zumindest kontrollieren zu können.

Wenn Sie für die offenen Fragen vorab ein Codeschema entwickeln, das die Hauptpunkte enthält, denen die möglichen Antworten auf die offenen Fragen zugeordnet werden können, können Sie diese Hauptpunkte nummerieren, die offenen Antworten dem jeweiligen Hauptpunkt zuordnen und die offene Frage als synthetische Variable mit Mehrfachnennungen zahlenmäßig in den Datensatz aufnehmen. Lautet die Frage z.B.

„In welcher Gegend Deutschlands würden Sie in diesem Herbst am liebsten Ihren Urlaub verbringen?" werden Sie alle möglichen Antworten bekommen, von „Schwarzwald" über „Harz" über „Pfälzer Weinstraße" bis hin zu „thüringische Seenplatte" oder „irgendwo nahe Berlin". Interessiert Sie für die Auswertung nur, ob die gewünschte Gegend im Norden, Westen, Süden oder Osten der Republik liegt, würden Sie über ein Codeschema wie z.B. „Norden" = 1, „Westen = 2, Süden = 3, „Osten" = 4" eine neue, synthetische Variable schaffen, welche die in der offenen Frage genannten Regionen einer dieser vier Kategorien zuordnen lassen. So bilden Sie eine synthetische, aber quantitative Variable, die in den Datensatz aufgenommen wird, wie eine „normale" Variable behandelt und mit anderen Variablen im Zusammenhang überprüft werden kann.

Das beschriebene Beispiel ist natürlich eher einfacher Natur. Lautet Ihre Frage z.B. „Wie stellen Sie sich Gott vor?“ würde das Codeschema, das Sie der Auswertung der Frage zugrunde legen, vermutlich differenzierter ausfallen, wodurch die Verkodung wesentlich schwieriger sein würde.

Urliste bzw. Rohdaten
Zurück zur eigentlichen Datenerfassung: Unabhängig davon, ob Sie jetzt für die Dateneingabe die Fragebogen- oder die Listeneingabe gewählt haben, sind Ihre Daten jetzt in der sog. Urliste der Befragung abgelegt (der Begriff „Urliste“ ist ein Terminus aus GrafStat, ganz allgemein sprechen wir von den „Rohdaten“). In der Urliste sind die Befragungsfälle zeilenweise abgelegt; in der obersten Zeile finden sich die Variablenbezeichnungen, die Sie bei der Eingabe in der „Kurzform“ vorgegeben haben.

Die Urliste wird Sie an eine Excel-Tabelle erinnern. Wollten Sie zur Erfassung der Daten mit Excel arbeiten, müssten Sie die Namen der Variablen ebenfalls in der ersten Zeile und die einzelnen Befragungsfälle ebenfalls zeilenweise eintragen. Nur dass Sie jetzt für jeden Fall und jede Variable direkt den Wert, der sich aus dem Fragebogen ergibt, manuell in die Tabelle eintragen, was deutlich fehleranfälliger ist als die Erstellung der Urliste in GrafStat über Fragebogen- oder Listeneingabe. Das Ergebnis wäre am Ende allerdings das Gleiche, nämlich eine in Tabellenform angeordnete Datenmatrix, in der die einzelnen Befragungsfälle zeilenweise abgelegt sind. Aufgrund der besseren Übersichtlichkeit bei der Dateneingabe und damit verbunden dem geringeren Risiko falscher Eingaben präferiere ich für schulische Zwecke die Dateneingabe mit GrafStat.

Weil jetzt immer wieder von falschen Eingaben die Rede war und davon, wie fehleranfällig die Eingabe der Daten sein kann, wollen wir uns im Folgenden mit der Frage beschäftigen, welche Fehler bei der Dateneingabe auftreten können und wie wir sie erkennen und beheben können.

4. Datenfehler und wie wir damit umgehen

Die häufigsten Datenfehler sind *fehlende Werte („missing data“) und „wild codes“;* ein wild code liegt vor, wenn ein Wert im Datensatz auftritt, den es gar nicht geben dürfte, z.B. wenn bei einer Fünfer-Skala im Fragebogen der Wert 8 im Datensatz erscheint. Darüber hinaus können Filterfehler, *formale* und *inhaltliche Inkonsistenzen* auftreten. Filterfehler (siehe dazu weiter unten den Abschnitt „Filterprüfungen und Filterfehler“) können Sie schon bei der Datenerfassung eliminieren, aber auch im Datensatz noch auffinden und beheben. Wie das Auffinden von Filterfehlern setzt auch das Erkennen von Inkonsistenzen den Abgleich von zwei oder mehr Variablen voraus. Fehlende Werte und wild codes erkennen Sie bereits beim ersten Blick in die *Grundauszählung*.

Grundauszählung
Als *Grundauszählung* oder *Randverteilung* bezeichnen wir die Auszählung *aller* Variablen des Datensatzes für *alle* Befragungspersonen. Aus ihr ersehen wir, wie viele Personen (absolut und/oder relativ) auf die Ausprägungen der einzelnen Variablen verteilt sind, wie viele Personen z.B. bei der Variablen „Geschlecht“ „männlich“ oder „weiblich“ sind, wie sich die Personen auf die 5 Punkte einer 5er-Skala zu einer Frage verteilen usw. – für alle Fragen über alle BefragungsteilnehmerInnen.

Wir beginnen die Datenbereinigung am besten mit einer Durchsicht der Grundauszählung, genau genommen sehen wir uns in der Grundauszählung als Erstes die

Variable „Fallnummer“ an. In dieser Variablen darf es jeden Fall nur ein einziges Mal geben; finden sich hier eine oder mehrere Fallnummern mehr als einmal, müssen Sie in die Originalfragebogen gehen und nachsehen, was passiert sein könnte. Oft ist es so, dass sich eine Doppelvergabe (z.B. der Fallnummer 107) dadurch erklärt, dass eine angrenzende Fallnummer (z.B. 106 oder 108) irrtümlicherweise nicht vergeben wurde, also im Datensatz fehlt. Der Blick in die Originalfragebogen klärt dieses Missverständnis schnell auf. Die korrekte Verkodung der Fallnummern ist die Voraussetzung für das Erkennen von Datenfehlern und – vor allem – für die Behebung dieser Fehler in der Datenbereinigung. Sehen Sie also in den Originalfragebogen nach und korrigieren Sie den Fehler im Datensatz manuell.

Erst wenn Sie das gemacht haben, sollten Sie sich die Grundauszählung der kompletten Daten anschauen. Welche Fehler werden Sie dabei entdecken? Am einfachsten sind missing data und wild codes zu erkennen.

Zunächst stellt sich aber noch die Frage, wie Sie eine Grundauszählung erstellen. In GrafStat geht das ganz schnell und komfortabel. Unter „Daten auswerten und präsentieren“ finden Sie den Button „Grundauswertung“. Wenn Sie den anklicken, erhalten Sie die Grundauszählung mit absoluten und relativen Häufigkeiten für alle Variablen und darüber hinaus einige statistische Kennwerte (für Skalenfragen z.B. Mittelwert und Median, für numerische Variablen Mittelwert, Minimum und Maximum). Ein besonderer Vorteil von GrafStat: In der Grundauszählung werden fehlende Werte explizit ausgewiesen („ohne Antwort“), und aufgrund der programminternen Regeln bei der Fragebogenerstellung bleiben Ihnen auch wild codes erspart, da Sie den Bereich gültiger Werte im Fragebogen festgelegt haben. Dennoch wollen wir diese Datenfehler kurz darstellen.

Missing data

Missing data liegen dann vor, wenn die Anzahl aller gültigen Antworten auf eine Frage niedriger ist als die Anzahl der Personen, die diese Frage hätten beantworten sollen. Machen wir ein ganz einfaches Beispiel: Bei einer Frage mit drei Antwortmöglichkeiten stehen im Datensatz 42 Fälle mit der Antwort „ja“, 18 Fälle mit „vielleicht“ und 37 Fälle mit „nein“, insgesamt also 97 gültige Werte. Da wir 100 Personen befragt haben, fehlen die Angaben von 3 Personen, bleiben also 3 missings. Wir sehen uns jetzt im Datensatz die Spalte mit der entsprechenden Variablen an und identifizieren die Fälle, für die kein Wert existiert. Dann schauen wir nach, um welche Fallnummern es sich handelt und ziehen für diese drei Fälle die Originalfragebogen heran. Sind die missings bereits im Fragebogen enthalten, belassen wir sie auch im Datensatz; sind im Fragebogen aber für den einen oder anderen Fall gültige Werte enthalten, und der Fehler ist bei der Dateneingabe erfolgt, korrigieren wir die entsprechenden Fälle im Datensatz manuell.

Wild codes

Wild codes sind ebenfalls auf den ersten Blick zu erkennen. Wir prüfen bei jeder Frage, ob es im Datensatz Fälle gibt, die einen Wert enthalten, den es aufgrund des Fragebogens nicht geben darf. Wir finden z.B. bei einer Frage mit 5er-Skala im Datensatz den Wert 8. Wir suchen im Datensatz wieder nach der Fallnummer für diesen Fall und holen uns auch hier den Originalfragebogen. Meist finden wir dann den „richtigen“ Wert, der bei der Datenerfassung falsch eingegeben worden ist (wie schnell erwischt man im Tastaturblock die 8 statt der 5!), im Fragebogen und korrigieren den Fehler im Datensatz manuell. Sollte das nicht eindeutig möglich sein, weil z.B. im Fragebogen zwei

Kästchen der Skala angekreuzt sind, müssen wir ein missing data vergeben, weil wir nicht zwei Skalenpunkte gleichzeitig verwerten können.

Wild codes sind also eher einfach zu korrigieren. Was wir auf diese Weise aber nicht erkennen können, ist der Fall, in dem ein gültiger Wert im Fragebogen (z.B. der Wert 4 auf der 5er-Skala) im Datensatz falsch, aber mit *einem anderen gültigen Wert* (z.B. mit 3 oder 5) eingegeben wurde.

Das Auffinden falsch eingetragener, aber gültiger Werte lässt sich ohne den immensen Aufwand einer doppelten Verkodung und dem Abgleich dieser beiden Verkodungen oder durch manuelles Überprüfen jeder Zahl des Datensatzes unter Verwendung der Originalfragebogen nicht leisten. Wenn Sie mich fragen: Von kürzeren Fragebogen und/oder geringen Fallzahlen abgesehen zu viel Aufwand für ein schulisches Forschungsprojekt. Und selbst bei kommerziellen oder akademisch verfassten Umfrageprojekten wird diese aufwändige Doppelbearbeitung der Datensätze nicht geleistet; dauert zu lange und kostet zu viel Geld.

Filterprüfungen und Filterfehler

Von einem „Filter" im Fragebogen sprechen wir dann, wenn das Beantworten einer Frage von der Antwort auf eine bestimmte vorhergehende Frage abhängig gemacht wird. Fragen wir z.B. danach, ob jemand Kinder unter 18 Jahren im Haushalt hat; die darauf folgenden Fragen zum Alter und zum Schulbesuch der Kinder können sinnhaft nur an Personen gestellt werden, die diese „Filterfrage" mit „Ja" beantwortet haben. Wer angibt, keine Kinder unter 18 Jahren im Haushalt zu haben, muss die Nachfragen zu den Kindern überspringen, wir sagen: wird überfiltert, und macht mit der nächsten Frage weiter, auf die er gefiltert wird, im Beispiel Fragen im Anschluss an die Nachfragen zu den Kindern.

Fragenfilter sind recht fehleranfällig, insbesondere beim Selbstausfüller. Es gibt im Fragebogen zwei Arten von Filterfehlern: Jemand beantwortet eine Frage, die er aufgrund des Filters eigentlich gar nicht hätte beantworten sollen. Oder jemand sollte aufgrund des Filters eine Frage beantworten, tut es aber nicht. Beide Fälle erkennen wir in der Grundauszählung: Im ersteren Fall gibt es bei der Frage, auf die gefiltert wurde, mehr Nennungen als durch den Filter definiert, im zweiten Fall weniger als durch den Filter vorgegeben.

Zur Korrektur der falschen Daten suchen wir in der Urliste die Variable, deren Frage den Filter einleitet, und prüfen dann die Antworten auf die gefilterten Folgefragen; in den Fällen, in denen wir die Fehler erkennen, schauen wir auch hier in die Originalfragebogen und korrigieren die Daten manuell.

Waren die bisher besprochenen Fehler im Datensatz mehr oder weniger augenscheinlich, müssen wir uns jetzt noch mit Fehlern beschäftigen, die zu entdecken Nachdenken im Vorfeld der Datenbetrachtung erforderlich macht: formale und inhaltliche Inkonsistenzen. In beiden Fällen müssen Sie vorab überlegen, bei welcher Kombination von Variablen Ihres Fragebogens Inkonsistenzen im Antwortverhalten nicht auszuschließen sind.

Formale Inkonsistenzen

Formale Inkonsistenzen im Datensatz treten bei Antworten auf mindestens zwei Fragen auf, die objektiv überprüfbar sind oder zumindest überprüfbar sein könnten. Sie müssen also zunächst überlegen, welche Variablenpaare davon betroffen sein könnten, und Daten zu diesen Variablen miteinander abgleichen. Nehmen wir als Beispiel die Variablen „Alter in Jahren" und „Anzahl der Kinder unter 18 Jahren". Wenn jetzt in

den Daten für eine Befragungsperson das Alter mit 25 angegeben ist und die Anzahl der Kinder mit 8, so ist diese Datenkombination nicht völlig unmöglich, aber doch sehr unwahrscheinlich. Oder Sie gehen von einem Zusammenhang zwischen der Variablen „Gewicht in kg“ und „Größe in cm“ aus: Finden Sie in den Daten eine Person, die 161 cm groß ist und 115 Kilo wiegt, ist ebenfalls anzunehmen, dass hier bei mindestens einer der Variablen ein Eingabefehler unterlaufen ist. Sie merken in beiden Beispielen: Die Werte für jede einzelne Variable sind durchaus zulässig, aber die Kombination der Daten aus beiden sollte überprüft werden.

Am einfachsten geht das, wenn Sie die Urliste so umsortieren, dass die Sie interessierenden Variablen im Datensatz in nebeneinanderstehenden Spalten abgebildet werden, oder Sie ziehen die entsprechenden Daten aus dem Datensatz heraus und legen sie in Kopie in separaten Dateien ab. So oder so haben Sie die Daten der interessierenden Variablen nebeneinander und damit im Blickfeld.

Suchen Sie in der Urliste diejenigen Kombinationen von Daten heraus, die Ihnen nicht plausibel erscheinen, ersehen Sie aus der Variable „Fallzahl“ die Fallnummern, nehmen Sie sich die Originalfragebogen vor und schauen Sie nach, wie die Angaben im Fragebogen tatsächlich sind. Stellen Sie fest, dass die Angaben dort nicht richtig in den Datensatz übertragen worden sind, korrigieren Sie den Datensatz manuell.

Aber was machen Sie, wenn die Angaben im Datensatz die Angaben im Fragebogen richtig wiedergeben, wenn also, wie im Beispiel oben, bei der entsprechenden Befragungsperson das Alter mit 25 und die Anzahl der Kinder mit 8 eingetragen ist? Sie können natürlich festlegen, dass Sie die Daten so übernehmen, wie sie im Fragebogen enthalten sind, auch wenn das sehr unplausibel erscheint. Sie können aber auch festlegen, dass Sie den plausibleren Wert, also das Alter 25 Jahre beibehalten und den weniger plausiblen Wert, also die Anzahl der Kinder mit 8 als missing verkoden. Diese Entscheidung obliegt Ihnen ganz allein, aber wenn Sie die Daten unverändert lassen, sollten Sie später, wenn Sie die Ergebnisse berichten, darauf hinweisen, dass Sie das Problem erkannt und warum Sie es so behandelt haben, wie Sie das getan haben.

Inhaltliche Inkonsistenzen

Bei inhaltlichen Inkonsistenzen ist die Problemlage vergleichbar mit den formalen Inkonsistenzen, nur dass es hier nicht um den Zusammenhang zwischen objektiv überprüfbaren Variablen geht, sondern um den Zusammenhang zwischen Variablen, die nicht überprüfbar sind, also im Bereich von Einstellungen oder Wertorientierungen oder ähnlichen nicht direkt erkennbaren, latenten Merkmalen. Beispiel: Im Datensatz wurde für eine Person bei Parteipräferenz „Die Linke“ verkodet, auf der Links-Rechts-Skala zur Einschätzung der eigenen politischen Grundhaltung der Skalenwert 7 als Wert für „rechts“. Dies erscheint zunächst unplausibel, also gehen Sie – mit Hilfe der Variablen „Fallnummer“ – in den Originalfragebogen und schauen nach. Liegt ein Eingabefehler vor, können Sie den im Datensatz manuell korrigieren.

Aber was machen Sie, wenn die Angaben im Datensatz die Angaben im Fragebogen richtig wiedergeben? Ganz einfach: Sie tun nichts, sondern lassen die Daten im Datensatz unverändert stehen. Die Welt ist komplex, vielleicht kann man wirklich politisch rechts stehen und dennoch die Partei „Die Linke“ gut finden. Zumindest können Sie nicht belastbar nachweisen, dass es diese Möglichkeit nicht geben könnte. Inhaltliche Inkonsistenzen sind „objektiv“ nicht auflösbar, sollten also im Datensatz so bleiben, wie sie im Fragebogen enthalten sind.

5. Zur Vorgehensweise im Unterricht

Zur Vorgehensweise im Unterricht empfehle ich, die Vorbereitung der Datenerfassung – Welche Fragebogenfehler, welche Datenfehler gibt es? Nach welchen Datenfehlern sollen wir suchen? Welche Daten sollen auf Inkonsistenzen überprüft werden?, etc. – in der gesamten Lerngruppe vorzunehmen. Das sichert nicht nur in gleicher Weise den Erkenntnisgewinn für alle SchülerInnen, sondern führt auch zu einer gemeinsamen Grundlage für die anstehende konkrete Datenerfassung und Datenbereinigung.

Die Datenerfassung selbst sollte dann in Kleingruppen erfolgen, von der Paginierung der Fragebogen, über die Fehlersuche und Fehlerkorrektur in den Originalfragebogen, die Dateneingabe und die Datenbereinigung bis hin zur Fertigstellung des Datensatzes. Die Datensätze der einzelnen Arbeitsgruppen können dann ganz am Ende zu einem gemeinsamen Datensatz zusammengefasst werden.

6. Zum Schluss

Wir haben jetzt die Angaben in unseren Fragebogen in eine Datei übertragen, die Daten nach Datenfehlern abgeprüft und die aufgefundenen Fehler bereinigt. Wenn man sich mit dem Thema „Befragungen“ nicht als MethodenforscherIn beschäftigt, sondern – wie Sie das vermutlich tun – Befragungen als Vehikel ansieht, um inhaltliche Fragestellungen zu bearbeiten, nähern Sie sich jetzt dem eigentlichen Ziel Ihrer schulischen Forschungsarbeit: Auswertung der Daten, Interpretation der Auswertungsergebnisse und Präsentation der zentralen Befunde. Mit dieser Thematik werden wir uns im nächsten Beitrag zur Reihe „Schüler forschen“ beschäftigen.

Literatur

Porst, R (2014): Sekundäranalyse und Zugang zu sozialwissenschaftlichen Daten. S. 553-562 in Gesellschaft • Wirtschaft • Politik 63, Heft 4

Porst, R. (2015): Fragebogenkonstruktion (I) – Grundlagen, Arten von Fragen, Arten von Skalen. S. 239-250 in Gesellschaft • Wirtschaft • Politik 64, Heft 2

Anmerkungen

1 Ruth Holthof ist als Studienrätin im Fach Sozialkunde am Eleonoren-Gymnasium in Worms tätig.

2 http://www-01.ibm.com/software/de/stats24/

3 http://www.sas.com/de_at/software/sas9.html

4 http://www.stata.com/products/

5 Der Verweis auf die Programme Excel und GrafStat und im weiteren die Konzentration auf GrafStat bedeutet natürlich nicht, dass es keine anderen, in Verfügbarkeit und Preis entsprechenden Programme gäbe, die Sie in Ihrem Forschungsprojekt einsetzen könnten; diese Auswahl ist alleine der subjektiven Erfahrung des Autoren mit diesen Programmen geschuldet.

6 Sekundäranalytisch bedeutet, dass die Daten, die wir auswerten wollen, nicht von uns selbst erhoben wurden, sondern von anderen Personen, die ihre Daten Dritten zur weiteren Auswertung zur Verfügung stellen (s. dazu Porst 2014).

7 Zu den unterschiedlichen Fragenarten siehe Porst (2015).