© Lukas pexels.com

Datencheck

So bereiten Sie Ihre Daten richtig für die Auswertung vor

von Daniela Keller

Sie haben Ihre erhobenen Daten vorliegen und freuen sich: Endlich geht es los mit der statistischen Auswertung. Doch Stopp! Vergessen Sie bei all der Vorfreude nicht den Datencheck! Wenn Sie ihn überspringen, fällt Ihnen das womöglich später vor die Füße. Es kann passieren, dass Sie während der Auswertung auf Ungereimtheiten in den Daten stoßen. Diese müssen Sie dann bereinigen und alle ihre bisher durchgeführten Auswertungen wiederholen. Das kostet unnötig Zeit und Nerven. Mit einem durchdachten Datencheck können Sie sich diesen Ärger sparen.

Der Datencheck besteht aus gezielten deskriptiven Analysen, deren Ergebnisse Sie allerdings nicht direkt für ihren Bericht verwenden. Stattdessen treffen Sie basierend auf diesen Ergebnissen Entscheidungen über die Bereinigung Ihres Datensatzes. Mit dem Datencheck stellen Sie sicher, dass Ihre Daten eine gute Qualität aufweisen und für die geplanten Auswertungen geeignet sind.

Welche Schritte Sie genau bei Ihrem Datencheck gehen, hängt sehr vom Studiendesign und der Art Ihrer Daten ab. Ich gebe Ihnen die Schritte an die Hand, die ich in vielen Fällen für sinnvoll erachte. Bedenken Sie aber, dass dieser Prozess nicht in Stein gemeißelt ist und immer individuell angepasst werden sollte.

Schritt 1: Anzahl der Zeilen und Spalten und Datentyp
Nachdem Sie Ihre Daten in die Statistiksoftware importiert haben, werfen Sie zunächst einen Blick auf die Anzahl der Zeilen und Spalten der Datentabelle. Sind diese in Ordnung? Ist beim Import nichts schiefgelaufen?

Außerdem betrachten Sie das Datenformat: Alle Variablen, die als Zahl eingegeben wurden, sollen auch als Zahl von der Statistiksoftware erkannt werden. Wenn Werte, die eigentlich Zahlen sein sollen, als Text erscheinen, gab es ein Problem beim Einlesen der Daten, beispielsweise durch Tippfehler verursacht.

Schritt 2: Deskriptive Untersuchung der einzelnen Variablen
Nun berechnen Sie für jede einzelne Variable die deskriptive Statistik. Insbesondere Minimum und Maximum sind hier wichtig. Anhand dieser Werte können Sie sehen, ob es ungewöhnliche Messwerte gibt, die auf Tippfehler oder andere Probleme hinweisen. Oder ob die Werte in einem plausiblen Bereich liegen.

Wenn Sie beispielsweise nur erwachsene Personen in Ihre Studie einschließen wollten, beim Alter nun aber der Minimalwert 15 ist, dann ist das entweder ein Tippfehler oder die Person hat fälschlicherweise teilgenommen. In beiden Fällen müssen Sie sich darum kümmern und entscheiden, was mit den Werten dieser Person geschehen soll.

Auch die Anzahl der fehlenden Werte pro Variable sollten Sie hier beachten. Hohe Anzahlen von fehlenden Werten können auf Probleme bei der Dateneingabe hinweisen, denen Sie nachgehen sollten.

Schritt 3: Untersuchung der Zeitvariablen bei Online-Umfragen
Wenn Ihre Daten aus einer Online-Umfrage stammen, haben Sie meist auch Angaben der Zeit, die die Person mit dem Ausfüllen Ihres Fragebogens zugebracht hat. Darauf sollten Sie ein besonderes Augenmerk legen.

Überlegen Sie sich, was eine realistische Dauer für das Ausfüllen Ihres Fragebogens ist. Und ermitteln Sie davon ausgehend, was dagegen sehr abwegige Zeiten (viel zu kurz oder viel zu lang) wären.

Untersuchen Sie, wie lange die Teilnehmenden fürs Ausfüllen benötigt haben (wieder Minimum und Maximum) und schließen Sie Fälle aus, die außerhalb der von Ihnen gesteckten Grenzen liegen.

Schritt 4: Untersuchung fehlender Werte
Fehlende Werte sind ein häufiges Problem in der Datenanalyse und sollten sorgfältig behandelt werden. Es gibt sehr tiefgehende Analysen, die man zur Untersuchung der Struktur fehlender Werte einsetzen kann. Hier will ich nur auf die grundlegendste Analyse pro Fall (Zeile im Datensatz) eingehen, um davon ausgehend gegebenenfalls Fälle auszuschließen.

Bestimmen Sie die Anzahl der fehlenden Werte pro Fall (also pro Zeile), eventuell auch für Variablensets (Fragebogenteile) separat. Überlegen Sie sich wieder, wie viele fehlende Werte Sie zulassen möchten und legen Sie damit eine Obergrenze fest, ab der Sie einzelne Datensätze ausschließen.

Schritt 5: Suche nach Mustern im Antwortverhalten
Bei der Arbeit mit Fragebögen-Daten kann es hilfreich sein, nach Mustern im Antwortverhalten zu suchen, die auf ein achtloses Ausfüllen des Fragebogens hindeuten. Wenn beispielswiese eine Person bei Likert-Items immer die Mitte oder immer den kleinsten Wert angekreuzt hat, hat sie das eventuell getan, um schnell durch den Fragebogen zu kommen und sich weniger Gedanken machen zu müssen. Mit solchen Daten wollen Sie in Ihrer Auswertung nicht arbeiten.

Diese Muster können Sie erkennen, indem Sie pro Zeile für ein bestimmtes Variablenset (z.B. Likert-Items zum gleichen Konstrukt) die Spannweite (Maximum minus Minimum) oder die Standardabweichung berechnen. Im Anschluss betrachten Sie diese Streuparameter und finden so auffällige Zeilen (auffällig = kleiner Wert = kleine Streuung = wenig unterschiedlich geantwortet).

Zusammen mit den Informationen zur Zeitmessung (siehe Schritt 3) oder mit der Anzahl der fehlenden Werte (Schritt 4) können Sie entscheiden, ob Sie die Daten dieser Zeile aus Ihrem Datensatz löschen wollen.

Es gibt kein Standard-Vorgehen!
Sie haben es sicher schon bemerkt: Ich kann Ihnen hier keine harten Kriterien für die Entscheidungen an die Hand geben. Sie sollen selbst festlegen, ab wann es zu viele fehlende Werte sind oder welcher Messwert unrealistisch ist. Das geht (leider) nicht anders. Beim Datencheck wird sehr viel inhaltlich und nach gesundem Menschenverstand entschieden.

Deshalb ist es umso wichtiger, dass Sie Ihr Vorgehen gut dokumentieren. Notieren Sie sich alle Entscheidungen

und die Gründe dafür während des Datenchecks. So sind die Schritte erstmal für Sie nachvollziehbar. Später werden Sie diese Notizen brauchen, um Ihr Vorgehen transparent zu berichten. Zudem ergeben sich aus diesen Überlegungen oft Punkte, die Sie in Ihrer Arbeit diskutieren sollten.

Es ist auch sehr hilfreich, wenn Sie schon in diesem frühen Stadium eine Idee davon haben, wie Sie weiter mit der Datenauswertung vorgehen werden: Welche Analysen und Methoden sind geplant? Sowohl fehlende Werte als auch Ausreißer sind bei einigen Methoden problematisch, bei anderen wiederum nicht.

Und ganz wichtig: Behalten Sie immer im Blick, dass die Datenbereinigung Ihre Stichprobe auch inhaltlich verändern kann. Stellen Sie sich vor, Sie haben eine große Anzahl an fehlenden Werten beim Alter. Gehen wir davon aus, dass eher ältere Frauen ihr Alter seltener angeben. Dann würden Sie durch den Ausschluss von Personen mit fehlender Altersangabe vor allem ältere Frauen aus Ihrem Datensatz löschen. Das hat zur Folge, dass der resultierende Datensatz eher männlich und jünger ist. Solche Zusammenhänge müssen Sie im Blick behalten.

Gehen Sie also – wie immer bei der Datenanalyse – mit Verstand und Bedacht ans Werk. Und trauen Sie sich, Entscheidungen zu treffen. Es gibt beim Datencheck nicht immer ein Richtig und ein Falsch.

© Uschi Mattke

Die Autorin

Daniela Keller ist leidenschaftliche Statistik-Expertin und berät Studierende und Wissenschaftler*innen zu allen Themen der statistischen Datenanalyse. Während ihres Studiums der Diplom-Mathematik gründete sie mit Kommilitonen eine studentische statistische Beratung und arbeitete anschließend selbstständig in diesem Feld. Neben Einzelberatungen und Workshops unterstützt sie ihre Kund*innen seit 2019 mit der Statistik-Akademie, ihrem Online-Mitgliederbereich für alle, die Statistik verstehen und selbstständig anwenden wollen. Ihr Blog (www.statistik-und-beratung.de/blog) und ihr YouTube-Kanal sind Fundgruben für leicht verständlich aufbereitetes Statistikwissen für die Praxis.