

In die Tiefe: Datenqualität und -aufbereitung

Ein kompakter kurzer Einstieg in Stichprobenkontrolle, Skalenniveaus, fehlende Werte und die Bildung neuer Variablen.

Kurz gesagt

Datenaufbereitung bedeutet, erhobene Daten vor der eigentlichen Analyse zu prüfen, zu ordnen und für die Auswertung vorzubereiten. Dabei geht es unter anderem um die Qualität der Stichprobe, plausible Werte, fehlende Angaben und eine nachvollziehbare Dokumentation.

1. Datenqualität prüfen

Nach der Datenerhebung sollte zunächst geprüft werden, ob die Daten plausibel und für die weitere Auswertung geeignet sind. Eine wichtige Möglichkeit ist die Stichprobenkontrolle: Sie hilft dabei, systematische Verzerrungen zu erkennen und die Plausibilität der Daten zu kontrollieren.

Plausibilitätsprüfung

Bei einer Plausibilitätsprüfung wird kontrolliert, ob die erhobenen Werte logisch und nachvollziehbar sind. Offensichtlich unplausible oder fehlerhafte Angaben (z.B. 1865 für das Geburtsjahr) können die spätere Analyse verzerren. Auffällige Werte sollten nicht automatisch gelöscht, sondern es sollte zunächst auf mögliche Eingabe-, Übertragungs- oder Erhebungsfehler geprüft werden. Wird ein Wert korrigiert oder entfernt, sollte diese Entscheidung nachvollziehbar dokumentiert werden.

Systematische Verzerrungen

Neben einzelnen fehlerhaften Angaben können auch systematische Verzerrungen auftreten. Ist eine bestimmte Personengruppe im Vergleich zur Grundgesamtheit deutlich unterrepräsentiert, kann, sofern noch Ressourcen vorhanden sind, eine gezielte Nacherhebung sinnvoll sein. Dadurch kann die Aussagekraft der Ergebnisse verbessert werden.

Dokumentation der Erhebung

Auch die Rahmenbedingungen der Datenerhebung sollten dokumentiert werden. Veränderungen oder besondere Ereignisse während der Erhebung können die Antworten beeinflussen und sind deshalb wichtig, wenn die Ergebnisse später eingeordnet werden. In Museen und Science Centern können beispielsweise Ausfälle oder Veränderungen von Ausstellungsbereichen und Exponaten, Bauarbeiten, geänderte Wegführungen oder besondere Veranstaltungen eine Rolle spielen. Auch

unterschiedliche Erhebungsbedingungen – etwa der Ort und Zeitpunkt der Befragung – sollten festgehalten werden. So lässt sich später besser nachvollziehen, ob Besonderheiten in den Ergebnissen möglicherweise mit den Bedingungen während der Erhebung zusammenhängen.

Wichtig

Datenqualität betrifft nicht nur einzelne falsche Werte. Auch die Zusammensetzung der Stichprobe und die Bedingungen, unter denen die Daten erhoben wurden, können die Aussagekraft der Ergebnisse beeinflussen.

2. Skalenniveaus

Nicht alle erhobenen Daten enthalten die gleiche Art von Informationen. Das Skalenniveau beschreibt, welche Eigenschaften die Werte einer Variable besitzen und wie sie miteinander verglichen werden können. Manche Variablen bestehen aus verschiedenen Kategorien, andere lassen sich zusätzlich in eine Rangfolge bringen und bei wieder anderen können auch die Abstände zwischen den Werten sinnvoll interpretiert werden. Das Skalenniveau ist deshalb eine wichtige Grundlage für die Datenaufbereitung und die spätere Auswertung. Es bestimmt, welche statistischen Kennwerte und Analyseverfahren für die jeweiligen Daten sinnvoll eingesetzt werden können. Bevor die eigentliche Analyse beginnt, sollte daher geklärt werden, ob eine Variable nominal, ordinal oder metrisch skaliert ist. Die folgende Übersicht zeigt die wichtigsten Unterschiede zwischen den drei Skalenniveaus und gibt Beispiele für typische Variablen.

Skalenniveau	Was bedeutet das?	Was ist möglich?	Beispiele
Nominal	Werte lassen sich nur verschiedenen Kategorien zuordnen. Es gibt keine sinnvolle Reihenfolge.	Kategorien und Häufigkeiten vergleichen	Geschlecht, Besuchsgrund, Wohnort, genutztes Verkehrsmittel
Ordinal	Werte lassen sich in eine Rangfolge bringen, aber die Abstände zwischen ihnen sind nicht eindeutig messbar	Auswertungen auf Basis der Rangfolge	Schulnoten, Zufriedenheitsstufen, Einstufungen von „stimme gar nicht zu“ bis „stimme voll zu“
Metrisch	Werte lassen sich ordnen und die Abstände zwischen ihnen sinnvoll messen und vergleichen	z.B. Mittelwerte, Standardabweichungen oder Korrelationen	Alter, Besuchsdauer, Anzahl der Museumsbesuche

Nominales Skalenniveau

Bei einem nominalen Skalenniveau werden die erhobenen Werte verschiedenen Kategorien zugeordnet. Zwischen diesen Kategorien besteht keine natürliche Rangfolge, sie sind also weder höher noch niedriger als andere Kategorien. Deshalb können vor allem die Häufigkeiten und Anteile der einzelnen Kategorien miteinander verglichen werden. Berechnungen wie Mittelwert und Standardabweichung sind bei nominalen Daten daher nicht sinnvoll.

Ordinales Skalenniveau

Auch bei einem ordinalen Skalenniveau werden Werte Kategorien zugeordnet. Diese besitzen jedoch eine sinnvolle Rangfolge. Es lässt sich also sagen, dass eine Ausprägung höher oder niedriger als eine andere ist. Wie groß der Abstand zwischen zwei Ausprägungen tatsächlich ist, lässt sich jedoch nicht eindeutig bestimmen. Auswertungen können deshalb die Reihenfolge der Werte berücksichtigen, ohne von gleich großen Abständen zwischen ihnen auszugehen.

Metrisches Skalenniveau

Bei einem metrischen Skalenniveau liegen Zahlenwerte vor, deren Abstände sinnvoll interpretiert und miteinander verglichen werden können. Es lässt sich also nicht nur bestimmen, welcher Wert größer oder kleiner ist, sondern auch, wie groß der Unterschied zwischen zwei Werten ist. Dadurch sind weiterführende statistische Berechnungen möglich, beispielsweise die Berechnung von Mittelwerten und Standardabweichungen oder die Untersuchung von Zusammenhängen.

Kurz gesagt

Nominal = Kategorien ohne Rangfolge

Ordinal = Kategorien mit Rangfolge

Metrisch = Werte mit messbaren Abständen

3. Fehlende Werte erkennen und behandeln

Fehlende Werte liegen vor, wenn für einzelne Personen oder Variablen keine Angaben vorhanden sind. Dafür kann es verschiedene Gründe geben: Eine Frage wurde nicht beantwortet, eine Befragung wurde vorzeitig beendet, es trat ein technisches Problem auf oder eine Frage wurde aufgrund der Filterführung gar nicht angezeigt. Im Datensatz sind fehlende Werte häufig als leere Zellen erkennbar.

Warum sind fehlende Werte wichtig?

Fehlende Werte sind nicht automatisch problematisch. Sie können jedoch die Aussagekraft einer Auswertung beeinflussen, insbesondere wenn viele Angaben fehlen. Je mehr Werte fehlen, desto weniger Informationen stehen für die Analyse zur Verfügung. Dadurch kann sich beispielsweise die Anzahl der Fälle verringern, auf denen ein Ergebnis beruht.

Deshalb sollte zunächst geprüft werden, wie viele Werte fehlen, bei welchen Variablen sie auftreten und ob sich mögliche Gründe für die fehlenden Angaben erkennen lassen.

Nicht jede fehlende Angabe ist zufällig

Bei fehlenden Werten ist nicht nur wichtig, wie viele Angaben fehlen, sondern auch warum sie fehlen. Es kann vorkommen, dass bestimmte Fragen von bestimmten Personengruppen häufiger nicht beantwortet werden. Werden in einem solchen Fall alle Personen mit fehlenden Angaben aus der Auswertung ausgeschlossen, kann sich die Zusammensetzung der Stichprobe verändern. Bestimmte Gruppen sind dann möglicherweise stärker oder schwächer vertreten als zuvor. Dadurch können die Ergebnisse verzerrt werden. Aus diesem Grund sollte vor der Entscheidung über den Umgang mit fehlenden Werten geprüft werden, ob sich bestimmte Muster bei den fehlenden Angaben erkennen lassen.

Wie kann mit fehlenden Werten umgegangen werden?

Es gibt nicht die eine richtige Vorgehensweise für den Umgang mit fehlenden Werten. Welche Methode geeignet ist, hängt unter anderem von der Fragestellung, den geplanten Analysen sowie vom Umgang und den möglichen Ursachen der fehlenden Angaben ab. Grundsätzlich gibt es verschiedene Möglichkeiten:

Vorgehensweise	Vorteil	Zu beachten
Unvollständige Fälle löschen	Analysen basieren auf einem vollständigen Datensatz	Die Stichprobe kann deutlich kleiner werden und wichtige Informationen können verloren gehen.
Je Auswertung vollständige Angaben nutzen	Es werden möglichst viele vorhandene Informationen verwendet	Die einzelnen Auswertungen können auf unterschiedlichen Fallzahlen beruhen; die Fallzahl sollte deshalb angegeben werden
Fehlende Werte ersetzen	Fehlende Angaben können durch geschätzte oder berechnete Werte ergänzt werden	Die Wahl der Methode muss begründet werden; genannt werden z.B. Mittelwert, Median oder der „nächstgelegene Nachbar“

Welche Vorgehensweise gewählt wird, sollte nachvollziehbar begründet und dokumentiert werden. Je nach Datensatz kann es auch sinnvoll sein, verschiedene Methoden miteinander zu kombinieren.

Transparenz

Wenn Auswertungen auf einer unterschiedlichen Anzahl gültiger Fälle beruhen, sollte die jeweilige Fallzahl angegeben werden. So lässt sich nachvollziehen, auf wie vielen Beobachtungen ein Ergebnis tatsächlich basiert.

4. Neue Variablen bilden

In manchen Fällen kann es sinnvoll sein, aus mehreren vorhandenen Variablen eine neue Variable zu bilden. Das kann beispielsweise ein Summenscore sein, bei dem die Werte mehrerer zusammengehöriger Fragen addiert werden, oder ein gemeinsamer Mittelwert aus mehreren Fragen, die dasselbe Merkmal erfassen. Auf diese Weise können mehrere Einzelinformationen zu einem übergeordneten Wert zusammengefasst werden. Das kann dabei helfen, komplexere Merkmale abzubilden, Ergebnisse übersichtlicher zu machen und die weitere Auswertung zu vereinfachen. Die Bildung einer neuen Variable sollte dabei immer inhaltlich begründet und nachvollziehbar dokumentiert werden.

Wichtig

Neue Variablen sollten nicht beliebig gebildet werden. Ihre Bildung sollte inhaltlich begründet und nachvollziehbar dokumentiert sein.

5. Orientierung: So kann bei der Datenaufbereitung vorgegangen werden

Bei der Datenaufbereitung werden die erhobenen Daten Schritt für Schritt geprüft und für die spätere Analyse vorbereitet. Die folgende Übersicht fasst die wichtigsten Punkte zusammen und kann als erste Orientierung dienen:

- 1. Datenqualität prüfen:** Daten auf unplausible oder fehlerhafte Werte kontrollieren und mögliche Verzerrungen der Stichprobe berücksichtigen.
- 2. Erhebung dokumentieren:** Besondere Rahmenbedingungen und Ereignisse festhalten, die die erhobenen Daten beeinflusst haben könnten.
- 3. Skalenniveau bestimmen:** Prüfen, ob die einzelnen Variablen nominal, ordinal oder metrisch skaliert sind.
- 4. Fehlende Werte prüfen:** Feststellen, wo und in welchem Umfang Angaben fehlen, mögliche Muster betrachten und einen geeigneten Umgang damit festlegen.

5. **Neue Variablen bilden (falls erforderlich):** Prüfen, ob für die geplante Auswertung sinnvoll und begründet neue Variablen gebildet werden sollen.
6. **Entscheidungen dokumentieren:** Änderungen und Entscheidungen bei der Datenaufbereitung nachvollziehbar festhalten.

Merksatz

Eine sorgfältige Datenaufbereitung schafft die Grundlage dafür, dass die anschließende Analyse nachvollziehbar und aussagekräftig durchgeführt werden kann.