Datenmanagement

Datenmanagement beinhaltet das Erstellen von Datenbanken und Tabellen, sowie die Aufbereitung der Daten in eine Form, die von den gängigen Statistikprogrammen verstanden wird. Außerdem sollen die Daten so vorliegen, dass statistische Analysen an ihnen vorgenommen werden können. Das heißt zum Beispiel, dass die Daten gesäubert sind, Teilmengen gebildet werden können, Variablen umbenannt wurden und auf fehlende Werte geachtet wird.

Im Folgenden beschreiben wir einige Grundsätze, die bei jeder Art von Datenmanagement zu beachten sind.

fu:stat bietet regelmäßig Schulungen für Hochschulangehörige sowie für Unternehmen und weitere Institutionen an. Die Inhalte reichen von Statistikgrundlagen (Deskriptive, Testen, Schätzen, lineare Regression) bis zu Methoden für Big Data. Es werden außerdem Kurse zu verschiedenen Software-Paketen gegeben. Auf Anfrage können wir auch gerne individuelle Inhouse-Schulungen bei Ihnen anbieten.

Inhaltsverzeichnis

- Struktur des Datensatzes
  - Variablen
  - Einträge
- Datensätze
  - Erstellen eines eigenen Datensatzes
  - Arbeiten mit einem vorgefertigten Datensatz
- Falsche Einträge erkennen/Datenvalidierung
- Behandlung Ausreißer und Fehlende Werte
- Große Datensätze/Big Data
- Datenmanagement Statistikprogramme
  - Verfügbare Programme
  - Programmspezifische Probleme/Aspekte

Struktur des Datensatzes

Die Struktur eines Datensatzes sollte in Form einer Tabelle sein. Die Spalten stehen hierbei für die Variablen, d.h. für die Messgrößen, also z.B. Körpergewicht oder Größe. Die Zeilen sind die einzelnen Beobachtungen, z.B. Personen. Die Einträge stellen dann den Wert der jeweiligen Variable für eine Beobachtung dar. Dies könnte zum Beispiel das Körpergewicht von Person 3 (75kg) sein. Das folgende Bild des ALLBUS Datensatzes ist beispielhaft für die Tabellenstruktur.

Unbekanntes Makro: 'vot

	ID	GESCHL	GEBJAHR	BERUFSTAETIG	ARBEITSSTD	ARZTBES	RAUCH	GRO	GEW	SCHULABSCHLUSS	SCHULABSCHLUSS_V	SCHULABSCHLUSS_M
1	1	WEIBLICH	1980	HAUPTBERUFL.GANZTAGS	40.0	0	NEIN	170	68	HOCHSCHULREIFE	VOLKS-,HAUPTSCHULE	FACHHOCHSCHULREIFE
2	2	WEIBLICH	1964	HAUPTBERUFL.GANZTAGS	43.0	1	NEIN	163	75	MITTLERE REIFE	VOLKS-,HAUPTSCHULE	VOLKS-,HAUPTSCHULE
3	3	MAENNLICH	1957	HAUPTBERUFL.GANZTAGS	70.0	0	JA	174	75	MITTLERE REIFE	VOLKS-,HAUPTSCHULE	VOLKS-,HAUPTSCHULE
4	4	MAENNLICH	1952	NEBENHER BERUFSTAE.	NA	2	NEIN	172	115	MITTLERE REIFE	VOLKS-,HAUPTSCHULE	VOLKS-,HAUPTSCHULE
5	5	WEIBLICH	1954	HAUPTBERUFL.HALBTAGS	20.0	1	NEIN	168	87	MITTLERE REIFE	FACHHOCHSCHULREIFE	FACHHOCHSCHULREIFE
6	6	MAENNLICH	1958	HAUPTBERUFL.GANZTAGS	60.0	0	NEIN	182	90	MITTLERE REIFE	VOLKS-,HAUPTSCHULE	MITTLERE REIFE
7	7	MAENNLICH	1947	NICHT ERWERBSTAETIG	NA	1	NEIN	178	101	VOLKS-,HAUPTSCHULE	VOLKS-,HAUPTSCHULE	VOLKS-,HAUPTSCHULE
8	8	WEIBLICH	1988	HAUPTBERUFL.GANZTAGS	40.0	6	NEIN	187	86	HOCHSCHULREIFE	HOCHSCHULREIFE	FACHHOCHSCHULREIFE
9	9	MAENNLICH	1955	NICHT ERWERBSTAETIG	NA	0	NEIN	180	100	FACHHOCHSCHULREIFE	VOLKS-,HAUPTSCHULE	VOLKS-,HAUPTSCHULE
10	10	MAENNLICH	1961	HAUPTBERUFL.GANZTAGS	35.0	8	NEIN	190	115	MITTLERE REIFE	VOLKS-,HAUPTSCHULE	VOLKS-,HAUPTSCHULE
11	11	MAENNLICH	1947	NICHT ERWERBSTAETIG	NA	1	NEIN	174	82	VOLKS-,HAUPTSCHULE	VOLKS-,HAUPTSCHULE	VOLKS-,HAUPTSCHULE
12	12	WEIBLICH	1968	HAUPTBERUFL.GANZTAGS	70.0	1	NEIN	168	56	MITTLERE REIFE	VOLKS-,HAUPTSCHULE	VOLKS-,HAUPTSCHULE
13	13	WEIBLICH	1962	HAUPTBERUFL.GANZTAGS	40.0	1	NEIN	162	54	MITTLERE REIFE	NA	NA
14	14	WEIBLICH	1956	HAUPTBERUFL.GANZTAGS	59.0	0	NEIN	167	80	MITTLERE REIFE	VOLKS-,HAUPTSCHULE	VOLKS-,HAUPTSCHULE
15	15	WEIBLICH	1964	NEBENHER BERUFSTAE.	NA	5	NEIN	162	63	VOLKS-,HAUPTSCHULE	VOLKS-,HAUPTSCHULE	VOLKS-,HAUPTSCHULE
16	16	MAENNLICH	1950	HAUPTBERUFL.GANZTAGS	58.0	3	NEIN	173	82	HOCHSCHULREIFE	VOLKS-,HAUPTSCHULE	VOLKS-,HAUPTSCHULE
17	17	WEIBLICH	1959	NICHT ERWERBSTAETIG	NA	15	NEIN	155	60	VOLKS-,HAUPTSCHULE	VOLKS-,HAUPTSCHULE	VOLKS-,HAUPTSCHULE
18	18	MAENNLICH	1953	HAUPTBERUFL.GANZTAGS	45.0	8	NEIN	175	105	MITTLERE REIFE	VOLKS-,HAUPTSCHULE	VOLKS-,HAUPTSCHULE
19	19	WEIBLICH	1989	HAUPTBERUFL.GANZTAGS	42.0	2	JA	160	58	FACHHOCHSCHULREIFE	VOLKS-,HAUPTSCHULE	MITTLERE REIFE
20	20	WEIBLICH	1946	NICHT ERWERBSTAETIG	NA	10	NEIN	158	58	MITTLERE REIFE	NA	VOLKS-,HAUPTSCHULE
21	21	MAENNLICH	1955	HAUPTBERUFL.GANZTAGS	40.0	2	JA	179	83	MITTLERE REIFE	VOLKS-,HAUPTSCHULE	VOLKS-,HAUPTSCHULE

Variablen

Jede Variable hat ein Skalenniveau. Hierunter versteht man, ob die Merkmalsauprägungen nominal-, ordinal oder intervallskaliert sind. Nominale oder ordinale Variablen werden auch als kategoriale Variablen bezeichnet. In fast allen Programmen lässt sich das Skalenniveau für jede Spalte separat einstellen.

Variablennamen

Definieren Sie sich selbst ein Konventionssystem für die Variablenbenennung, die selbsterklärend ist. Die Variablennamen sollten kurz und prägnant sein, z.B. *gro* für Größe und *gew* für Gewicht. Längere Beschreibungen sollten separat gespeichert werden. Eine mögliche Benennung ist die Nummer im Fragebogen oder eine allgemeine aufsteigende Nummerierung, z.B. F1, F2, F3 oder V01, V02, V03. Falls man mit längeren Datensätzen arbeitet, ist die Kombination aus selbst gewähltem Präfix, Stamm und Suffix in Variablenamen verständlicher, z.B. nach Jahr, Fragenmodul oder Thematik.

Einfach-/Mehrfachnennung

Sollen bei einer Umfrage Mehrfachnennung möglich sein, so muss für jede der möglichen Antworten eine eigene Variable angelegt werden.

Dummy-Variablen

Manche Statistikprogramme "verstehen" kategoriale Variablen mit mehr als zwei verschiedenen Merkmalsausprägungen, bei anderen Programmen muss eine kategoriale Variable mit k Merkmalsausprägungen in k (oder k-1) Dummy-Variablen aufgeteilt werden, die jeweils nur 0 (Merkmal nicht vorhanden) und 1 (Merkmal vorhanden) als Merkmalsausprägungen haben.

Eine Variable pro Merkmal

Häufig werden neben den eigentlichen Variablen auch Merkmalskombinationen analysiert. Wird einer Versuchsperson bspw. eine Kombination aus Bild und Text vorgelegt, so ist man auch an der Kombination von Bild und Text interessiert. Diese Interaktion wird über eine Variable gemessen, die alle Kombinationen von Bild und Ton wiedergibt. Um jedoch den Effekt der einzelnen Merkmale zu messen, braucht man im Datensatz auch immer die Originalmerkmale, also eine Variable für Bild und eine Variable für Text. Also: Für jedes Merkmal eine Variable und manchmal auch eine für die Kombination mit anderen Merkmalen.

Einträge

Umlaute

Umlaute, Akzente sowie Sonderzeichen abgesehen von '.', '_' und '-' sollten generell vermieden werden, da diese teils zu Fehlermeldungen führen, bzw. der Text von Statistikprogrammen nicht richtig erkannt wird. Dies gilt ebenso für Variablen- und Dateinamen.

Punkt und Komma

Die meisten Programme arbeiten mit der internationalen Schreibweise von Dezimalzahlen, bei der ein Punkt statt des deutschen Kommas benutzt wird. Entsprechend sollten Einträge auch in dieser Form vorgenommen werden.

Maßeinheiten

Maßeinheiten sollten nicht Teil der Einträge sein. Statt 1000€ als Eintrag, sollte 1000 eingetragen werden und die Einheiten in der Variablenbeschreibung erwähnt werden (Haushalteinkommen in Euro).

Fehlende Werte

Fehlende Werte (missing values) sollten sorgfältig behandelt werden, um ungewollte Auswirkungen auf die Analyse zu vermeiden. Deshalb sollten Fehlende Werte bei jeder Variable vollständig und eindeutig definiert/codiert werden. Eine leider weit verbreitete Unsitte ist die Konvention für fehlende Werte je nach Variable out-of-range Zahlen zu benutzen, z.B. -999.99 oder 99. Bei einigen Statistikprogrammen werden fehlende Werte durch NA (Not Available) oder '.' gekennzeichnet.

Eigentlich sollte man mindestens zwei Zeichen für fehlende Werte benutzen, da es zwei unterschiedliche Gründe für fehlende Angaben gibt: "Trifft nicht zu" oder "Keine Angabe". Viele Statistik Programme unterstützen die Benutzung von mehreren Missing Codes.

Datensätze

Erstellen eines eigenen Datensatzes

Schon vor der ersten Befragung oder dem ersten Experiment, sollte der Aufbau des Datensatzes klar sein. Hierduch lassen sich eventuelle spätere Unklarheiten vermeiden.

Fragebogengestaltung: Fehler und Tipps zur Fehlervermeidung

Arbeiten mit einem vorgefertigten Datensatz

Dokumentation

Bei der Arbeit mit einem vorgefertigten Datensatz, wie zum Beispiel dem ALLBUS Datensatz, der auf fu:stat:thesis als Beispiel benutzt wird, ist im Allgemeinen als erstes die beigefügte Dokumentation zu Rate zu ziehen. Hieraus sollte man auf jeden Fall folgende Informationen gewinnen:

- Codierung von fehlenden Werten
- Skalenniveaus
- Wertebereiche
- Dimensionalität/Maßeinheiten
- Informationsverlust aufgrund von Anonymisierung und Vertraulichkeit.

Sollten diese Informationen nicht in der Dokumentation vorliegen, so sollten sie so gut wie möglich aus den Daten selbst erschlossen werden.

Einlesen des Datensatzes

Beim Einlesen des Datensatzes muss zuerst darauf geachtet werden, in welchem Dateiformat die Daten vorliegen. Dies lässt sich einfach an der Dateiendung erkennen und natürlich muss das Statistikprogramm dieses Dateiformat auch einlesen können. Wenn dies nicht möglich ist, muss das Dateiformat geändert werden. So können z.B. Excel-Tabellen auch im .csv Format abgespeichert werden, was nahezu universell lesbar ist.

Die meisten Statistikprogramme bringen einen "Wizard" zum Einlesen von Daten mit. Dieser erlaubt oftmals das Dateneinlesen komfortabel zu gestalten und eventuelle Optionen einfach zu ändern.

Überprüfung

Wenn der Datensatz eingelesen ist, sollte zuerst überprüft werden, ob das richtig geschehen ist. Ein Blick auf die erste Zeile zeigt zum Beispiel, ob die Kopfzeile, die die Variablennamen enthält, richtig erkannt worden ist, oder ob sich nur Variablennamen in den Einträgen befinden.

- Passen die eingelesenen Daten zur Dokumentation? Befindet sich Text in numerischen Variablen?
- Säubern des Datensatzes nach den Regeln im Paragrafen oben.

Falsche Einträge erkennen/Datenvalidierung

Nachdem der eigene Datensatz erstellt worden ist oder ein Datensatz eingelesen worden ist, suchen Sie nach möglichen falschen Einträgen, Übertragunsfehlern oder Messfehlern:

- Checken von möglichen Vertauschungen, z.B. 79 statt 97 während der Dateneingabe /Dateneinlesens.
- Überprüfen, ob alle numerische Variablen als numerisch richtig eingelesen sind, z.B. 0 (Null) wird O während der Dateneingabe.
- Überprüfen, ob die Bereichswerte korrekt für jede Variable eingetragen sind, z.B. Antworten außerhalb des Bereichs von möglichen Antworten, wie beispielsweise ein Alter von 1000.

Behandlung Ausreißer und Fehlende Werte

Nachdem alle obengenannte Fehler erfolgreich entdeckt worden sind, erfolgt eine erste explorative Analyse. Hier werden z.B. Ausreißer oder extreme Ausreißer identifiziert. Siehe hierfür den Artikel Identification of Outliers. Eine Möglichkeit für die Behandlung fehlender Werte wird im Wiki und hier detailliert diskutiert.

Große Datensätze/Big Data

Bei sehr großen Datensätzen bietet es sich an, die Daten erst mit einem Programm zu verarbeiten, das komplexe Suchabfragen und Datenbearbeitung für große Datenbanken beherrscht. Dies tun zum Beispiel die Programme MS Access und SQLite, die die Programmiersprache SQL benutzen. Der Wikipedia Artikel bietet einen guten Überblick über die einfachen Befehle, die sehr mächtige Datenabfragen erlauben. Eine Vielzahl von weiteren Tutorials ist im Internet frei verfügbar.

Außerdem bietet die FU-Berlin ein Cluster für High-Performance Computing an, das sich für AnwenderInnen mit hohem Speicher- und Rechenbedarf eignet.

Datenmanagement Statistikprogramme

Bei der Erhebung, Erfassung, Erstellung und Analyse eines Datensatzes werden verschiedene Programme und Statistikprogramme verwendet. Sie unterscheiden sich unter anderem darin, ob das Programm tabellenorientiert (e.g. Excel, Minitab) oder programmierorientiert (e.g. R, SAS, STATA) oder einfach ist, weil es eine Variablenansicht und einen Dateiansicht (e.g. SPSS, JMP) bietet.

Verfügbare Programme

Folgende Programme eignen sich zur Ersterstellung, Erhebung und Erfassung von Datensätzen:

- Excel, SPSS, Minitab, JMP, EpiData
- SQL/MS Access/SQLite/Oracle

Zur Bearbeitung und Analyse von Datensätzen werden folgende Programme häufig verwendet:

- R
- SPSS, SAS, STATA, JMP

Programmspezifische Probleme/Aspekte

- Fehlende Werte: z.B. in R, STATA und SPSS, "NA" oder "."
- Datumsprobleme in Excel. Excel erkennt Dezimalzahlen teils als Datumsangaben, dies hängt von den Ländereinstellungen des Betriebssystems ab.
- Keine Farben verwenden in Excel. Diese Informationen gehen beim Einlesen in die meisten Statistikprogramme verloren.
- Formeln in Excel, sowie Verweise auf andere Tabellen, werden eventuell nicht erkannt.