IMBEI
Institut für Medizinische Biometrie, Epidemiologie und Informatik

Obere Zahlbacher Straße 69
D-55131 Mainz
Tel.: 06131-173252
Fax: 06131 172968

Empfehlungen zur Struktur und Erfassung von Daten

Die folgenden kurzen Anmerkungen sollen als Empfehlung verstanden werden. Eine Nichtbeachtung kann jedoch bei der EDV-gestützten Dateneingabe sowohl zu merklichem Mehraufwand führen als auch die spätere Auswertung der Daten erheblich komplizieren. In den Unterlagen zu den SPSS-Kursen, die vom IMBEI regelmäßig angeboten werden, ist eine ausführliche Anleitung zum Umgang mit Daten in exemplarischer Form enthalten.

Datenstruktur

Die gängigen statistischen Auswertungsprogramme (u.a. auch die am IMBEI verfügbaren Programmpakete IBM SPSS Statistics® und SAS® setzen voraus, dass die zu verarbeitenden Rohdaten in einer "rechteckigen" Datenstruktur angeordnet sind. Im einfachsten Fall befinden sich die Beobachtungseinheiten (z.B. Patienten oder Tiere) in den Zeilen und in den Spalten die erhobenen Variablen (Merkmale) für jede Beobachtungseinheit. Die ersten Felder jeder Zeile sind üblicherweise solchen Variablen zugeordnet, mit denen sich die jeweilige Beobachtungseinheit identifizieren lässt. Wenn als Beobachtungseinheiten z.B. Patienten anzusehen sind, könnten dies etwa die Variablen "IDNR", "AGE", "SEX" usw. sein. Daran schließen sich die Felder an, in denen die Messungen weiterer Merkmale erfasst werden.

Werden einzelne Merkmale für jede Beobachtungseinheit zu zwei Zeitpunkten wiederholt erhoben, wie etwa bei Messung des diastolischen Blutdrucks unmittelbar vor und zwei Stunden nach Gabe eines Medikamentes, so muss für dieses Merkmal für jeden Messzeitpunkt ein eigenes Feld und somit eine gesonderte Variable zugeordnet werden, also etwa die Variablen "DBP1" bzw. "DBP2". Werden alle Merkmale pro Beobachtungseinheit mehr als einmal erhoben, bietet es sich an, diese Erhebungen zeilenweise zu erfassen. Dann werden zwei verschiedene Identifikations-Codes vergeben: einer für jede Beobachtungseinheit (z. B. die Patienten-ID) und einer pro Erhebung (z. B. die Nummer der Untersuchung). Je nach der Art der geplanten statistischen Auswertung ist abzuwägen, in welcher Form Messwiederholungen sinnvoll erfasst werden (Hilfestellung bietet die statistische Beratung).

IDNR	AGE	WEIGHT	DBP1	DBP2	...
971265	25	76,0	85	80	
975621	30	56,1	69	72	
964521	54	84,3	76	75	
...					

IDNR	TIME	AGE	WEIGHT	DBP	...
971265	1	25	76,0	80	
971265	2	25	76,0	78	
971265	3	25	76,0	78	
975621	1	30	56,1	72	
975621	2	30	56,1	74	
...					

Regeln für die Einrichtung von Variablennamen

- Möglichst kurze Variablennamen erzeugen
- Variablenname muss mit einem Buchstaben beginnen
- Variablenname darf außer Buchstaben nur die Ziffern 0 bis 9 und das Zeichen "_" enthalten
- Ansonsten keine Umlaute, kein **ß**, keine Sonderzeichen und keine math. Operatoren verwenden
- Variablenname darf auch keine Leerzeichen enthalten
- Variablenname darf nicht doppelt vorhanden sein

IMBEI
Institut für Medizinische Biometrie, Epidemiologie und Informatik

Obere Zahlbacher Straße 69
D-55131 Mainz
Tel.: 06131-173252
Fax: 06131 172968

Außerdem gilt

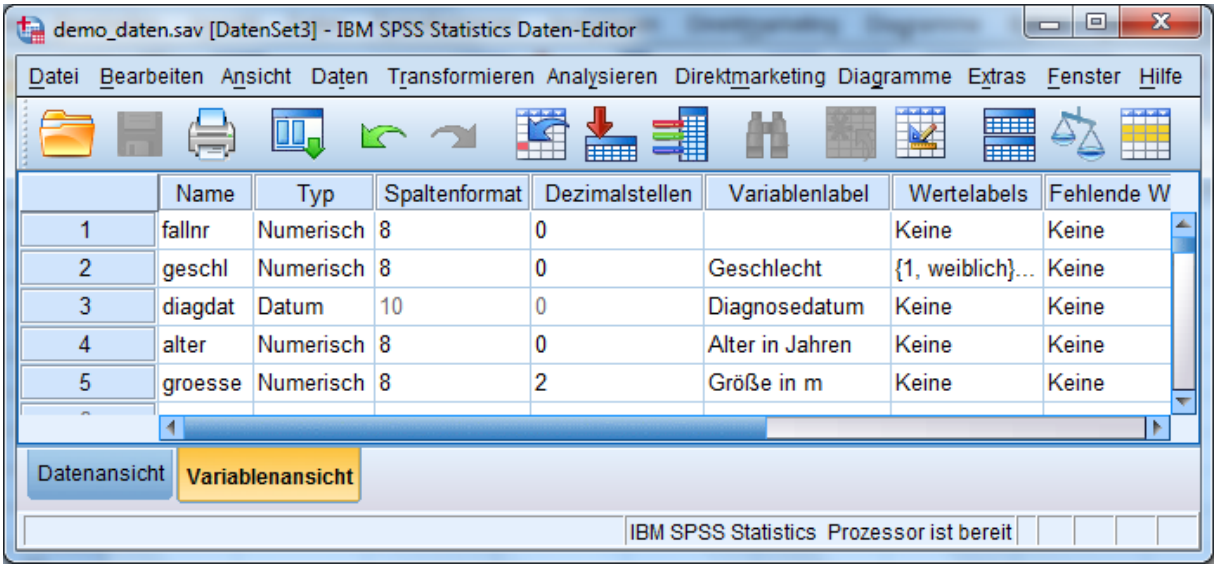
- Klartext ist nicht unmittelbar auswertbar und muss deshalb sinnvoll kodiert werden
- Kalenderdaten nicht als Textfelder (z. B. Juni 97) sondern als Datumsfelder definieren und eingeben
- Neben der eigentlichen Rohdatendatei ist eine vollständige Liste der Variablen und des jeweiligen Wertebereichs notwendig (sog. *"Datenbeschreibung"*)
- Personendaten müssen grundsätzlich anonymisiert werden (Namen durch ID-Nr. ersetzen)

Datenerfassung mit IBM SPSS Statistics®

Für eine Datenauswertung mit IBM SPSS Statistics® empfiehlt es sich, bereits die Rohdaten mit dem Datenblattwerkzeug von IBM SPSS Statistics® zu erfassen. In der Version 20 ist hierzu wie folgt vorzugehen.

Variablen definieren

Dies geschieht im **Daten-Editor**-Fenster, der zwei Ansichtfenster hat: ein Variablensichtfenster, wo zunächst die Variablen erzeugt werden und ein Datensichtfenster, wo danach die Werte einzugeben sind. Zuerst wird also das Variablensichtfenster benutzt. Dort sind zeilenweise die Varibalennamen einzutragen (Beispiel):



Nach Eintragen des Variablennamens kann jeweils schon mit RETURN bestätigt werden. Das System aktiviert dann die Vorbelegungen **numerisch** für den **Datentyp** und 8 für das **Spaltenformat**. Weitere Spezifikationen für eine Variable sind in der entprechenden Zelle vorzunehmen. Die Variable **diagdat** obigen Beispiels muss den Datentyp **Datum** erhalten, weil dort Tagesdaten einzugeben sein werden. Ein Linksklick in die Zelle **Typ** aktiviert den Auswahldialog **Variablentyp definieren**, wo eine entsprechende Auswahl (wie hier: **Datum**) zu markieren und mit **OK** zu bestätigen ist. Das Spaltenformat wird automatisch eingestellt und ist nur in Ausnahmefällen eigenständig zu ändern. Die Dezimalstellen können über den Zellendialog (Linksklick in Zelle **Dezimalstellen**) angepasst werden. Das **Variablenlabel** ist ein Klar-Name für die Variable und kann im normalen Textformat beschrieben werden. **Wertelabels** sind für die Textausgabe numerischer Verschlüsselungen in späteren Auswertungen vorgesehen (z. B. „weiblich“ für 1 und „männlich“ für 2).

IMBEI
Institut für Medizinische Biometrie, Epidemiologie und Informatik

Obere Zahlbacher Straße 69
D-55131 Mainz
Tel.: 06131-173252
Fax: 06131 172968

Regeln für die Daten-Eingabe

Nachkommastellen mit dem in Windows eingestellten Dezimaltrennzeichen abtrennen
Datums-Variablen immer im Format **TT.MM.JJJJ** eingeben (z.B.: 16.12.1998)
In jeder Zelle darf nur ein Variablenwert stehen!
Umgang mit Fehlwerten: Zelle unbedingt leer lassen!
Keine leeren Spalten und keine leeren Zeilen innerhalb der Tabelle!

Abspeichern der Daten

Speichern Sie die Daten unter einem selbstgewählten Namen in einem Ordner ab, den Sie auch wiederfinden. Die Dateinamenerweiterung **.sav** fügt IBM SPSS Statistics® automatisch hinzu.

Verwendung von MS-EXCEL® zur Datenerfassung

Prinzipiell ist die Erfassung von Rohdaten auch mit MS-EXCEL® möglich. Damit die Daten hinterher korrekt in ein statistisches Programmpaket übertragbar sind, müssen jedoch eine Reihe von Regeln beachtet werden. Zusätzlich zu den Regeln, die oben bezüglich der Variablennamen schon für IBM SPSS Statistics® angeführt sind, gilt vor allem:

- Alle Daten möglichst in eine solitäre Tabelle der Arbeitsmappe eingeben
- Felder mit numerischen Variablen dürfen nur Ziffern, Vorzeichen sowie Dezimaltrennzeichen enthalten
- Datums-Variablen müssen auch das Datums-Format haben: TT.MM.JJJJ
- Die Tabelle darf keine Blöcke mit neu berechneten Variablen enthalten
- Keine Spalten / keine Zeilen verbergen (nicht "ausblenden")!
- Keine überflüssigen Verschönerungen wie Farben, Rahmen etc. in die Tabelle einbringen

Das folgende Beispiel bildet eine Daten-Eingabetabelle in MS-EXCEL® ab:

A6 f_x

	A	B	C	D	E
1	fallnr	geschl	diagdat	alter	groesse
2	1	1	11.12.2011	57	1,76
3	2	2	04.11.2009	63	1,82
4	3	1	17.06.2006	58	1,68

Das Tabellenkalkulationprogramm wird also nur als Aufnahmeplattform für die später im Statistikpaket auszuwertenden Daten benutzt.

Die Daten-Tabelle des Tabellen-Kalkulationsprogramms kann in IBM SPSS Statistics® mit der Befehlsfolge **Datei, Öffnen** (und Dateityp *.xlsx eintellen) eingelesen werden, wenn alle technischen Voraussetzungen gegeben sind. Variablen-Labels und Werte-Labels (falls erforderlich) sind dann im Statistikprogramm zuzuweisen. Jedes der im IMBEI benutzten Programmpakte bietet diese Möglichkeiten.