

# Vorbereiten eines Datensatzes für die statistische Beratung

Abteilung Biostatistik  
Institut für Sozial- und Präventivmedizin  
Universität Zürich  
<http://www.biostat.uzh.ch>

## 1 Vorbereiten eines Datensatzes zur Analyse

In diesem Kapitel geben wir einige Richtlinien, wie Daten so erhoben werden, dass sie anschliessend mit gängigen Statistik-Programmen ausgewertet werden können.

**Datenschutz** Datenfiles, die z.B. dem Statistiker zur Auswertung zur Verfügung gestellt werden, sollten unter keinen Umständen vom Datenschutzgesetz geschützte Informationen enthalten. Insbesondere dürfen die Patienten in medizinischen Studien nicht identifizierbar sein, d.h. **Patientennamen müssen durch eindeutige Identifikationsnummern ersetzt werden**. Beachte: Oft geschieht die unerwünschte Übermittlung von Patientennamen in nicht gebrauchten und “vergessen gegangenen” Tabellenblättern in einem Excel-Dokument.

### 1.1 Datenformat

**Allgemeine Datenstruktur.** Die gängigen statistischen Auswertungsprogramme (u.a. R, SPSS, Stata, SAS) setzen voraus, dass die zu verarbeitenden Rohdaten “rechteckig” angeordnet sind. Darunter versteht man, dass in einer Tabelle die erhobenen Daten für jeden Fall (d.h. für die Beobachtungseinheiten wie z.B. Patienten) in genau der gleichen Abfolge und Zahl von Merkmalen aufgelistet sind. Die zu einem Fall gehörigen Angaben werden jeweils in einer Zeile zusammengefasst. Die Messungen eines bestimmten Merkmals werden in der gleichen Spalte untereinander für jeden Fall eingetragen. Eine so strukturierte Datei enthält demnach je Fall (=Beobachtungseinheit) eine Zeile und je Merkmal (=Variable) eine Spalte. Die ersten Felder jeder Zeile sind üblicherweise solchen Variablen zugeordnet, mit denen sich die jeweilige Beobachtungseinheit identifizieren lässt. Häufig ist dies eine einzige Variable, etwa eine Identifikationsnummer für die Patienten einer Stichprobe. Daran schliessen sich die Felder an, in denen die Werte weiterer Merkmale erfasst werden.

**Messwiederholungen.** Werden einzelne Merkmale für jede Beobachtungseinheit zu verschiedenen Zeitpunkten wiederholt erhoben, wie etwa Messungen des Gewichts bei Baseline und nach einer Behandlung, so muss für dieses Merkmal für jeden Messzeitpunkt eine eigene Spalte zugeordnet werden. Siehe dazu Tabelle 1.

PatNr	Geschlecht	Alter	Baseline	nach_Behandlung
1	f	45	57	62
2	m	36	65	64
3	m	51	70	65
⋮	⋮	⋮	⋮	

Tabelle 1: Datenstruktur, eine Messung pro Beobachtungseinheit.

Werden fast alle Merkmale pro Beobachtungseinheit mehr als einmal erhoben, so kann es je nach geplanter statistischer Auswertung sinnvoller sein, pro Beobachtungseinheit und Zeitpunkt eine eigene Zeile zu erfassen. In diesem Fall sind zwei verschiedene Identifikations-Kodes nötig: einer für jede Beobachtungseinheit (z.B. die Patienten-Nummer) und einer pro Erhebung (z.B. Zeitpunkt, die Nummer der Untersuchung). Folglich ist eine zusätzliche Variable zur Erkennung der verschiedenen Zeitpunkte nötig, siehe Tabelle 2.

PatNr	Zeitpunkt	Geschlecht	Gewicht
1	1	f	57
1	2	f	62
2	1	m	65
2	2	m	64
3	1	m	70
3	2	m	65
⋮	⋮	⋮	⋮

Tabelle 2: Datenstruktur, mehrere Messungen pro Beobachtungseinheit.

Welches Datenformat wann geeigneter ist, wird am besten vor der Datenerhebung festgelegt. Bedingung dafür ist natürlich, dass bereits dann bekannt ist, wie die Daten genau ausgewertet werden sollen. Dies sollte aber eigentlich immer der Fall sein.

**Weitere Anforderungen an die Daten.** Bei der Dateneingabe sind des weiteren zu beachten:

- Die Variablennamen müssen in der ersten Zeile stehen. Abgesehen von dieser ersten Zeile darf die Tabelle nur Datenwerte enthalten (keine Zwischenresultate, Grafiken, Formeln usw.).
- Variablennamen sollen eindeutig und möglichst kurz sein (maximal 16 Zeichen), sollten mit einem Buchstaben beginnen und dürfen keine Umlaute, keine Wortzwischenräume und keine Sonderzeichen (z.B. , %, #, -) ausser dem Tiefstrich ( \_ ) enthalten.
- Die erste Variable soll die eindeutige Fall-Kennung enthalten (z.B. PatNr, IDNR, UPN, SERNO).

- Felder mit numerischen Variablen dürfen nur Ziffern, das Vorzeichen + oder – sowie entweder Dezimalpunkt oder Dezimalkomma enthalten. Ob Zahlen korrekt formatiert sind, kann in Excel durch Setzen der Standardausrichtung festgestellt werden: Als Zahl formatierte Zahlen werden dabei rechtsbündig ausgerichtet, als Text formatierte dagegen linksbündig.
- Falls Variablenwerte nicht nur Ziffern, sondern auch alphanumerische Zeichen enthalten (also Zeichenketten wie z.B. T1a oder CO2), sollten Umlaute und Sonderzeichen vermieden werden, um eine einwandfreie Übertragung der Daten zu gewährleisten.
- Kalenderdaten sollten nicht als Textfelder (z.B. August 2000) eingegeben werden.
- Klartext ist unter keinen Umständen unmittelbar auswertbar und muss deshalb sinnvoll kodiert werden.
- Falls Werte kodiert werden, sollen die Codes numerisch sein (z.B. 1 für männlich, 0 für weiblich). Codes für die gleichen Antwortkategorien sollen für alle Variablen gleich sein, z.B. 0=nein, 1=ja. Geordnete Merkmale sollen aufsteigend (ordnungserhaltend) kodiert werden.
- Falls Werte fehlen, so sind die entsprechenden Felder am besten leer zu lassen. Werden fehlende Werte durch einen speziellen Code gekennzeichnet (z.B. 99999, NA, MV, etc.), muss dieser Code denselben Typ haben wie die betrachtete Variable (also z.B. “999” für eine numerische Variable und “NA” für eine Textvariable) und darf unter keinen Umständen als echte Beobachtung möglich sein.
- Idealerweise erstellt man zu einem Datenfile ein Beiblatt, das Bedeutung und Codierung der Variablen erklärt.
- Beim Import in ein Statistikprogramm gehen sämtliche Formatierungen (Farben, Schriftarten usw.) verloren. Wesentliche Informationen dürfen deshalb nie nur durch unterschiedliche Formatierungen dargestellt werden, sondern müssen immer aus dem unformatierten Inhalt der Zellen erkennbar sein. Gross- und Kleinschreibung wird ebenfalls nicht von allen Programmen unterschieden.

Dieses Dokument ist als Anhang B Bestandteil des Skripts *Einführung in die Biostatistik*. Dieses Skript kann von <http://www.biostat.uzh.ch/teaching/lecturenotes/scripts.html> heruntergeladen oder im Studentenladen der UZH gekauft werden.