Anleitung zur Datenerfassung in Excel

1. Einleitung

Datenbanken wie z.B. Access bieten umfangreiche Möglichkeiten bei der Datenerfassung, sind daher jedoch komplex in der Anwendung. Das einfachere Programm zur Erfassung von Daten ist Excel. Bei der Datenerfassung mit Excel sollten jedoch bestimmte Vorgehensweisen beachtet werden, um die anschließende Auswertung sowohl in Excel als auch bei einer Datenübernahme in andere Statistikprogramme zu erleichtern. Die folgenden Angaben für Excel beschreiben das Vorgehen bei der Datenerfassung.

2. Datenstruktur

2.1.Beobachtungseinheit und Merkmal

- Merkmale **spaltenweise**
- Beobachtungseinheiten (z.B. Patient, Maus, Muskelzelle) **zeilenweise**
  Jede Beobachtungseinheit erhält eine **eindeutige Identifikationsnummer** (ID), die in der ersten Spalte einer Tabelle eingetragen wird (hier: PatID)

Tabellengrundstruktur in Excel

	A	B	C	D	E	F
1	PatID	Gruppe	Alter	Geschlecht	Gewicht	Groesse
2	1	2	45	1	70	170
3	2	2	18	1	65	149
4	3	1	32	2	85	200
5	4	1	21	1	50	158
6	5	2	70	2	90	190

- *Datenblatt:*
  Mehrere Tabellen zu benutzen ist für die Auswertung mit Statistikprogrammen (wie z.B. SAS) kein Problem; getrennte Tabellenblätter für Demographie, Anamnese, Laborwerte, usw.
  Wichtig: Auf jedem Blatt muss für z.B. einen Patienten, **dieselbe ID** vergeben werden!
  Für eine Auswertung mit Excel sind mehrere Tabellenblätter nicht zu empfehlen.
  Es wäre daher im Allgemeinen wünschenswert alles in einer Tabelle zu erfassen.
- *Formatierung:*
  Komplett leere Spalten oder Zeilen, mehrzeilige Spaltenüberschriften, Farben und Rahmen vermeiden

  ➔ **Einfache Datenstruktur**

2.2. Datenschutz

Vertrauliche Patientendaten (z.B. Name, Adresse, Krankenkasse) dürfen aus Datenschutzgründen nicht in den Daten enthalten sein. Bei der Auswertung der anonymisierten Daten wird als Kennung für den Patienten die **eindeutige Identifikationsnummer (ID)** verwendet.

2.3. Datenstruktur bei Verlaufsdaten

Verlaufsdaten sind Beobachtungsmessungen, die zu verschiedenen Zeitpunkten entstehen. Diese können auf 2 Arten erfasst werden:

Datenstruktur 1

	A	B	C	D	E	F	G
1	PatID	Gruppe	Puls_Visite1	Puls_Visite2	Puls_Visite3	Schmerz_Visite1	Schmerz_Visite3
2	1	2	65	70	60	3	2
3	2	2	90	75	60	1	0

- Die Merkmale (Messungen zu verschiedenen Zeitpunkten) sind **nebeneinander** angeordnet.
- geeignet für **Vergleiche** (Gegenüberstellung: Schmerzen bei Aufnahme und Schmerzen bei Abschluss)

Datenstruktur 2

	A	B	C	D	E
1	PatID	Gruppe	Visite	Puls	Schmerz
2	1	2	1	65	3
3	1	2	2	70	
4	1	2	3	60	2
5	2	2	1	90	1
6	2	2	2	75	
7	2	2	3	60	0

- Die Daten (Messungen zu verschiedenen Zeitpunkten) stehen **untereinander**, im Bsp. 3 Zeilen pro Patient (eine Zeile für jede vorhandene Visite eines Patienten).
- lässt sich mit statistischen Programmen wie **SAS** gut auswerten
- geeignet für die **Verlaufsdarstellung** (v.a. grafisch)

3. Variablen

- Variablennamen stehen in der **ersten Zeile** der Tabelle
- die Namen sollten **kurz** und **aussagekräftig** sein
- möglichst **keine** Sonderzeichen, wie Umlaute oder Satzzeichen verwenden (sie bereiten evtl. Probleme bei der Datenübernahme in Statistikprogramme)
  Ausnahme: Unterstriche (underscores) dürfen benutzt werden, z.B. OP_Datum
- **nicht** mit einer Zahl beginnen (z.B. 1Visite)
- jeder Variablenname darf nur **einmal** verwendet werden, auch wenn dieselbe Variable in mehreren Tabellenblättern vorkommt.
  Ausnahme: die Identifikationsvariable muss in allen Tabellenblättern gleich lauten, z.B. PatID.

4. Dateneingabe und Kodierung

4.1 Allgemeines

- generell **immer** den Originalwert erfassen, ggf. sind Umrechnungen mit Excel oder SAS einfach durchführbar und weniger fehleranfällig (z.B. Gewicht und Körpergröße erfassen und BMI dann berechnen lassen)
- *Freitextangaben:*

  sind ungeeignet, weil durch unterschiedliche Schreibweisen verschiedene Kategorien entstehen
  - z.B. Geschlecht mit „m“ oder „M“ erfasst → ergibt 2 Kategorien
    → **Lösung:** Vorgegebene strukturierte Kodierungen zur Auswahl
  - Kommentare zu Dateneinträgen sind **nur** zulässig wenn sie in einer **extra Spalte** stehen

- *strukturierte Kodierung:*

  - *Beispiel:* für „nein“ eine 0 verwenden und für „ja“ eine 1
  - Kodierung darf **nur aus Zahlen** bestehen
  - **sinnvoll** vergeben, d.h. für die gleichen Antwortmöglichkeiten von verschiedenen Merkmalen, die gleichen Kodierungen verwenden und die Kodierung sollte immer aufsteigend sein
  - z.B. Merkmal Schmerzempfinden:

    keine Schmerzen= [0] leichte Schmerzen= [1]

    mittlere Schmerzen= [2] starke Schmerzen= [3]

- *Mehrfacheinträge in einer Tabellenzelle:*

  mehrere Angaben in einer Zelle/Spalte können **nicht** ausgewertet werden

  - *Beispiel:* Blutdruck (systolisch/diastolisch)

    [120/80] = nicht auswertbar → Einteilung in 2 Variablen: systolisch= [120] diastolisch= [80]

  - *Beispiel:* Maßeinheit bei Erythrozytenwert

    [4,0 µl] = nicht auswertbar → Eingabe in Variable: Erys = [4.0]

    (Angabe der Einheit bei der Variablenbezeichnung, s. 4.3)

- *Format:*

  Jede Spalte darf **nur** Angaben/Werte in **einem** Format enthalten:

  - numerisch (numeric): unbedingt ein einheitliches Dezimalzeichen verwenden
  - alphanumerisch (character): möglichst keine Umlaute und Sonderzeichen
  - Datum (date): einheitliches Datumsformat: TT.MM.JJJJ, z.B. 07.12.2011

- *Mehrfachantworten / Mehrfachauswahl / Mehrfachnennung:*

  Bei Fragestellungen mit Mehrfachauswahl soll **jede Antwortmöglichkeit** in eine **eigene Spalte** eingetragen werden, sonst können die Daten nicht richtig ausgewertet werden.

Beispiel: **Formulardarstellung**:

Begleiterkrankungen
(Mehrfachauswahl)
Asthma ☐
Bluthochdruck ☐
Diabetes ☐

Umsetzung in Excel:

Beispiel: Dateneingabe bei Mehrfachauswahl (jeweils mit nein/ja (0/1) kodiert)

	A	B	C	D	E
1	PatID	Gruppe	Begl_Asthma	Begl_Hochdruck	Begl_Diabetes
2	1	2	1	0	1
3	2	2	0	1	1
4	3	1	0	0	0
5	4	1	1	1	0
6	5	2	1	1	1

4.2 Annotated CRF (kommentierter Fragebogen)

Manchmal befinden sich die Angaben zur Kodierung (1=männlich, 2=weiblich) und evtl. auch der Variablenname [geschl] bereits im zugehörigen Formular/Fragebogen.

Beispiel:

Geschlecht: männlich ○ 1
[geschl] weiblich ○ 2

4.3. Variablenliste

In einem extra Tabellenblatt sollten **alle** Variablen mit ihren Ausprägungen **übersichtlich strukturiert** dargestellt werden:

Struktur des Tabellenblatts Variablenliste

	A	B	C	D	E
1	Tabelle	Variablenname	Variablenbezeichnung	Kodierung	Auspraegung
2	**PatDaten**	**PatID**	Patientennummer		
3	PatDaten	Gruppe	Testgruppe	1	Medikament
4	PatDaten			2	Placebo
5	PatDaten	Visite	Untersuchungen	1	Aufnahme
6	PatDaten			2	Zwischenuntersuchung
7	PatDaten			3	Abschluß
8	PatDaten	geschl	Geschlecht	1	männlich
9	PatDaten			2	weiblich
10	PatDaten	Schmerz	Schmerzempfinden	0	keine Schmerzen
11	PatDaten			1	leichte Schmerzen
12	PatDaten			2	mittlere Schmerzen
13	PatDaten			3	starke Schmerzen
14	PatDaten	Begl_Asthma	Asthma	0	nein
15	PatDaten			1	ja
16	PatDaten	Begl_Hochdruck	Bluthochdruck	0	nein
17	PatDaten			1	ja
18	PatDaten	Begl_Diabetes	Diabetes	0	nein
19	PatDaten			1	ja
20	PatDaten	Alter	Alter (Jahre)		
21	PatDaten	Gewicht	Gewicht (kg)		
22	PatDaten	Groesse	Größe (cm)		
23	**Labor**	**PatID**	Patientennummer		
24	Labor	Erys	Erythrozyten (µl)		
25	Labor	Hb	Hämoglobin (g/dl)		

➔ **Die Variablenliste ist für eine fehlerfreie Datenerfassung und Auswertung ein unentbehrliches Hilfsmittel.**

5. Fehlende Werte (Missings)

Fehlende Werte müssen korrekt erfasst werden, sonst kann es zu Problemen bei der Berechnung von statistischen Maßzahlen kommen.

- fehlende Werte dürfen **nie** mit einer **„0“** oder einem anderen **Zeichen** gekennzeichnet werden. Die „0“ ist trotz alledem eine **numerische Zahl** und wird mit in die Berechnung übernommen.

➔ **bei fehlenden Werten entsprechende Excel-Zellen unbedingt leer lassen**

Literatur

Hain, Johannes: Leitfaden zur Datenerfassung in Excel. Würzburg, Universität, Studienarbeit.
Online im Internet: URL: http://www.statistik-mathematik.uni-wuerzburg.de/fileadmin/10040800/user_upload/studberatung/excel_leitfaden.pdf

RRZN/Leibniz Universität Hannover: Excel 2010 Grundlagen. Hannover; v. Auflage 2 November 2010; erhältlich für Studierende an allen Hochschulen und Universitäten