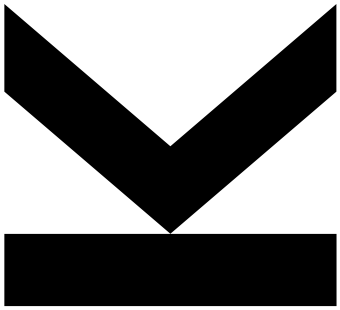
MEDIZINISCHE FAKULTÄT

Statistische Beratung und Datenanalyse

Kompetenzzentrum für Klinische Studien (KKS Linz)
am Zentrum für Klinische Forschung (ZKF)
Medizinische Fakultät

T +43 732 2468 3309
medstat@jku.at
jku.at/med/medstat

STATISTISCHE BERATUNG UND DATENANALYSE



Anforderungen an Datensätze

JOHANNES KEPLER UNIVERSITÄT LINZ
Altenberger Straße 69
4040 Linz, Österreich
jku.at
DVR 0093696

Um einen reibungslosen Ablauf von statistischen Analysen mit dem Team Statistik des KKS Linz zu gewährleisten, ist eine geeignete Struktur und Dokumentation der Datenerfassung und Datenaufbereitung wesentlich. Zur Einhaltung der Good Scientific Practice wird angeführt, welche Vorgaben an Datensätze einzuhalten sind, um eine effiziente Analyse der Daten mithilfe statistischer Software wie SPSS, R, u.a. zu ermöglichen. Viele der nachfolgenden Anforderungen beziehen sich im Speziellen auf eine Dateneingabe mithilfe von Microsoft Excel.

Die nachfolgenden Anforderungen sind Voraussetzungen für eine allfällige im Zuge der statistischen Beratungen und Datenanalyse als notwendig erachtete Übergabe von Datensätzen an das Team Statistik des KKS Linz zur statistischen Auswertung.

- Die Daten werden in Form einer Matrix in Excel eingegeben, eine Zeile entspricht einer Erhebungseinheit (z.B. Patient*in), eine Spalte entspricht einer Variablen. Für jede Variable ist eine Spalte vorzusehen (nicht: Datum und Grund für Ende der Therapie in einer Spalte).
- Die erste Zeile der Matrix muss die Variablennamen enthalten, ansonsten dürfen in der Tabelle nur Datenwerte stehen (keine Grafiken, Leerzeilen etc.).
- Die erste Variable (Spalte) enthält eine eindeutige Patientenkennung, z.B. Patientennummer.
- Informationen über Patient*innen wie Namen, Adressen, Telefonnummern, etc., dürfen nicht im File enthalten sein.
- Jede Zelle entspricht dem Datenwert einer/s bestimmten Patient*in (dessen Kennung in der 1. Spalte angegeben ist) in einer bestimmten Variablen (deren Namen in der ersten Zeile angegeben ist).
- Variablennamen dürfen nur die folgenden Zeichen enthalten:
    - Buchstaben A-Z
    - Zahlen 0-9
    - Punkt „.“, oder Underscore „_“ als Trennzeichen

  Das bedeutet, dass keine Umlaute, Leerzeichen und Sonderzeichen in Variablennamen erlaubt sind. Variablennamen müssen mit einem Buchstaben beginnen und sollen möglichst kurz (wenn möglich nicht länger als 20 Zeichen) sein.
- Jeder Variablenname darf nur einmal verwendet werden- Groß - und Kleinschreibung wird in vielen Statistik-Programmen nicht unterschieden, daher nur Groß- oder Kleinbuchstaben verwenden.
- Die Kodierung aller Variablen ist z.B. in einem eigenen Excel-Blatt festzuhalten. Dazu sind Spalten mit „Variablename“, „Inhalt“, „Kodierung“ und „Skalenniveau“ vorzusehen. Mustervorlagen für einen Codierbogen stellt das Team Statistik auf Anfrage gerne zur Verfügung.
    - Beispiel: sex; Geschlecht; 0=männlich, 1=weiblich, 2= divers, Skalenniveau: nominal
- Als Codes sind Zahlen zu verwenden, Codes für gleiche Antworten sollen für alle Variablen identisch sein, z.B. 0=nein, 1=ja.
- Bei Mehrfachantworten ist für jede Antwortmöglichkeit eine Spalte vorzusehen.
- Bei fehlenden Werten und nur bei diesen ist die entsprechende Zelle leer. Das heißt, wenn ein Wert nicht bekannt ist, ist die Zelle leer und es darf auch kein Blank (Leerzeichen) eingegeben werden. Für Variable mit Werten 0/1 (oder nein/ja) bedeutet eine leere Zelle nicht 0 (bzw. nein), sondern dass der Wert unbekannt (missing) ist.

- Weitere Codierungen für fehlende Werte wie z.B. 999, 9999 müssen ausdrücklich im Codierbogen festgehalten werden. Wir empfehlen jedoch diese Zellen leer zu halten.
- Die Zellen dürfen nur Werte jedoch keine Einheiten enthalten, daher muss die Maßeinheit einer Variablen für alle Einheiten die gleiche sein (nicht: einmal Monate, einmal Jahre).
- Die Zellen müssen das richtige Format haben, d.h. Zahlenformat für Zahlen, Datumsformat für Datumswerte. Datumswerte sind in der Form dd.mm.yyyy (z.B. 24.06.2008) einzugeben. Das Format einer Zelle kann mit Format->Zelle gewählt werden.
- Konsistente Verwendung von Kommazeichen im ganzen Datensatz. Entweder „,“ oder „.“ jedoch dürfen nicht beide Varianten verwendet werden, auch nicht bei verschiedenen Variablen. Wir empfehlen „,“ als Kommazeichen zu verwenden.
- Keine Tausender Trennzeichen (Blank oder „,“ oder „.“) verwenden.
- Alle Daten sind möglichst in einem Tabellenblatt anzuführen.
- Kommentare dürfen nicht in die Spalte mit dem Datenwert eingetragen werden. Wenn Kommentare notwendig sind, ist dafür eine eigene Spalte vorzusehen.
- Zellen dürfen keine Fußnoten, Färbungen etc. enthalten.
- Zur Berechnung des Alters werden sowohl die Geburtsdaten als auch das Datum, zu dem das Alter bestimmt werden soll (z.B. Diagnose, Start der Behandlung) benötigt. Die Berechnung des Alters erfolgt mit dem Statistikprogramm.
- In den Daten sollen keine Spalten enthalten sein, die aus anderen existierenden Spalten berechnet wurden, z.B. BMI aus Größe und Gewicht. Diese können durch Statistikprogramme fehlerfrei bestimmt werden.
- Bei Messwiederholungen bietet es sich an 2 Tabellenblätter zu erstellen. Dies tritt auf, wenn Daten zu verschiedenen Zeitpunkten erhoben werden. Ein Tabellenblatt enthält dabei die Informationen zum/zur z.B. Patient*in (Werte, die sich über die Zeit hinweg nicht verändern). Das andere enthält die Charakterisierung der wiederholten Messungen (z.B. Nummer der Messung: 1,2,.. oder Zeitpunkt: baseline, 1 Monat, …) und die Messwerte. In einer Zeile wird die Information zu einem Messzeitpunkt erfasst.
- Analog ist vorzugehen, wenn Messungen z.B. durch verschiedene Rater*innen durchgeführt werden: in diesem Fall ist eine Spalte zur Charakterisierung des/r Rater*in erforderlich und die Bewertung jedes/r Rater*in in einer Zeile zu erfassen.