

Organisatorisches:

Das vorliegende Dokument beschreibt, welche inhaltlichen und technischen Bedingungen klinische Datensätze erfüllen sollen, um möglichst problemlose statistische Analysen zu ermöglichen. Für die Analyse Ihrer Daten empfehlen wir Ihnen die Statistiksoftware IBM® SPSS® Statistics, die Sie als Angehörige/r der WWU über das ZIV kostenfrei nutzen können. Nähere Informationen finden Sie unter

<https://www.uni-muenster.de/ZIV/Software/SPSSAllgemeineInformationen.html>

Inhaltliche Bedingungen an Datensätze:

Aufbau des Datensatzes:

1. Stets Datum des Datenstandes vermerken, z.B. im Dateinamen
2. In der Regel eine Zeile pro Patient/Proband/Probe (Ausnahme: longitudinale Daten / Messwiederholungen, dann abklären) und eine Spalte pro Variable (manchmal auch „Item“ genannt)
3. Eindeutige ID in separater Spalte
 - **DON'T:** sich auf Zeilennummer (z.B. in Excel) verlassen. Immer ID-Spalte anlegen.
 - **DON'T:** Namen von Probanden / Akronym verwenden (die Daten müssen mindestens pseudonymisiert sein)
 - **DON'T:** die gleiche ID für mehrere Patienten (z.B. in verschiedenen Therapiegruppen) verwenden
4. ID muss in allen Dateien / Tabellenblättern eindeutig sein
5. Bei Erhebung in Excel nach Möglichkeit nicht zu viele Tabellenblätter anlegen, da beim Einlesen in SPSS stets nur ein einzelnes Tabellenblatt eingelesen wird.

Variablenamen / Itemnamen:

1. Variablenamen kurz und eindeutig, Sonderzeichen und Umlaute vermeiden, Ausnahme: _ (Unterstrich) ist erlaubt. Der Variablenname soll aber mit einem Buchstaben beginnen.
 - **DON'T:** Codierung in Variablenamen einbauen
2. Eindeutige Variablenamen verwenden: nicht den gleichen Namen mehrfach verwenden, Messwiederholungen geeignet durch Endung (z.B. _0, _1, _2 usw.) kennzeichnen
 - **DON'T:** Gleiche Variablen in unterschiedlichen Dateien / Tabellenblättern / Dateiversionen unterschiedlich benennen
 - **DON'T:** Unterschiedliche Variablen in unterschiedlichen Dateien / Tabellenblättern / Dateiversionen gleich benennen

3. Variablenlabels vergeben (SPSS): kurze, prägnante Beschreibung der Variablen, eindeutig und nicht zu lang

Plausibilitätschecks:

Es empfiehlt sich, die Daten vor Beginn der eigentlichen Analyse zu prüfen, um Fehler vor Beginn der Analysen zu korrigieren und so häufiges Wiederholen / Korrigieren der Analysen zu vermeiden.

1. Stetige Variablen: der Größe nach sortieren, größte und kleinste Werte prüfen (alternativ: Minimum, Maximum berechnen, Boxplot machen, ...)
2. Bei Umcodieren / Bilden von Kategorien: Kreuztabelle mit ursprünglicher Variable anlegen und Zuordnung prüfen
3. Variablen mit Datumsangaben: der Größe nach sortieren und prüfen, ob Werte zu weit in Vergangenheit / Zukunft liegen
4. Kategoriale Variablen (und evtl. auch metrische Variablen): Häufigkeitstabelle anlegen und prüfen. Ist die Anzahl fehlender Werte ok?
5. Falls Werte einer Variablen von anderer Variablen abhängen, dann Kreuztabelle anlegen und prüfen (Bsp. Raucher ja/nein vs. Anzahl Zigaretten pro Tag – Nichtraucher sollten einen fehlenden Wert oder eine Null aufweisen)

Technische Bedingungen an Datensätze:

Damit Daten von Statistikprogrammen adäquat eingelesen werden können, müssen diese gewisse technische Bedingungen erfüllen (siehe unten). Diese Bedingungen hängen teils auch von der Wahl der Statistiksoftware ab. Welche Statistiksoftware für Ihre Analysen erforderlich ist, stellt sich oftmals erst im fachlichen Beratungsgespräch mit dem Statistiker heraus.

In der Regel sind SPSS-Datensätze zu bevorzugen, da die Erstellung von SPSS-Datensätzen deutlich weniger Fehlermöglichkeiten bereithält als die Erstellung von Excel-Tabellen und SPSS-Datensätze von gängiger Statistiksoftware relativ problemlos eingelesen werden können. Wenn Sie die Möglichkeit haben, erstellen Sie also SPSS-Datensätze. Entweder indem Sie Ihre Daten direkt mittels SPSS in eine SPSS-Tabelle eingeben oder indem Sie eine Excel-Tabelle erstellen und diese adäquat in SPSS importieren. Dieser Ansatz erfordert, dass Sie über die kostenpflichtige Software SPSS verfügen (die Sie aber ggf. auch über das ZIV kostenfrei beziehen können, siehe oben). Bei Excel-Tabellen müssen Sie mit mehr Schwierigkeiten beim Einlesen rechnen, und damit mit mehr Aufwand bei der Nachbearbeitung / Korrektur Ihrer Datentabellen, die der Biometriker von Ihnen fordern wird.

Bei manchen speziellen Analysetechniken kann der Einsatz von Statistiksoftware erforderlich sein, die ein weiteres Spektrum als SPSS abdeckt. Die Statistiksoftware R erfüllt diese Bedingung. Bei R kann man sowohl Excel-Dateien, als auch SPSS-Dateien einlesen. Auch hier sind SPSS-Dateien zu bevorzugen.

Beschrieben werden 3 Fälle:

- A. Wie sollen Excel-Datensätze aufgebaut sein, damit diese in die Statistiksoftware R eingelesen werden können?
- B. Wie sollen Excel-Datensätze aufgebaut sein, damit diese in die Statistiksoftware SPSS eingelesen werden können, und damit auch als SPSS-Datensatz abgespeichert werden können?
- C. Wie sollen SPSS-Datensätze aussehen? Diese können auch in R und anderer Statistiksoftware wie z. B. SAS verwendet werden.

A. Excel-Datensätze für die Software „R“:

1. Die Variablennamen ...
 - müssen in der ersten Zeile des Datenblatts stehen.
 - müssen eindeutig sein.
 - dürfen keine Umlaute, Sonderzeichen (außer Unterstriche) und Leerzeichen enthalten und müssen immer mit einem Buchstaben beginnen. Am besten einfach nur Buchstaben, Zahlen und Unterstriche verwenden.
2. Direkt unter die Variablennamen kommen die Tabelleneinträge.
3. Bei den (Tabellen-)Einträgen sind Datumsangaben, numerische Werte und Zeichenfolgen/Strings erlaubt:
 - Datumsangaben sind in Excel als Datumsformate zu formatieren (über rechte Maustaste → Zellen formatieren ...).
 - Numerische Werte / Zahlen sollten auch als Zahlen formatiert werden (über rechte Maustaste auf Spalte klicken → Zellen formatieren ...). Achten Sie darauf, dass die Einträge danach auch in Excel richtig dargestellt werden.
 - Bei numerischen Angaben sind Einträge wie „20 mg“ nicht erlaubt. Als Eintrag nimmt man 20, und aus der Variablenbeschriftung sollte hervorgehen, dass es sich um Angaben in mg handelt.
 - Numerische Dezimalwerte sind bei deutschen Excel-Versionen mit einem Komma zu schreiben (also beispielsweise 20,5 und nicht 20.5).
 - Zeichenfolgen sind nur im Sinne kategorialer Variablen erlaubt, also beispielsweise „maennlich“ und „weiblich“ um das Geschlecht abzutragen. Uneinheitliche Beschreibungen können ggf. eingelesen, aber nicht analysiert werden: z. B. „maennlich“ und „maennlich “, also unterschiedlich wegen eines Leerzeichens, oder „maennlich“ „Maennlich“ wegen uneinheitlicher Schreibweise (Statistiksoftwareprogramme sind häufig case sensitive, unterscheiden also zwischen Groß- und Kleinschreibung).
4. Bei den Einträgen sind Umlaute, Sonderzeichen usw. nicht erlaubt (also nicht zu verwenden: „männlich“).
5. In den Einträgen sind keine Kombinationen erlaubt wie „Dosis 20 mg + Radiotherapie“. So etwas gehört in separate Spalten / Variablen, außer es wird damit ein Therapiearm oder Ähnliches beschrieben. (Bemerkung: Das Sonderzeichen „+“ ist dabei aber nicht gestattet.)
6. Hinterlegte Formeln sind nicht erlaubt.
7. Hinterlegte Farben, grafische Anpassungen usw. sind nicht erlaubt.
8. Kommentartexte sollten vermieden werden.
9. Die Excel-Tabelle muss abschließend als csv-Datei (Trennzeichen getrennt) abgespeichert werden. Bitte überprüfen, ob man die erstellte Datei auch in Excel geöffnet richtig angezeigt bekommt („richtig“ bedeutet u. a. als Tabelle mit mehreren Spalten, also nicht alle Einträge in einer Spalte).

B. Excel-Datensätze für die Software „SPSS“:

1. Die Variablennamen ...
 - müssen in der ersten Zeile des Datenblatts stehen.
 - müssen eindeutig sein.
 - dürfen keine Umlaute, Sonderzeichen (außer Unterstriche) und Leerzeichen enthalten und müssen immer mit einem Buchstaben beginnen. Am besten einfach nur Buchstaben, Zahlen und Unterstriche verwenden.
2. Direkt unter die Variablennamen kommen die Tabelleneinträge.
3. Bei den (Tabellen-)Einträgen sind Datumsangaben und numerische Werte erlaubt.
Zeichenfolgen/Strings sind nicht erlaubt.
 - Datumsangaben sind in Excel als Datumsformate zu formatieren (über rechte Maustaste → Zellen formatieren ...).
 - Numerische Werte / Zahlen sollten auch als Zahlen formatiert werden (über rechte Maustaste auf Spalte klicken → Zellen formatieren ...). Achten Sie darauf, dass die Einträge danach auch in Excel richtig dargestellt werden.
 - Bei numerischen Angaben sind Einträge wie „20 mg“ nicht erlaubt. Als Eintrag nimmt man 20, und aus der Variablenbeschriftung sollte hervorgehen, dass es sich um Angaben in mg handelt.
 - Kategoriale Variablen werden ebenfalls mit numerischen Werten eingetragen (0, 1, 2, ...). Der Hintergrund ist, dass nach dem Einlesen in SPSS die Kategorienbezeichnungen wieder hinterlegt werden müssen (z. B. wenn 1 für „female“ steht, und 2 für „male“). Bei einer Variablen mit dem Inhalt „nein / ja“, „liegt nicht vor / liegt vor“, empfiehlt sich die Codierung 0 (nein) / 1 (ja). Das geht in SPSS über „Werte“ oder „value labels“ (s.u.)
 - Zeichenfolgen sind nur als Ausnahme im Sinne von Anmerkungen erlaubt, die in die statistische Analyse aber nicht einfließen können.
4. In den Einträgen sind keine Kombinationen erlaubt wie „Dosis 20 mg + Radiotherapie“. So etwas gehört in separate Spalten / Variablen, außer es wird damit ein Therapiearm oder Ähnliches beschrieben.
5. Hinterlegte Formeln sind nicht erlaubt.
6. Hinterlegte Farben, grafische Anpassungen usw. sind nicht erlaubt.
7. Kommentartexte sollten vermieden werden.
8. Die Excel Tabellen können xlsx-Format haben oder als csv-Datei (Trennzeichen getrennt) abgespeichert werden.
9. Lesen Sie nun die Excel-Tabelle in SPSS ein (Datei → Öffnen ...).
10. Variablenbeschriftungen und Wertelabels / Kodierungen werden hinterlegt. Das geht in SPSS über „Werte“ oder „value labels“..., siehe Anleitung unter <https://www.youtube.com/watch?v=sSoOY99XqZ4>.
11. Speichern Sie den entsprechenden SPSS-Datensatz ab. Der SPSS-Datensatz (mit Endung .sav) kann nun für die Analyse mittels SPSS verwendet werden.

C. SPSS-Datensätze:

1. Bei den Variablennamen (bei SPSS schlicht „Name“) sind Umlaute, Sonderzeichen (außer Unterstriche) und Leerzeichen nicht gestattet und sie beginnen immer mit einem Buchstaben. Am besten einfach nur Buchstaben, Zahlen und Unterstriche verwenden.
2. Bei den (Tabellen-)Einträgen sind Datumsangaben und numerische Werte erlaubt. Zeichenfolgen/Strings werden zwar vom Programm auch akzeptiert, können in der Regel aber nicht in die statistischen Analysen einfließen.
3. Kategoriale Variablen werden numerisch eingetragen und mit den entsprechenden Wertelabels / Kodierungen hinterlegt (über „Werte“ oder „value labels“..., siehe Anleitung unter <https://www.youtube.com/watch?v=sSoOY99XqZ4>).
4. Beim Hinterlegen von Wertelabels sind Umlaute und Sonderzeichen nicht erlaubt.
5. Bei kategorialen Variablen muss jede Kategorie mit einem Wertelabel hinterlegt werden. Wenn es beispielsweise bei einer Variablen die Kategorien 1, 2, 3 gibt, welche hinterlegt werden mit „Schweregrad1“, „Schweregrad2“, „Schweregrad3“, und dann noch die Kategorie 99 hinzukommt (um beispielsweise zu kennzeichnen, dass der entsprechende Wert in den Akten fehlt), dann ist auch die Kategorie 99 beispielsweise mit „missing“ zu hinterlegen (über „Fehlende Werte“...). Ansonsten gibt es Schwierigkeiten beim Einlesen der Daten.
6. Das Hinterlegen von Kategorien, die in den Daten gar nicht vorkommen, sollte möglichst vermieden werden.