Einführungskurs in SPSS

Susanne Heim
<susanne.heim@stat.uni-muenchen.de>

Christian Heumann
Raum 339, Di 13:00 - 14:00
<chris@stat.uni-muenchen.de>

SS 2009

Organisatorisches

- Termine: 9 - 17 Uhr (**nach Absprache**); 6.-9. April
- Theorie- und Übungsblöcke im Wechsel
- Datensätze, Folien und Übungsblätter im Internet unter http://www.statistik.lmu.de/∼chris/spss/homepage/
- Teilnahmebestätigung ohne Leistungsnachweis
- 3 ETCS bei Abgabe der Hausarbeit bis zum 15. Mai 2009

IuK-Pool

- Kennung für Studierende der Statistik/BWL/VWL/Wipäd gegen Vorlage des Studenten– und Personalausweises in der WiWi-Bibliothek
- Übungsdateien im Ordner `Trainingslaufwerk` / `k108` sind lokal zu kopieren
- langfristiges Speichern ist nur auf eigenem Netzlaufwerk möglich
- Laptops dürfen nicht ans Netzwerk angeschlossen werden; WLAN ist verfügbar
- Essen und Trinken ist nicht gestattet (Kamera!)

Inhalt

- Anmerkungen zum Programmpaket SPSS
- Datenmanagement (Eingabe und -bearbeitung, Verschmelzen und Aggregieren von Datendateien)
- Syntaxbefehle und Variablentransformation
- deskriptive, explorative und induktive Datenanalyse
- Kreuztabellen, Mittelwertvergleiche, t-Test, Korrelationsanalyse, Regressionsanalyse, Faktoren- und Clusteranalyse
- graphische Verfahren

SPSS: Allgemeines

- Entwicklung 1968 im akademischen Umfeld

  ➔ Statistical Package for Social Sciences

- Universitäten, private Wirtschaft, öffentliche Verwaltungen

- Soziologie, Psychologie, BWL/VWL, Biologie, Medizin, u. a.

- weltweit am meisten verbreitetes Statistikprogramm

  ➔ Superior Performing Software Systems
  ➔ Statistical Product and Service Solutions
  ➔ heute nunmehr das Akronym

Es gibt einen Statistik Blog.
Informationsseite für Studenten:
http://www.spss.com/de/vertical_markets/academia.htm

SPSS: Versionen

- Version 17.0 für Windows XP oder Windows Vista, Mac OS X und Linux

- Kategoriale Regression (Base), multinomiale logistische oder nichtlineare Regression (Regression), GLMs/GEEs (Advanced Statistics); mehrere geöffnete Datensätze innerhalb einer Sitzung ($>$ 14.0); Lesen/Schreiben des Stata-Formats, Ergebnisse als PDF exportierbar, Plugin für R Code und Python
  What's new? http://www.spss.com/statistics/changes.htm

- Demo-Version für 30 Tage von der Homepage:
  http://www.spss.com/de/downl.cfm

- Campuslizenz beim LRZ erhältlich: EUR 50,- + EUR 5,- Datenträger; Gültigkeit: 01.04.2009 - 31.03.2010

SPSS: Vor- und Nachteile

– kommerziell, d. h. jüngere statistische Methoden i. A. nicht verfügbar

\+ Windows-konform, erweiterter Anwenderkreis mit UNIX/Linux Version

\+ komfortable GUI, anwenderfreundlich

+/– 'intuitiv' bedienbar

+/– publikumsreifer Output

– Graphiken nur begrenzt automatisch manipulierbar

\+ automatische Erstellung von Programmcode (Syntax)

– unhandliches Programmiertool: wenig entwicklerfreundlich

– weniger statistische Verfahren als andere kommerzielle Software

\+ umfassendes Hilfesystem

SPSS Struktur: Basissystem mit Erweiterungsmodulen

SPSS Base	SPSS Decision Trees
SPSS Advanced Statistics	SPSS Conjoint
SPSS Regression	SPSS Custom Tables
SPSS Exact Tests	SPSS Complex Samples
SPSS Categories	SPSS Data Preparation
SPSS Forecasting	SPSS Text Analysis for Surveys
SPSS Programmability Extension	SPSS Missing Values
SPSS Neural Networks	SPSS EZ RFM

Zusatzprodukte: Sample Power, SPSS Data Entry, Amos, . . .

SPSS: Fenster und Dateien

- Daten–Editor
  - Daten–/Variablenansicht (Wechsel mit `Ctrl + T`)
  - mehrere gleichzeitig geöffnete Datensätze ($>$ 14.0)
  - Schließen des letzten Datenfensters beendet SPSS
  - `<data>.sav` ($\rightarrow$ to `save`)
- Viewer/Ausgabe
  - Output–Navigator
  - Objekte bearbeitbar, z. B. Tabellen formatieren
  - `<output>.spo` ($\rightarrow$ `SPSS` object)
- Syntax–Editor
  - `<syntax>.sps` ($\rightarrow$ `SPSS` syntax)

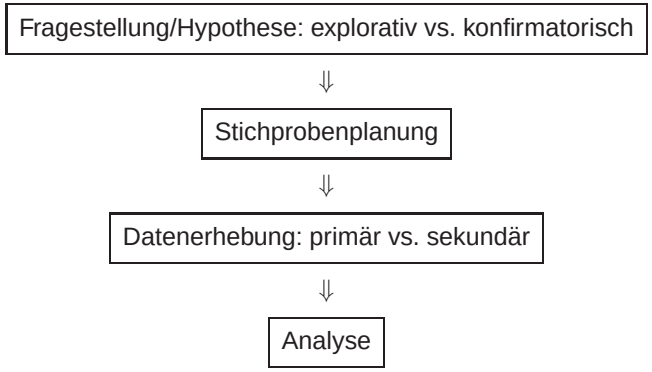
`Alt + Tab` schneller Wechsel des aktiven Fensters

SPSS: Hilfesystem

- Menüpunkt `Hilfe > Themen` mit üblichen Registrierkarten `Inhalt`, `Index` und `Suchen`. Zusätzlich: `Favoriten` z.B. Keyboard Shortcuts
- kontextbezogene Hilfe bei Prozeduraufruf übers Menü
- Ausgabefenster: Doppelklick auf Tabellenkopf
- Syntaxfenster: Hilfe zu markiertem Befehl mittels speziellem Icon oder über `Hilfe > Befehlssyntax`
- Menüpunkt `Hilfe > Algorithmen` für z. B. Teststatistiken
- `Hilfe > [Lernprogramm, Fallstudien, Statistics Coach]`
- Handbücher, wenig geeignet zum Einstieg in die Kommandosprache

- `Hilfe > SPSS Developer Central`
- USENET Newsgroup `comp.soft-sys.stat.spss`

statistische Untersuchung/Studie



Datenstruktur

SPSS: Fälle, Untersuchungseinheiten

- Menschen: Fragebögen, Messungen
- Unternehmen
- Objekte: Messungen

SPSS: Variablen, Untersuchungsmerkmale

- demographisch: Alter, Geschlecht, Familienstand, Schulabschluss
- klinisch: Blutzucker, Hormonkonzentration, Gewicht
- soziologisch: Freizeitaktivität, Mediennutzung, Kaufverhalten
- wirtschaftlich: Umsatz, Gewinn vor Steuer
- technisch: Materialeigenschaften, experimentelles Design

SPSS: Datenmatrix, Relationale Datenstruktur

- Zeile = Untersuchungseinheit,
- Spalte = Untersuchungsmerkmal

Variablenarten

- stetige Variablen (Alter, Gewicht): reelle Zahlen
- ordinale Variablen (Schulnoten, Klassifizierungen): größer/kleiner

⇒ Merkmalsausprägungen werden Zahlen zugeordnet

- nominale Variablen (Geschlecht, Farben): gleich/ungleich

⇒ Merkmalsausprägungen werden willkürlich Zahlen zugewiesen

- Datumsangaben
- Text

Die Skala legt die erlaubten Operationen fest.

⇒ Variablenkodierung, d. h. zulässige Skalenabbildung
(Dummykodierung, numerische Repräsentation der Antwort)

Variablendefinition

Variablenansicht des Daten-Editors erlaubt Spezifikation von 10 Attributen

- Name der Benutzervariablen (siehe Hilfe):
  - max. 64 Zeichen
  - erstes Zeichen Buchstabe (no Case-Sensitivity)
  - beliebig Ziffern, Buchstaben, Sonderzeichen (`@`, `#`, `$`, `.`, `_`), keine Leerzeichen, Vermeidung von Umlauten und "ß"
  - nicht `.` oder `_` als Endung
  - keine SPSS-Syntax wie `ALL`, `AND`, `BY`, `EQ`, `GE`, `GT`, `LE`, `LT`, `NE`, `NOT`, `OR`, `TO`, `WITH`

Hilfsvariablen:	temporär gültig, nicht speicherbar/auswertbar, 'hash' als erstes Zeichen (z. B. `#aux1`)
Systemvariablen:	`$DATE`, `$TIME`, `$CASENUM`, `$SYSMIS`

Variablendefinition

- Typ (Numerisch, String, Datum, Scientific, ...)
- Format (Dezimalstellen, Spaltenformat)
- Labels für Namen und Werte
- Fehlende Werte (systemdefiniert, benutzerdefiniert)
- Spalten und Ausrichtung
- Messniveau (metrisch, ordinal, nominal)

Fehlende Werte

– gehen nicht in die Auswertungen ein

systemdefiniert:

- numerische Variablen:
  - keine Zahl oder Leerzeichen liefert ein `$SYSMIS`
  - Darstellung als `'.'`
- Stringvariablen: Interpretation als leerer String

benutzerdefiniert: Deklaration in der Variablenansicht

- Differenzierung der Ursachen eines fehlenden Wertes,
  etwa keine Antwort geben wollen oder können, da nichts zutraf.
  ➔ "Ersatzwert" definieren, z. B. für Einkommen:
  '-9' Verweigerung, '-99' Nicht angetroffen
- fehlender Wert vom Typ String:
  'KA' kodiert für 'Keine Angabe'
  'ka' wird als String gelesen, z. B. PKW–Kennzeichen Karlsruhe.

Kleine Studie im Kurs

- ca. 15 Teilnehmer
- 5 Variablen: Alter, Geschlecht, Gewicht, Größe, Salatkonsum

Geburtsdatum	_____ dd-mm-jjjj
Geschlecht	☐ männlich ☐ weiblich
Gewicht	_____ kg
Körpergröße	_____ cm
Ich esse Salat	_____ – mal pro Woche.

- Fragestellungen/Hypothesen:
  - Gibt es einen Zusammenhang zw. Geschlecht und Salatkonsum?
  - Wiegen Salatliebhaber signifikant weniger als Salatverweigerer?
  - Welchen Einfluss hat der Salatkonsum auf das Gewicht in Abhängigkeit vom Geschlecht und der Körpergröße?

Übung 1

1. Füllen Sie den ausgeteilten Fragebogen aus.

2. Erzeugen Sie eine SPSS–Datei zur Eingabe der Fragebögen und definieren Sie sinnvolle Variablenattribute.

3. Geben Sie 5 Fragebögen ein (ausgefüllte Fragebögen austauschen).

4. Generieren Sie eine druckbare Variablenübersicht mittels `Datei > Datendatei > Information anzeigen > Arbeitsdatei` im Ausgabe-Fenster.

5. Speichern Sie die Datendatei in Ihrem Verzeichnis ab.

`ENTER` zeilenweises Füllen der Datenmatrix
`Tab` spaltenweise Füllen der Datenmatrix

Daten einlesen I

`Datei > Öffnen > Daten`

- dBase
- Lotus, Excel
- SQL Abfragen, ODBC
- Systat, SAS, Stata
- andere SPSS Versionen

`Datei > Textdaten lesen`

- unter Verwendung eines Wizards
- ASCII frei, fest
- Tab–delimited, Semikolon-delimited
- Format für zukünftige Verwendung als *.tpf–Datei speicherbar

Daten einlesen II

Excel-Format: `salat07_r.xls`

- von SPSS für Excel verwendete Kodierung für fehlenden Wert: #NULL!
- in Excel benutzerdefiniertes Format für Datumsangabe ist international: tt-mmm-jjjj
- Formatierung als europäische 'tt.mm.jjjj'–Angabe innerhalb Excel (`Format > Zellen`) oder SPSS (Variablenansicht)
- deutsche Version von Excel interpretiert Komma stets als Dezimalzeichen; Excel bietet über `Daten > Externe Daten importieren > Daten importieren...` einen Ausweg, indem Dezimalzeichen gesondert angegeben werden kann.

Daten einlesen III

"freies" ASCII: `salat07_r.dat, salat07_r.csv`

- disjunkte Spezifikation von Trennzeichen, Dezimalzeichen und fehlendem Wert
- übliche Trennzeichen: Tabulator, Komma, Semikolon, Leerzeichen
- Format per Editor kontrollieren beim Import/Export
- unsichtbare Steuer-/Formatierungszeichen: `<TAB>`, `<CR>`, `<SPACE>`
- leere Zelle kodiert vorzugsweise für fehlenden Wert
  $\Rightarrow$ Stringerkennung unproblematisch
- variable Feldbreite
- Variablennamen in erster Zeile möglich
- SPSS Datenimport Assistent
  z. B. 'mm/tt/jjjj'–Datumsformat für Variable Geburtsdatum spezifizieren; ändern in der SPSS Variablenansicht

Daten einlesen IV

"freies" ASCII

- Lesen von Dezimalzahlen abhängig von Einstellung im Betriebssystem; Änderung des Dezimalzeichens auf Windows XP unter `Systemsteuerung > Regions- und Sprachoptionen > Regionale Einstellungen > Anpassen > Zahlen`
- SPSS liest beim 1. Aufruf die Konventionen aus der Systemsteuerung

festes ASCII `salat_r.asc`, in R: fixed width format (*.fwf)

- feste Feldbreite, folglich keine Variablennamen in erster Zeile
- leere Zelle kodiert vorzugsweise fehlenden Wert

Übung 2

1. Öffnen Sie jeweils die xls-, dat-, csv- und asc-formatierten Datensätze `salat07_r` in SPSS. Welche Schwierigkeiten treten auf und wie können diese gelöst werden?

2. Erweitern Sie die Daten aus Übung 1 um eine Variable für den Studienbeginn. Übertragen Sie die Datumseigenschaften der Geburtstagsvariablen unter Verwendung des Dialogs `Data > Copy Data Properties ...`. Erfragen Sie die Daten für Ihre 5 Beispielfälle, tragen Sie diese in die Datenmatrix ein und speichern Sie erneut.

Datentransformation

- `Daten > Fälle sortieren`
  auf-/absteigend sortieren nach einer oder mehreren Variablen
- `Daten > Transponieren`
- `Daten > Fälle auswählen`
  filtern nach einer oder mehreren Variablen
- `Daten > Umstrukturieren` (über Assistenten)

Variablentransformation

- kopieren, verschieben (markieren und drag & drop), löschen
- Variable vervielfältigen/Attribute übertragen
- `Transformieren > Variable berechnen`
  für numerischen Ausdruck, u. a. verfügbare Funktionengruppen sind Datumsarithmetik und Zufallszahlen
- `Transformieren > Umkodieren` unter Nebenbedingungen möglich
- `Transformieren > Visuell Klassieren`
  einer ordinalen/metrischen Variable
- professionell: Rohdatendatei (`<daten>r.sav`) → Transformationsprogramm (`<daten>t.sps`) → Fertigdatendatei (`<daten>.sav`)

Übung 3: Variablentransformation

1. Lesen Sie den Datensatz `salat09.dat` ein und spezifizieren Sie die Variablenattribute soweit sinnvoll.
2. Identifizieren Sie anhand des Geburtstags, ob eventuell eine Fallverdopplung vorliegt. Verwenden Sie den zugehörigen Dialog `Daten > Doppelte Fälle ermitteln ...`. Machen Sie sich mit den Optionen des Dialogs vertraut. Welchen Effekt haben diese?
3. Fügen Sie eine Variable zur Fallidentifikation ein.
4. Berechnen Sie das Alter in Jahren unter Verwendung des Datumsassistenten. Erarbeiten Sie auch eine Lösung über Datumsarithmetik.
5. Generieren Sie eine Variable mit den Werten des Body-Mass-Index in der Einheit [kg/m^2].
6. Simulieren Sie jeweils das Körpergewicht nach zweimaliger Gewichts**zunahme** als weitere Variablen des Datensatzes.
   Speichern Sie den Datensatz.

Übung 4: Umstrukturierung

1. Organisieren Sie den in Übung 3 erweiterten Datensatz mit Hilfe des Umstrukturierungsassistenten, sodass die wiederholten Messwerte für Körpergewicht in einer einzigen Variablen erfasst werden und eine Hilfsvariable für die Messungen kodiert.
   Speichern Sie den Datensatz.

2. Strukturieren Sie den Datensatz aus Aufgabe 1 so, dass die Messwiederholungen wieder eigenständige Variablen darstellen.
   Hinweis: Falls Aufgabe 1 nicht gelöst wurde,
   laden Sie `salat07_messwh_umstrukturiert.sav`.