Eine Einführung in SPSS

Aufbau von SPSS 14

Bemerkung: SPSS 14 kann in den Subzentren in der Kopernikusgasse installiert werden, falls dies noch nicht geschehen ist. Dazu öffnet man den Application Explorer am Desktop, danach wählt man All und abschließend doppelklickt man auf dem SPSS Symbol.

Beim Starten von SPSS öffnet sich der Daten-Editor: Im Daten-Editor können Sie Daten eingeben und bearbeiten. Der Daten-Editor spiegelt die grundlegende Struktur einer SPSS-Datendatei wider. Jede Zeile stellt einen einzelnen Fall (eine Beobachtung) dar, jede Spalte beschreibt eine Variable.

So ist zum Beispiel jede Person in einer Umfrage ein Fall. Jede Frage ist demgegenüber eine Variable, und die Antwort der Person ist die Realisation der Variable für diese Person.

Im Daten-Editor können keine Berechnungen ausgeführt oder Formeln eingegeben werden – wie z.B. in einem Excel Spreadsheet. Verwenden Sie zum Berechnen neuer Werte und Variablen sowie zum Umkodieren von Daten das Menü "Transformieren".

Nachdem Daten in den Daten-Editor eingegeben oder aus einer externen Quelle importiert wurden, können sie einer statistischen Analyse unterzogen werden. Die Ergebnisse werden im Ausgabe-Navigator angezeigt. Der linke Ausschnitt des Navigators enthält eine Gliederungsansicht des Inhaltsfensters. Der rechte Ausschnitt enthält statistische Tabellen, Diagramme und Textausgabe. Der Ausgabe-Navigator verfügt die üblichen Copy-Paste Funktionalitäten.

Die über die graphische Benutzerschnittstelle zugänglichen Befehle entsprechen Befehlen in der SPSS Kommandosprache. SPSS Kommandos können direkt in den Syntax-Editor eingegeben und exekutiert werden (Produktionsmodus).

Grundlegend für die statistische Analyse ist die graphische Aufbereitung der Daten. Nach Generierung einer Graphik kann diese im Diagramm-Editor bearbeitet werden (Ausnahme: interaktive Plots).

Das Doppelklicken auf Objekte ist ein praktischer Shortcut, wenn Sie schnell auf viele Funktionen zum Bearbeiten von Diagrammen zugreifen möchten. Um zum Beispiel eine Diagrammbeschriftung zu ändern, doppelklicken Sie darauf. Dadurch wird das entsprechende Dialogfeld geöffnet.

Grundlegende Schritte der Datenanalyse mit SPSS:

Schritte	Menüs	Befehle
Dateneingabe: • Import von Daten • Eingabe mittels Daten-Editor • Datenexport	**Datei** **Bearbeiten**	Neu Öffnen Datenbank einlesen (Database Capture) Ascii-Daten einlesen (Read Text Data) Speichern unter
Bearbeitung: • Definieren • Sortieren • Transformieren	**Daten** **Transformieren**	Definieren Berechnen (Compute) Umkodieren (Recode)
Datenanalyse: • Statistische Analyse • Graphische Analyse	**Analysieren** **Grafiken**	Zusammenfassen (Analysieren→Deskriptive Statistiken→Explorative Datenanalyse; Reports→Case Summaries, Decriptive Statistics→ Explore) Tabellen (Tables, Custom Tables) Mittelwerte vergleichen (Compare Means) Balken (Bar) Kreis (Pie) Boxplot (Box) Streudiagramm (Scatter) Histogramm QQ-, PP-Plot Interaktive Plots
Prüfung der Ergebnisse: • Methodische Voraussetzungen • Interpretation • Iteration der Datenanalyse • Aufbereitung		

1. Einlesen von Daten

- Kleinere Datensätze können direkt mittels Dateneditor eingegeben werden. Der SPSS Dateneditor gleicht einem gewöhnlichen Spreadsheet. Bei der Eingabe ist darauf zu achten, dass die Zeilen stets Beobachtungen (Objekte) und die Spalten stets Variable beinhalten. Besteht der Datensatz beispielsweise aus Messungen des Gewichts und der Körpergröße von 100 Personen, dann trägt man die Größe und das Gewicht der i-ten Person in die ersten beiden Spalten der i-ten Zeile ein. Es resultiert eine (100 x 2)-Datenmatrix. Falls die 100 Personen in 50 Männer und 50 Frauen aufgeteilt sind, dann wird dies durch eine Indikatorvariable beschrieben, welche z.B. 0 für Männer und 1 für Frauen ist.

- Größere Datensätze werden häufig in SPSS importiert. Dafür stehen im Menü Datei drei Befehle zur Verfügung:

- **Öffnen**: Öffnen von standardisierten Files. Keine Einflussmöglichkeit durch den Benützer. Hauptsächlich für SPSS-Files, aber auch Excel, Lotus, Dbase und Tabulatoren-getrennte Ascii-Files.
- **Datenbank öffnen**: flexible ODBC Schnittstelle zu den gängigsten Datenbanken, z.B. Access, Excel 5.
- **Textdaten einlesen**: Daten liegen in Ascii-Form vor. Dieser Befehl wird zum Einlesen von Textdaten verwendet. Im Gegensatz zu früheren Versionen von SPSS wird man beim Einlesen durch einen Wizard unterstützt – der natürlich einiges wissen möchte:
  1. Gibt es ein vordefiniertes (predefined) Format für dieses File: ja/nein
  2. Fixe Breite (fixed width) oder Trennzeichen (delimited)? (Die erste Option bedeutet, dass die Variablenwerte in immer der gleichen, dafür vorgesehen Spalte stehen, die zweite, dass sie durch bestimmte Trennzeichen voneinander abgeteilt werden.)
  3. Variablennamen in der Kopfzeile usw.?
  4. Mit der Maus können bei fixer Breite die Spalten explizit durch Klicken angegeben werden.
  5. Vergeben von Variablennamen und Typen.

Beispiel: **Einlesen der Aimu Daten**.

Die **Datendefinition**:

Variablen:
1 ... Identifikationsnummer
2 ... Jahr der Erfassung
3 ... Alter in Jahren
4 ... Größe in cm
5 ... Gewicht in kg
6 ... VC in CL
7 ... FEV1 in CL
8 ... FEV1/VC in Prozent
9 ... Region (Faktor mit Stufen: A-Aichfeld; M-Murau)

Anzahl der Datenpunkte n=79.

Das Datenformat ist Ascii, mit fixer Breite der Variablenspalten. Wir verwenden zum Einlesen den Befehl „Textdaten einlesen“.

Schritte:
1. Angeben der einzulesenden Datei
2. Fragen des Wizards: Vordefiniert=No, fixe Breite=Yes, Variablennamen=No
3. Alle Fälle auswählen (Daten ab 1. Zeile, je Fall eine Zeile)
4. Angeben der Spalten durch Klicken
5. Eventuell Namen eingeben und Datenformat ändern
6. OK oder Umleitung in Syntax-Editor – dann sieht man die Kommando Syntax von SPSS für diesen Befehl

Wenn wir die Daten abspeichern möchten, dann gehen wir im Menü Datei auf „Speichern unter“. Die zur Verfügung stehenden Formate sind SPSS-, Excel-, Textformat u.a.

Bearbeitung der Daten:

Vergabe von geeigneten Namen und Formaten:
Schritte:

1. Auf Registerkarte Variablenansicht klicken
2. Vergabe des Namens
3. Typ angeben
4. Spaltenformat festlegen
5. Variablenlabel definieren, falls gewünscht (ausführlichere Bezeichnung der Variable).
6. Wertelabel angeben (z.B. Geschlecht ist mit 1/0 kodiert, dann kann in dieser Spalte angegeben werden, dass 1=weiblich, 0=männlich bedeutet.)
7. Messniveau festlegen
8. Restliche Spalten überprüfen

(Namen im Beispiel: „id“, „jahr“, „alter“, „gr_cm“, „ge_kg“, „fvc“, „fe“, „fvcfe“, „region“)

Alternativ kommt man durch zweimaliges Klicken auf den Spaltenkopf automatisch in die Variablenansicht.

Transformieren der Daten:

Sehr häufig steht man vor der Aufgabe, neue Datenwerte auf der Grundlage von numerischen Transformationen bestehender Variablen zu berechnen – zum Beispiel, wenn man den Logarithmus einer Variable analysieren möchte oder wenn man die Variable in andere Maßeinheiten umrechnen möchte.

Schritte:

1. Wählen Sie aus dem Menü "Transformieren" den Befehl "Berechnen" aus.
2. Geben Sie den Namen der Zielvariablen ein.
3. Geben Sie den numerischen Ausdruck ein: Sie können Variablen aus der Quelliste einfügen, Zahlen und Operatoren mit der Rechnertastatur eingeben und Funktionen aus der Funktionsliste einfügen.
4. „OK“. Die neue Variable wird am Ende der Datendatei eingefügt.

Umrechnen von fvc in fvc_l=fvc/100.

1. Transformieren
2. Berechnen
3. Zielvariable angeben
4. Im gegenüberliegenden Fenster die Berechung angeben – z.B. mit Hilfe des Rechners
5. „OK“

Die Schaltfläche „Einfügen“ („Paste“) erlaubt das Umleiten in den Syntax-Editor und liefert

```
COMPUTE fvc_l = fvc/100 .
EXECUTE .
```

(Die Verwendung des Editors empfiehlt sich bei aufwendigen Datenmanipulationen, die mehrfach durchgeführt werden sollen.)

Wir möchten nun die Daten nach den Werten einer stetigen Variable (z.B. „ge_kg“) in Klassen aufteilen, um danach eine Zielgröße in den einzelnen Gewichtsklassen untersuchen zu können. Zu diesem Zweck wird eine Variable „ge_kl“ wie folgt berechnet: „ge_kl“ ist gleich 1, 2, 3, 4 für Werte von „ge_kg“ zwischen 54-70, 71-78, 79-84,85-102.

Schritte:
1. „Transformieren“
2. „Umkodieren“
3. „In andere Variablen“
4. Variable „ge_kg“ selektieren
5. Name der Ausgabevariable auf „ge_kl“ setzen, „Zuweisen“ anklicken
6. „Alte und neue Werte“ anklicken
7. Bereich eingeben, neuen Wert angeben
8. 7. für alle Bereiche wiederholen
9. „Weiter“
10. „OK“

Anmerkung: Manche Statistische Prozeduren in SPSS sind nur möglich, wenn die Variablen numerisch kodiert sind. Bei der Variable „region“ ist dies z.B. nicht der Fall: „A“ steht für Aichfeld und „M“ für Murau – diese Variablen werden von SPSS als Zeichenketten interpretiert. Deshalb ist es zielführend, diese Variable in einen so genannten Faktor umzukodieren.

Schritte:
1. „Transformieren“
2. „Automatisch Umkodieren“
3. Variable „region“ selektieren
4. Name der neuen Variable auf „re_fak“ setzen, „Neuer Name“ anklicken
5. „Neuen Namen hinzufügen“
6. „OK“

Graphische Datenanalyse:

Wählen Sie aus dem Menü "Grafiken" den gewünschten Diagrammtyp aus.

Beispiel: Boxplot für Alter nach Region:
Schritte:
1. „Grafiken“.
2. „Boxplot“
3. Variante ist „einfach“, „nach Kategorien einer Variable“
4. „Definieren“
5. Als Variable „alter“ selektieren
6. „re_fak“ für die Kategorienachse selektieren
7. „OK“.

Nach dem Erstellen eines Diagramms können Sie dessen Darstellung durch Bearbeiten einer Vielzahl von Attributen verändern. Doppelklicken Sie auf dem

Diagramm, um den Diagramm-Editor zu öffnen. Sie können Titel, Beschriftung, Schriftarten oder Farben ändern, den Bereich der Skalenachse ändern, Achsen vertauschen und vieles mehr.
Viele statistische Verfahren beruhen auf der Normalverteilungsannahme. Daher müssen Methoden zur Verfügung stehen, die uns bei der Beurteilung dieser Voraussetzung unterstützen. Eine Möglichkeit besteht in der Anfertigung von QQ-Plots, bei denen die Quantile der Stichprobe gegen die Quantile einer theoretischen Verteilung aufgetragen werden. Wir wollen dies für „fvc_l“ tun.

Schritte:
1. „Grafiken“
2. „Q-Q“
3. „fvc_l“ auswählen
4. Verteilungsparameter aus den Daten schätzen lassen
5. „OK“

Im resultierenden Plot sollten die Punkte um eine Gerade (zufällig) streuen. Systematische Abweichungen deuten auf Verletzung der Verteilungsannahme hin. Im trendbereinigten Plot sollten die Punkte zufällig um die horizontale Achse streuen.

Aufgabe 1: Experimentieren Sie mit dem QQ-Plot:
1. Stellen Sie die Parameter der Normalverteilung selbst ein.
2. Verwenden Sie andere Referenzverteilungen.
3. Verwenden Sie andere Variablen. (Beispielsweise „alter“ – Vergleich mit Gleichverteilung. Stutzen Sie die Daten (mittels Befehl „Fälle auswählen“ in „Daten“) so, dass die Anpassung besser wird.)
4. Interpretieren Sie Ihre Resultate. Was bedeuten die Abweichungen im Einzelnen.

Das Fehlerbalkendiagramm:

Die graphische Analyse hat grundsätzlich eher explorativen Charakter. Nichstdestotrotz kommen auch hier bereits inferenzstatistische Erkenntnisse zum Einsatz. Anhand von Boxplots und Fehlerbalken ist bereits eine inferenzstatistische Beurteilung der Daten möglich.

Wir wollen nun Folgendes tun:
1. Die Daten nach der neuen Variable „gr_cm“ aufteilen. Wir bilden dazu eine neue Variable „gr_kl“ mit den Werten 1, 2, 3 und 4, falls „gr_cm“ zwischen 160-172, 173-176, 177-181, 182-195 liegt.
2. Analysieren Sie die Daten mittels Fehlerbalken: Beim Fehlerbalkendiagramm werden vom Mittelwert Balken aufgetragen, wobei deren Länge optional (a) ein Vielfaches der Standardabweichung, (b) des Standardfehlers des Mittelwerts bzw. (c) ein Konfidenzintervall sein kann.

Die Scatterplotmatrix (Matrix-Streudiagramm):
Möchte man die Zusammenhänge zwischen mehreren Variablen untersuchen bietet sich die Scatterplotmatrix als graphische Analysemethode an. Untersuchen Sie den Zusammenhang zwischen „fvc_l“, „alter“, „gr_cm“ und „ge_kg“ mittels Grafiken→Streu-/Punktdiagramm. Zeichnen Sie in den Graphen Ausgleichskurven ein (linear, quadratisch, loess) (Diagramm-Editor öffnen – dort unter Elemente – Anpassungslinie bei Gesamtwert).

Aufgabe 2: Führen Sie eine Analyse der AIMU Daten mittels Fehlerbalken und Scatterplotmatrizen durch. Interpretieren Sie ihre Resultate.

Bemerkung: Für gewisse Aufgaben eignen sich die interaktiven Plots besser (etwa 3-dimensionale Scatterplots, kumulative Histogramme).

Statistische Analyse:

Zusammenfassung, Häufigkeiten, Kennzahlen

Wenn Sie eine statistische Prozedur ausführen möchten, wählen Sie "Analysieren" aus der Menüleiste aus. Bestimmen Sie dann die „Kategorie“ der Prozedur (z.B. „Deskriptive Statistiken“) und schließlich den gewünschten Befehl („Häufigkeiten“).

Nach Auswahl der Variablen, die in die Analyse mit einbezogen werden sollen, und dem Setzen der gewünschten Optionen, wird mit „OK“ die Berechnung gestartet. Die Ergebnisse werden im Ausgabe-Navigator angezeigt.

Beispiel: Untersuchung der Häufigkeiten der Variable „alter“.
Schritte:
1. Menü „Analysieren“
2. Kategorie „Deskriptive Statistik“
3. Kommando „Häufigkeiten“
4. Auswahl einer Variable
5. Auswahl von zu berechnenden Statistiken
6. Auswahl von auszugebenden Diagrammen

SPSS Syntax erhält man durch Klicken auf die Schaltfläche „Einfügen“ im entsprechenden
Dialog und sieht in diesem Fall wie folgt aus:

```
FREQUENCIES
  VARIABLES=alter
  /NTILES=  4
  /NTILES= 10
/STATISTICS=STDDEV VARIANCE RANGE MINIMUM MAXIMUM SEMEAN MEAN MEDIAN MODE
  SUM SKEWNESS SESKW KURTOSIS SEKURT
  /HISTOGRAM  NORMAL .
```

Untersucht wird die Variable „alter“. Dabei werden Quartile und Perzentile berechnet. Zusätzliche werden die Statistiken Standardabweichung, Varianz, Spannweite, Minimum, Maximum, Mittelwert, Median, Modalwert, Summe, Schiefe, Kurtosis (mit jeweiligen Standardfehlern) ermittelt. Ein Histogramm wird gezeichnet und darüber die Normalverteilung als Referenzverteilung gelegt.

Vergleich von Mittelwerten:

Nach dem graphischen Vergleich von Mittelwerten in einzelnen Gruppen, kommen nun klassische inferenzstatistische Verfahren zum Einsatz. Dabei werden Hypothesentests durchgeführt und Konfidenzintervalle berechnet.

T-Test für eine Stichprobe:

Test der Hypothese, dass der Mittelwert von unabhängigen und nomalverteilten Zufallsvariablen mit einem angenommenen Wert übereinstimmt. Die Alternativhypothese umfasst Abweichungen in beide Richtungen. Dabei wird die Varianz als nicht bekannt angesehen.

Schritte:
1. „Analysieren“
2. „Mittelwerte vergleichen“
3. „T-Test bei einer Stichprobe“

Es soll getestet werden, ob der Mittelwert von „fvc“ gleich 500, 530, 536, 537, 560, 570, 571 ist (alpha=0.05). Wie lautet das 95%-Konfidenzintervall für den Mittelwert? Welcher Zusammenhang besteht zwischen dem Test und dem Konfidenzintervall für den Mittelwert von „fvc“?

Statistik bei einer Stichprobe

	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
FVC	79	553.4937	76.2704	8.5811

Konfidenzintervall (KI): mw +/- t(0.975,78) se(mw)

t(0.975,78) = 1.9908
KI = (536.41,570.57)

Test bei einer Sichprobe

	Testwert = 500					
					95% Konfidenzintervall der Differenz	
	T	df	Sig. (2-seitig)	Mittlere Differenz	Untere	Obere
FVC	6.234	78	.000	53.4937	36.4100	70.5773

KI für mw : +500 +500
KI = (536.41,570.57)

Nach Durchführung der Tests kommt man bezüglich des Zusammenhanges vom Konfidenzintervall und dem zweiseitigen Hypothesentest zu folgendem Schluss: Das 95%-Konfidenzintervall beinhaltet alle Werte mu_0, die bei dem zweiseitigen Test der Hypothese $mu=mu_0$ zum Niveau alpha=0.05 nicht zur Ablehnung führen. Umgekehrt kann ein Test der Nullhypothese $mu=mu_0$ zum Niveau Alpha durchgeführt werden, indem man prüft, ob der Testwert im (1-Alpha)*100% Konfidenzintervall für mu liegt.

Signifikanz- oder P-Wert:

Bei vielen Programmpaketen – wie auch dem vorliegenden – wird der Signifikanz-Wert (oder P-Wert) angegeben, der eine alternative Darstellung der Entscheidungsregel eines Tests liefert. Der Signifikanz-Wert ist definiert als

P(x1,…,xn) = kleinstes Signifikanzniveau, bei dem H0 gerade noch verworfen wird. Bei einem vorgegebenen alpha lautet die Entscheidungsregel

- P(x1,…,xn) <= alpha, dann wird H0 verworfen.
- P(x1,…,xn) > alpha, dann wird H0 beibehalten.

Der Signfikanzwert beim zweiseitigen T-Test errechnet sich nach der Formel P(x1,…,xn) = P(t) = 2(1-F(|t|)), wobei t die Realisation der (standardisierten) Teststatistik im T-Test ist und F die Verteilungsfunktion der t-Verteilung mit n-1 Freiheitsgraden.

Für die einseitige Fragestellung mit den Nullhypothesen mu<=mu0, resp. mu>=mu0 errechnet sich der P-Wert nach den Formel 1-F(t) resp. F(t). Bei SPSS ist nur der zweiseitige Signifikanzwert angegeben, der dann dementsprechend umgerechnet werden muss, falls ein einseitiges Testproblem vorliegt:

Das folgende Beispiel mag das illustrieren. Wir testen die Hypothese: Erwartungswert mu von „fvc“ = 538 zum Niveau alpha=0.05. Es resultiert:

Statistik bei einer Stichprobe

	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
FVC	79	553.4937	76.2704	8.5811

Test bei einer Sichprobe

	Testwert = 538					
					95% Konfidenzintervall der Differenz	
	T	df	Sig. (2-seitig)	Mittlere Differenz	Untere	Obere
FVC	1.806	78	.075	15.4937	-1.5900	32.5773

Das Signifikanzniveau des zweiseitigen Tests beträgt 0.075, somit kann die Nullhypothese zum Niveau 0.05 nicht abgelehnt werden. Der Test der Hypothese mu<=538 ergibt einen Signifikanzwert von 0.075/2=0.037. Dieser Wert führt dazu, dass die Hypothese zum Niveau von 0.05 zu verwerfen wäre. Die Hypothese mu>=538 dagegen liefert einen Signifikanzwert von 0.963. Das kann man natürlich keinesfalls ablehnen.

Interessanter ist der **T-Test für unabhängige Stichproben:**

Hierbei werden die Mittelwerte zweier Stichproben verglichen. Die Nullhypothese behauptet die Gleichheit der Mittelwerte. Die Varianz ist dabei nicht bekannt, aber in beiden Gruppen gleich. Falls letzteres nicht erfüllt ist, wird ein approximativer Test verwendet.

Schritte:

1. „Analysieren“
2. „Mittelwerte vergleichen“
3. „T-Test bei unabhängigen Stichproben“

4. Gruppierungsvariable eingeben
5. Unter Optionen das Konfidenzniveau für die Konfidenzintervalle festlegen

Definieren Sie die Variable „klein_groß“ mit Werten 1, 2 für „gr_cm“ zwischen 150-173, und 174-195. Untersuchen Sie ob ein Unterschied zwischen dem mittleren „fvc“-Werten in den beiden Gruppen besteht (Niveau = 5%).

Gruppenstatistiken

	klein_gross	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
fvc	1,00	28	497,9643	57,50780	10,86795
	2,00	51	583,9804	67,92746	9,51175

(95%-Konfidenzintervalle für die beiden Mittelwerte?)

Test bei unabhängigen Stichproben

		Levene-Test der Varianzgleichheit		T-Test für die Mittelwertgleichheit						
									95% Konfidenzintervall der Differenz	
		F	Signifikanz	T	df	Sig. (2-seitig)	Mittlere Differenz	Standardfehler der Differenz	Untere	Obere
fvc	Varianzen sind gleich	,757	,387	-5,673	77	,000	-86,01611	15,16279	-116,209	-55,82314
	Varianzen sind nicht gleich			-5,956	63,945	,000	-86,01611	14,44250	-114,869	-57,16343

Anstatt des F-Tests wird von vorneherein der Levene Test verwendet, um auf Gleichheit der Varianzen zu testen. Die Argumentation bleibt die gleiche. Zu einem Signifikanzniveau von alpha=0.05 kann die Hypothese der Gleichheit der Varianzen nicht abgelehnt werden. Der Wert der Teststatistik T beläuft sich auf –5.651 bei 77 Freiheitsgraden – ist also hochsignifikant. Die mittlere Differenz ist –89.2250 mit einem Standardfehler von 15.7886, woraus ein 95%-Konfidenzintervall für die mittlere Differenz von (-120.66, -57.78) resultiert.

Bei verbundenen oder gepaarten Stichproben kommt der **T-Test bei gepaarten Stichproben** zum Einsatz.