TECHNISCHE UNIVERSITÄT
CHEMNITZ

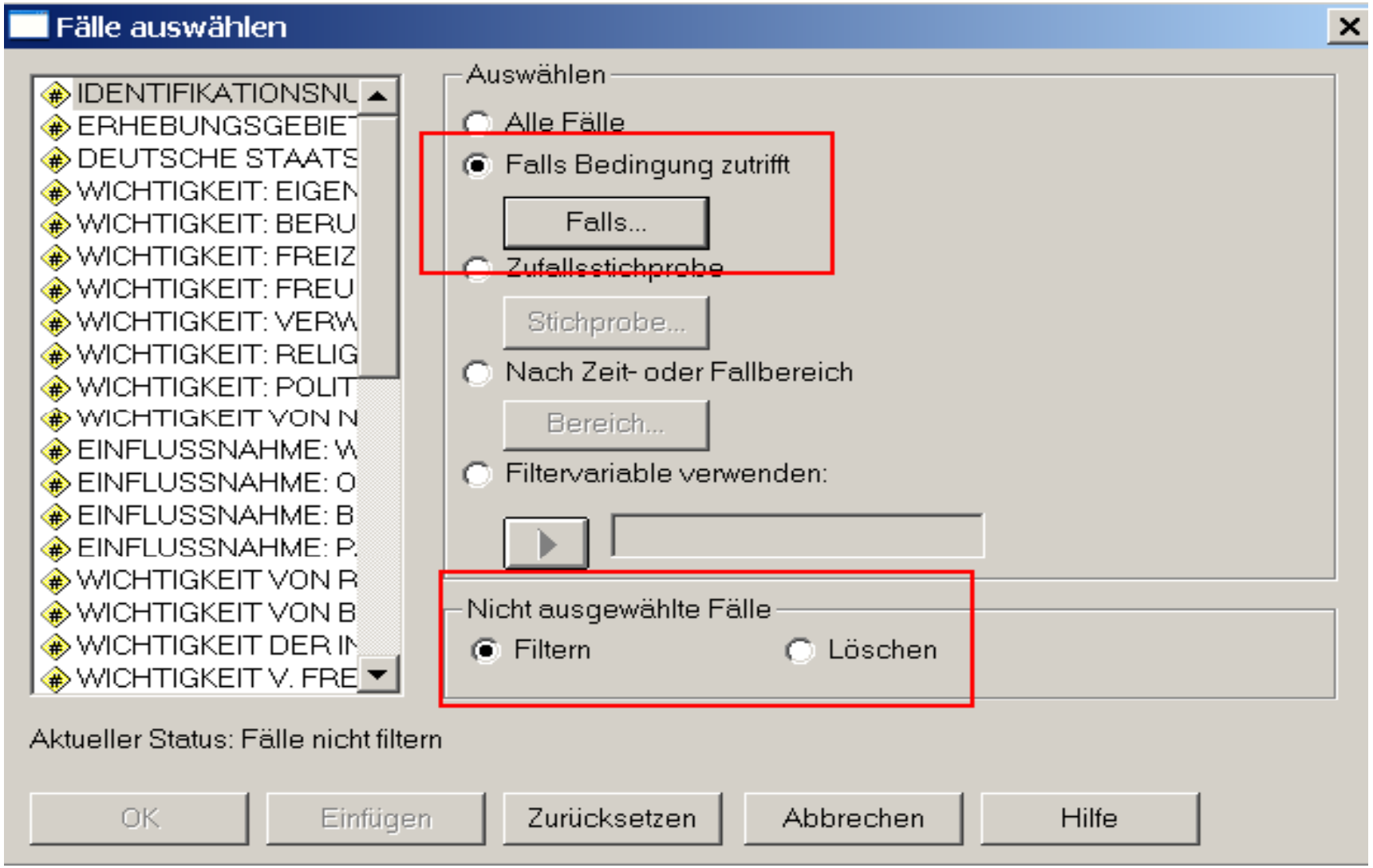
Institut für Soziologie
Professur für empirische Sozialforschung
Prof. Dr. Johannes Kopp
Dr. Daniel Lois

Datenhandling in SPSS

Stand: Mai 2010

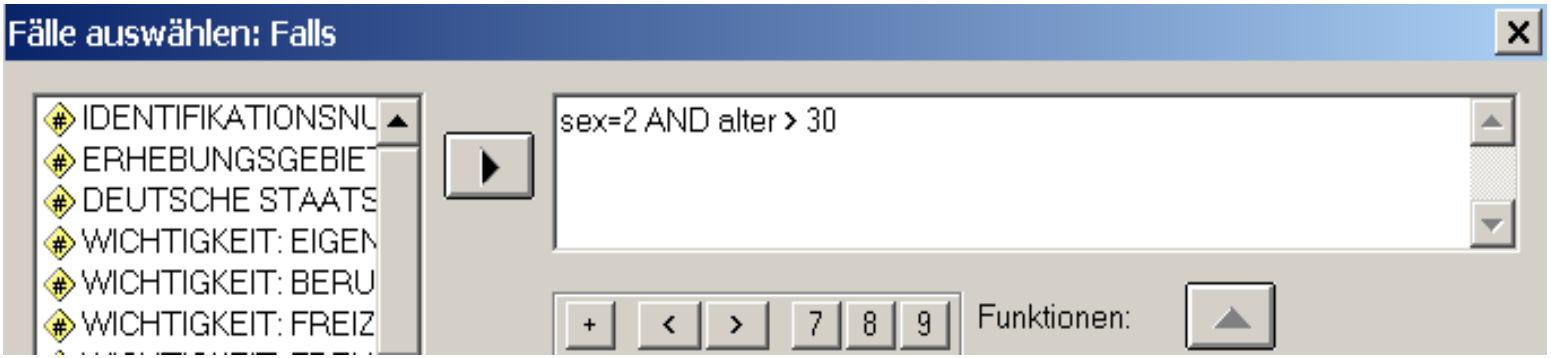
Inhaltsverzeichnis

Fälle auswählen



Fälle auswählen

- Ein häufig vorkommender Anwendungsfall besteht darin, für bestimmte Berechnungen nur eine ausgewählte Gruppe von Fällen zuzulassen.
- Zum Beispiel können nur die Fälle ausgewählt werden, die eine bestimmte *Bedingung* erfüllen. Wie solche Bedingungen formuliert werden, zeigt die nächste Folie
- Fallauswahl in SPSS im Menü *Daten – Fälle auswählen – Falls Bedingung zutrifft* oder mit dem Syntax-Befehl **SELECT IF**
- Sollen z.B. nur über 30jährige Frauen berücksichtigt werden, lautet die Bedingung:



Fälle auswählen

Vergleichs-operator	Bedeutung
=	gleich
~=	ungleich
<	kleiner
<=	kleiner oder gleich
>	größer
>=	größer oder gleich

- Bei den Bedingungen handelt es sich um logische Ausdrücke, bei denen verschiedene Vergleichsoperatoren verwendet werden (siehe Tabelle).
- Zwei (oder auch mehr) Vergleiche können durch die logischen Operatoren AND bzw. OR verknüpft werden (z.B. IF var1=1 AND var2=1). Ein logischer Ausdruck, der aus einer AND-Verbindung zweier Vergleiche besteht, ist genau dann „wahr“, wenn beide Vergleiche wahr sind. Wenn z.B. var1=1 ist, var2 dagegen ungleich 1, ist die formulierte Bedingung nicht gegeben (wahr).
- Ein logischer Ausdruck, der aus einer OR-Verbindung zweier Vergleiche besteht (z.B. IF var1=1 OR var2=1) ist genau dann wahr, wenn mindestens einer der Vergleiche wahr ist.

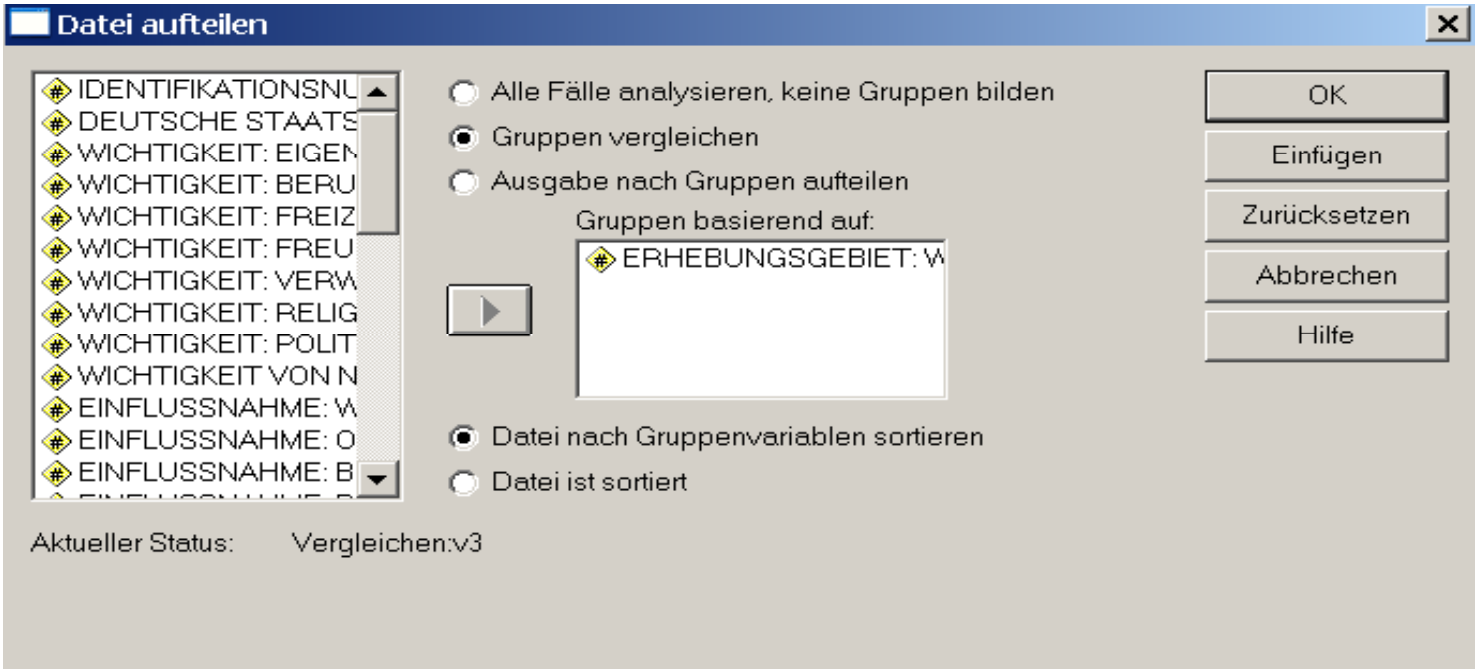
Fälle auswählen

Prozedur	Syntax
Fälle filtern: Frauen (sex = 2)	COMPUTE filter_$ = (sex = 2). FILTER BY filter_$.
Ostdeutsche (v5 = 2), die eine Konfession haben (konf = 1)	COMPUTE filter_$ = (v5 = 2 AND konf = 1). FILTER BY filter_$.
Ostdeutsche (v5 = 2) im Alter von 18-40 Jahren	COMPUTE filter_$ = (v5 = 2 AND RANGE (alter,18,40)). FILTER BY filter_$.
Ostdeutsche (v5 = 2) oder Westdeutsche älter als 40 Jahre	COMPUTE filter_$ = (v5 = 2 OR (v5=1 AND alter > 40)). FILTER BY filter_$.
Filter deaktivieren	FILTER OFF. USE ALL.
Nicht ausgewählte Fälle (hier: Männer) aus Datensatz löschen	SELECT IF sex = 2.

Datei aufteilen

- Außerdem besteht im Menü *Daten – Fälle auswählen* noch die Möglichkeit, eine Zufallsstichprobe aus den vorhandenen Fällen zu ziehen (z.B. 10% der Fälle auszuwählen)
- oder die Auswahl nach einer *Filtervariablen* zu richten, wobei alle Fälle ausgewählt werden, die bei dieser Filtervariablen einen anderen Wert als 0 oder fehlend haben
- Eine weitere Möglichkeit besteht darin, die Datendatei für Auswertungen in ein oder mehrere Fallgruppen aufzuteilen (*Daten – Datei aufteilen*)
- Im Folgenden Beispiel wurde die Datei z.B. nach dem Erhebungsgebiet (Ost- und Westdeutschland) aufgeteilt und die Option „Gruppen vergleichen“ gewählt
- Auswertungen wie z.B. Häufigkeitstabellen werden nun zu Vergleichszwecken stets zusammen angezeigt:

Datei aufteilen



Prozedur	Syntax
Datei aufteilen	SORT CASES BY X . SPLIT FILE LAYERED BY X.
Aufteilung wieder rückgängig machen	SPLIT FILE OFF.

Datei aufteilen

WICHTIGKEIT: BERUF UND ARBEIT

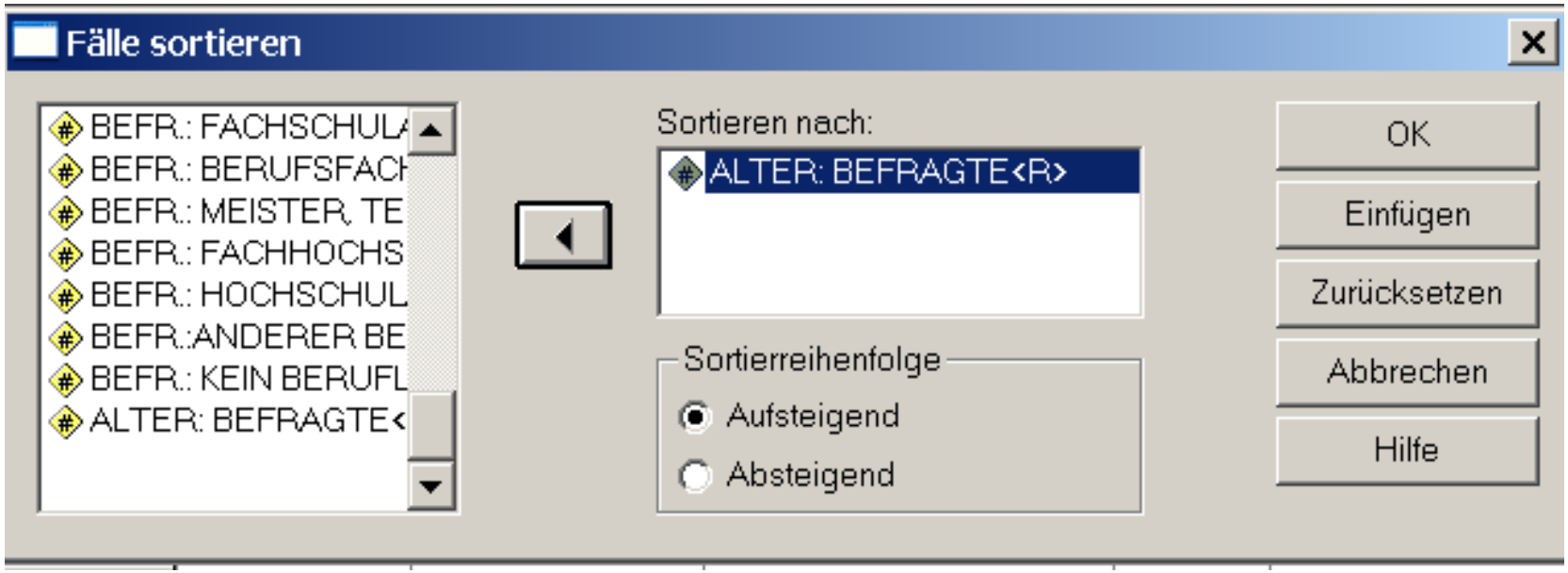
ERHEBUNGSGEBIET: WEST - OST			Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
ALTE BUNDESLAENDER	Gültig	UNWICHTIG	240	10,8	10,9	10,9
			109	4,9	4,9	15,8
			117	5,3	5,3	21,2
			197	8,9	8,9	30,1
			277	12,5	12,6	42,7
			439	19,8	19,9	62,6
		SEHR WICHTIG	824	37,3	37,4	100,0
		Gesamt	2203	99,6	100,0	
	Fehlend	KEINE ANGABE	9	,4		
	Gesamt		2212	100,0		
NEUE BUNDESLAENDER	Gültig	UNWICHTIG	138	13,5	13,5	13,5
			34	3,3	3,3	16,9
			39	3,8	3,8	20,7
			38	3,7	3,7	24,4
			91	8,9	8,9	33,3
			140	13,7	13,7	47,1
		SEHR WICHTIG	540	52,8	52,9	100,0
		Gesamt	1020	99,8	100,0	
	Fehlend	KEINE ANGABE	2	,2		
	Gesamt		1022	100,0		

Fälle sortieren

- Der Datensatz kann nicht nur nach bestimmten Kriterien aufgeteilt, sondern auch in einer bestimmten Reihenfolge *sortiert* werden (*Daten – Fälle sortieren* oder Syntax-Befehl **SORT CASES BY**)
- Bei der auf der nächsten Folie abgebildeten Einstellung wird der Datensatz z.B. aufsteigend nach dem Alter der Befragten sortiert. In der ersten Zeile des Dateneditors steht also der jüngste, in der letzten Zeile der älteste Befragte
- Die Syntax sieht so aus:

Prozedur	Syntax
Fälle aufsteigend nach X sortieren	SORT CASES BY X (A).
Fälle aufsteigend nach X und Y sortieren	SORT CASES BY X (A) Y (A).

Fälle sortieren



Fälle gewichten

- Je nach Stichprobe kann es darüber hinaus vorkommen, dass die Fälle **gewichtet,** d.h. mit einem bestimmten Faktor < 1 oder > 1 multipliziert werden müssen
- In SPSS steht nur eine Häufigkeitsgewichtung zur Verfügung
- Mit der Gewichtung einer Stichprobe kann z.B. erreicht werden, dass das Stichprobenprofil der Untersuchung einem gewünschten Profil wie bspw. der zugrundeliegenden Grundgesamtheit angenähert wird (sog. *Redressmentgewichtung*, siehe das weiter unten folgende Beispiel)

Fälle gewichten

- Von der Redressmentgewichtung unterscheidet man die *Design- oder Transformationsgewichtung*, die dann notwendig wird, wenn nicht alle Elemente der Grundgesamtheit die gleiche Chance hatten, in die Stichprobe zu gelangen
- Beispiel: Es werden Haushalte zufällig ausgewählt und pro Haushalt nur eine Person befragt
- In diesem Fall haben Personen in Mehrpersonenhaushalten eine geringere Auswahlchance als Personen in Einpersonenhaushalten; dies kann durch Gewichtung korrigiert werden

Fälle gewichten

Beispiel zur Konstruktion eines Redressmentgewichtes

Die Tabelle zeigt die relative Häufigkeit verschiedener allgemeiner Schulabschlüsse in einer Stichprobe und einer externen Vergleichsstichprobe (Daten frei erfunden!)

Höchster allgemeiner Schulabschluss	Stichprobe A	Externe Vergleichsstichprobe B (z.B. Mikrozensus)
	Prozent	
Kein Abschluss	6,7	10,0
Volks-, Hauptschule	26,7	33,0
Mittlere Reife	33,3	27,0
(Fach)Hochschulreife	33,3	30,0

► Es zeigt sich ein Bildungsbias: In Stichprobe A sind besser Gebildete überrepräsentiert.

Fälle gewichten

Diese Verzerrung soll durch ein Redressment-Gewicht korrigiert werden. „Soll-Durch-Ist-Prozedur“ wird angewandt: In jeder Zelle dividiert man die relative Häufigkeit der externen Tafel (Soll-Tafel, Stichprobe B) durch die relative Häufigkeit in der Ist-Tafel (Stichprobe A).

Soll-Wert	**Ist-Wert**	**Redressment-Gewicht (Soll / Ist)**
10,0	6,7	1,49
33,0	26,7	1,25
27,0	33,3	0,81
30,0	33,3	0,90

Alle Fälle der Ist-Datei, die zu dieser Zelle gehören, erhalten dann das so entstandene Gewicht als Multiplikator:

Person Nr.	**Höchster allgemeiner Schulabschluss**	**Redressment-Gewicht**
1	Kein Abschluss	1,49
2	(Fach)Hochschule	0,90
3	Volks-, Hauptsch.	1,25
4	Mittlere Reife	0,81

Fälle gewichten

- Eine *Redressment*gewichtung ist nur unter bestimmten Bedingungen sinnvoll:
- Erstens dann, wenn die üblichen Gewichtungsvariablen (z.B. Bildung) mit den anderen erhobenen Merkmalen (z.B. Parteipräferenz, Lebenszufriedenheit) eng korrelieren
- Wenn in diesem Fall der Bildungsbias durch Gewichtung beseitigt wird, werden auch Verzerrungen bei den anderen Merkmalen korrigiert
- Leider korrelieren in der Praxis die Gewichtungsvariablen nicht eng genug mit anderen Variablen, was den Sinn der Redressment-gewichtung in Frage stellt

Fälle gewichten

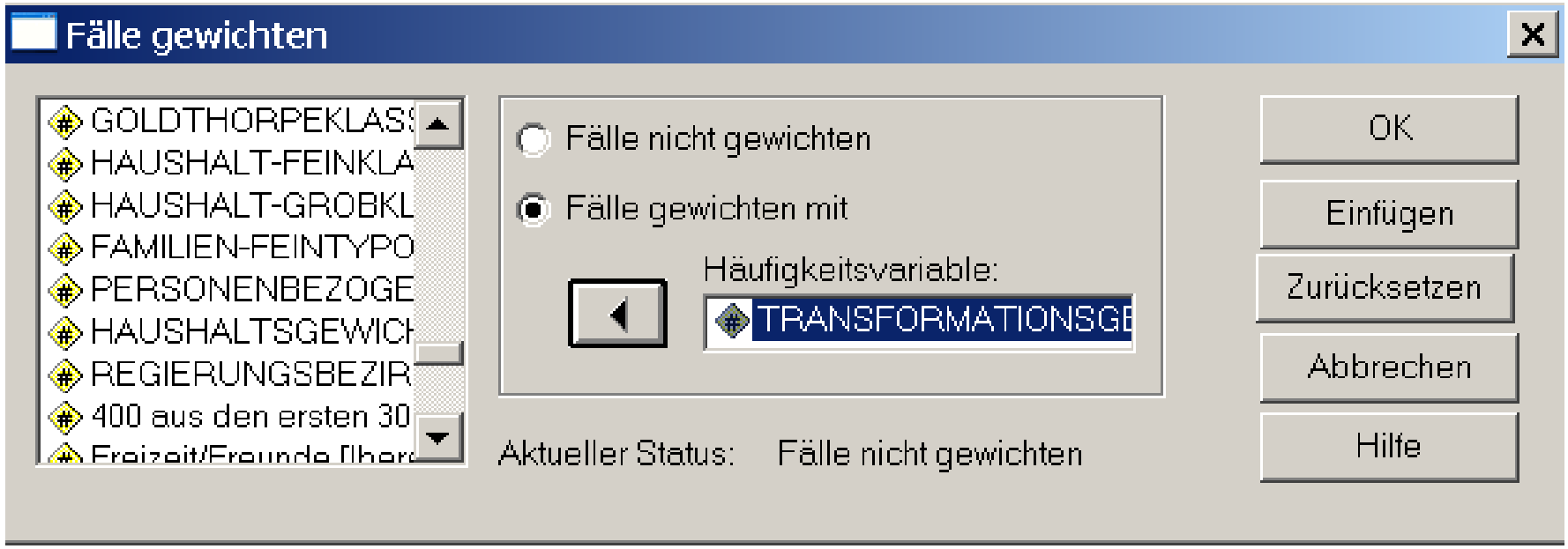
- Zweitens kann eine Redressmentgewichtung dann sinnvoll sein, wenn die Wahrscheinlichkeit einer Nichtteilnahme an der Befragung (Nonresponse) eng mit den Gewichtungsvariablen zusammenhängt
- Auch in diesem Fall würde der Nonresponse-Bias durch Gewichtung beseitigt
- Leider hängen die typischen soziodemografischen Gewichtungs-variablen wie Alter, Bildung usw. in der Praxis auch mit der Nonresponse-Wahrscheinlichkeit häufig nicht eng genug zusammen
- Gewichte werden oft mit dem Datensatz geliefert (z.B. im ALLBUS) und entsprechend in der Dokumentation des Datensatzes erläutert
- Die Eigenkonstruktion von Gewichten empfiehlt sich darüber hinaus in der Regel nicht (wenn, dann nur unter Anleitung einschlägiger Literatur!)

Fälle gewichten

- Die jeweilige Gewichtungsvariable, die im Datensatz bzw. der Datensatzdokumentation in der Regel beschrieben ist, kann unter *Daten – Fälle gewichten* angegeben werden (siehe nächste Folie)
- Alle statistischen Analysen werden dann nach der jeweiligen Variablen gewichtet, bis sie die Gewichtung ausschalten oder eine andere Gewichtungsvariable angeben
- Die Syntax sieht so aus:

Prozedur	Syntax
Fälle mit Variable X gewichten	WEIGHT BY X.
Gewichtung ausschalten	WEIGHT OFF.

Fälle gewichten



Datenaufbereitung: Vorbemerkungen

- Im Rahmen eines empirischen Projektes mit quantitativen Daten nimmt die Datenaufbereitung mit Abstand die meiste Zeit in Anspruch, wird oft unterschätzt und stellt unerfahrene Anwender häufig vor Probleme
- Die meisten Aufgaben lassen sich jedoch mit einfachem logischen Denken und einigen wenigen Syntaxbefehlen (in SPSS vor allem RECODE, COMPUTE und IF) bewerkstelligen
- Es wird dringend empfohlen, sich diese relativ einfachen Syntaxbefehle (und vor allem: die dahinter stehende Logik) einzuprägen und ihre Anwendung einzuüben
- Von einer Datenaufbereitung per Menü bzw. „klicken“ wird aus mehreren Gründen dringend abgeraten

Datenaufbereitung: Vorbemerkungen

- Empirische Forschung sollte nachvollziehbar und replizierbar sein; dies ist nur gewährleistet, wenn insbesondere die Datenaufbereitung in der Form eines Syntaxfiles dokumentiert wird
- Fehler in der Datenaufbereitung können, wenn man nur mit Menü arbeitet, im Nachhinein nicht mehr nachvollzogen werden
- Nur mit Hilfe von Syntax kann man einfach und schnell Modifikationen, z.B. bei der Operationalisierung von Variablen, vornehmen (das „Basteln“ an den Daten ist an der Tagesordnung; in den seltensten Fällen steht sofort die endgültige Lösung)

Datenaufbereitung: Vorbemerkungen

- Nur mit Hilfe von Syntax kann man vorläufige Datensätze auswerten und diese Aufbereitungen und Auswertungen dann auf den vollständigen Datensatz übertragen (arbeitet man nur mit Menü, muss man alles doppelt und dreifach machen, was unter Umständen Wochen dauern kann, mit Syntax dagegen Sekunden)
- Im Folgenden werden daher die Prozeduren zur Datenaufbereitung ausschließlich syntaxbasiert vorgestellt
- Bei der Daten*auswertung* empfiehlt es sich zwar ebenfalls, Syntaxfiles anzulegen
- Eine Häufigkeitstabelle oder Korrelation kann man jedoch durchaus auch mal kurz „klicken“, zumal die SPSS-Syntax hier häufig uneinheitlich bzw. umständlich und daher wenig einprägsam ist

Variable umkodieren

- Das Umkodieren von Variablen wird sehr häufig im Rahmen der Datenaufbereitung notwendig
- Zum Beispiel dann, wenn man bei metrischen Variablen wie Einkommen Klassen bilden will,
- …wenn bei kategorialen Variablen verschiedene Kategorien zusammengefasst werden sollen (z.B. verschiedene nicht-christliche Konfessionen zur Sammelkategorie „sonstige Konfession“)
- …wenn Wertelabels so umkodiert werden sollen, dass z.B. bei der Variablen Konflikte, die 5-fach abgestuft erfragt wurde, die Codierung umgepolt werden soll, so dass der Wert 5 für „hohes Konfliktniveau“ steht

Variable umkodieren

- Grundsätzlich zu unterscheiden ist zwischen dem Umkodieren in dieselbe Variable und in eine andere Variable
- Wird in dieselbe Variable umkodiert, wird die Variable X überschrieben. Man sollte vor Anwendung dieses Befehls genau überlegen, ob man die Originalversion von X noch braucht
- Wenn das der Fall ist, sollte X besser in eine andere Variable Y umkodiert werden (die Ursprungsvariable bleibt hierbei erhalten und eine neue Variable wird erstellt)
- Im Folgenden wird der **RECODE**-Befehl anhand von zwei Beispielen erläutert

Variable umkodieren

- Beispiel 1: Die ordinale CASMIN-Bildungsklassifikation soll in Bildungsjahre (metrisches Skalenniveau) umgerechnet werden und zwar nach folgendem Schema:

Code	CASMIN-Abschluss	Bildungsjahre
0	In school	12
1	Inadequately completed	8
2	General elementary school	9
3	Basic vocational qualification	11
4	Intermediate general qualification	10
5	Intermediate vocational	12
6	General maturity certificate	13
7	Vocational maturity certificat	15
8	Lower tertiary education	16
9	Higher tertiary education	18

Variable umkodieren

Prozedur	Syntax
Umkodierung der Variable casmin in dieselbe Variable	RECODE casmin (0 = 12) (1 = 8) (2 = 9) (3 = 11) (4 = 10) (5 = 12) (6 = 13) (7 = 15) (8 = 16) (9 = 18). EXECUTE.
Umkodierung von casmin in die neue Variable „bildungj“	RECODE casmin (0 = 12) (1 = 8) (2 = 9) (3 = 11) (4 = 10) (5 = 12) (6 = 13) (7 = 15) (8 = 16) (9 = 18) INTO bildungj. EXECUTE.

Nicht vergessen: Jede Befehlszeile mit einem Punkt abschließen.
Ausführung erfolgt erst durch „EXECUTE.“

Variable umkodieren

- Beispiel 2: Die metrische Variable Alter soll in vier Altersgruppen (18-25, 26-35, 36-45, 46-55 Jahre) umkodiert werden (Variable „alter_kat“); alle anderen Alterswerte sollen auf System-Missing gesetzt werden

Prozedur	**Syntax**
Umkodierung der Variable Alter in die neue Variable „alter_kat“	RECODE alter (18 thru 25 = 1) (26 thru 35 = 2) (36 thru 45 = 3) (46 thru 55 = 4) (ELSE = SYSMIS) INTO alter_kat. EXECUTE.

Nicht vergessen: Jede Befehlszeile mit einem Punkt abschließen.
Ausführung erfolgt erst durch „EXECUTE.“

Variable umkodieren

- Bei der neu erstellten Variablen „alter_kat“ ist nun in der Variablenansicht nicht ersichtlich, was die Codes 1-4 zu bedeuten haben. Es empfiehlt sich daher, per Syntax neue Variablen- und Wertelabels zu vergeben:

Prozedur	**Syntax**
Vergeben eines Variablenlabels „Alter kategorisiert“ für „alter_kat“	VARIABLE LABELS alter_kat „Alter kategorisiert“. EXECUTE.
Vergeben von Wertelabeln für „alter_kat“	VALUE LABELS alter_kat 1 ‘18-25 Jahre‘ 2 ‘26-35 Jahre‘ 3 ‘36-45 Jahre‘ 4 ‘46-55 Jahre‘. EXECUTE.

Nicht vergessen: Jede Befehlszeile mit einem Punkt abschließen.
Ausführung erfolgt erst durch „EXECUTE.“

Variable berechnen

- Häufig müssen Variablen durch bestimmte Berechnungsprozeduren (Addition, Multiplikation, Division) transformiert werden; dies erfolgt mit dem Syntax-Befehl **COMPUTE**.
- Auf der nächsten Folie finden sich zwei einfache Beispiele
- Beim ersten wird aus dem Geburtsjahr des Befragten (Variable „gebjahr“) und dem Erhebungsjahr (2007) das Alter des Befragten im Erhebungsjahr gebildet; im zweiten Beispiel auf dieser Basis das Alter im Jahr 1998
- Beim dritten Beispiel wird auf der Basis der Partnerschaftsdauer (Variable „pdauer“), die monatsgenau gemessen ist, die Partnerschaftsdauer in Jahren berechnet

Variable berechnen

Prozedur	Syntax
Berechnung des Alters des Befragten im Jahr 2007	COMPUTE alter = 2007 - gebjahr. EXECUTE.
Berechnung des Alters im Jahr 2008 auf der Basis vom oben berechneten Alter im Jahr 2007	COMPUTE alter98 = alter + 1. EXECUTE.
Umrechnung der monatsgenauen Partnerschaftsdauer in Jahre	COMPUTE pdauer_jahr = pdauer / 12. EXECUTE.

Nicht vergessen: Jede Befehlszeile mit einem Punkt abschließen.
Ausführung erfolgt erst durch „EXECUTE.“

Variable berechnen

- Beim Beispiel unten wird dargestellt, wie man die metrische Variable Alter so umformt, dass man einen nichtlinearen (u-förmigen oder glockenförmigen) Effekt des Alters auf eine metrische abhängige Variable (in einer linearen Regression) abbilden kann
- Es ist notwendig, a) das Alter zu zentrieren, indem man den arithmetischen Mittelwert (20 Jahre) abzieht und b) wird dieses zentrierte Alter dann quadriert (beide Terme gehen gemeinsam in die Regression ein):

Prozedur	**Syntax**
Zentrierung der Variable Alter (Mittelwert = 20 Jahre)	COMPUTE alterz = alter - 20. EXECUTE.
Quadrieren des zentrierten Alters	COMPUTE alterquad = alterz * alterz. EXECUTE.

Variable berechnen

- Der Compute-Befehl wird auch häufig eingesetzt, um eine Skala, die als Mittelwert von verschiedenen einzelnen Items gebildet wird, zu berechnen
- Auf der nächsten Folie sind als Beispiel verschiedene Items zur subjektiven Ähnlichkeit von Partnern dargestellt (jeweils 5-fach abgestuftes Antwortformat)
- Wenn diese Items zur Zusammenfassung in eine Skala geeignet sind (vgl. das Skript „Faktorenanalyse und Skalierung“), kann man aus diesen Variablen die neue Variable „matching“ als Mittelwert der Einzelitems berechnen

Variable berechnen

Variablenname	Item
matching1	„Mein Partner und ich haben die gleichen Ansichten über den Umfang mit Geld“
matching2	„Mein Partner und ich haben überwiegend die gleichen Freizeitinteressen“
matching3	„Mein Partner und ich haben die gleiche sexuelle Wellenlänge“

Prozedur	Syntax
Berechnung der Skala „matching“ (mindestens 1 Item hat einen gültigen Wert)	COMPUTE matching = MEAN (matching1, matching2, matching3). EXECUTE.
Berechnung der Skala „matching“ (mindestens 2 Items haben gültige Werte)	COMPUTE matching = MEAN.2 (matching1, matching2, matching3). EXECUTE.

Variable berechnen

- Noch häufiger als der Compute-Befehl wird im Rahmen der Datenaufbereitung der **IF**-Befehl angewendet, einem Spezialfall von Compute
- Mit dem If-Befehl kann eine Berechnung unter eine oder mehrere Bedingungen gestellt werden; nur wenn diese erfüllt sind, wird die Berechnung durchgeführt
- Verdeutlichen wir dies an zwei Beispielen. Beim ersten Beispiel auf der nächsten Folie soll aus den Variablen „zp_sex“ (Geschlecht der Zielperson, 1 = Mann, 2 = Frau) und „p_sex“ (Geschlecht des Partners der Zielperson, genauso codiert) die neue Variable „hetero“ gebildet werden (1 = heterosexuell, 0 = homosexuell)

Variable berechnen

Prozedur	Syntax

Berechnung der Variable „hetero“ mit Hilfe der Variablen „zp_sex“ und „p_sex“	IF (zp_sex = 1 AND p_sex = 1) OR (zp_sex = 2 AND p_sex = 2) hetero = 0.
	IF (zp_sex = 1 AND p_sex = 2) OR (zp_sex = 2 AND p_sex = 1) hetero = 1.
	EXECUTE.

Vergeben eines Variablenlabels	VARIABLE LABELS hetero "Homo- versus Heterosexuell". EXECUTE.
Vergeben von Wertelabels	VALUE LABELS hetero 0 'homosexuell' 1 'heterosexuell'. EXECUTE.

Nicht vergessen: Jede Befehlszeile mit einem Punkt abschließen.
Ausführung erfolgt erst durch „EXECUTE.“

Variable berechnen

- Das nächste Beispiel betrifft die Berechnung des sog. Inglehart-Index auf der Basis von 4 Items
- „Ruhe und Ordnung“ (siehe nächste Folie) sowie „Bekämpfung Preisanstieg“ sind Items, die Materialismus messen sollen; mit den Items „Bürgereinfluss“ und „freie Meinungsäußerung“ soll Postmaterialismus gemessen werden
- Inglehart hat diese vier Items zu einem 4-er Index kombiniert. Je nach Antwortmuster wird zwischen vier Typen unterschieden: reine Materialisten, reine Postmaterialisten, materialistischer sowie postmaterialistischer Mischtyp
- Nachfolgend finden sich eine Übersicht über die Typen und ein möglicher Syntaxbefehl zur Bildung des Index

Variable berechnen

WICHTIGKEIT VON RUHE UND ORDNUNG

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	AM WICHTIGSTEN	1377	42,6	43,4	43,4
	AM ZWEITWICHTIGSTEN	820	25,4	25,9	69,3
	AM DRITTWICHTIGSTEN	620	19,2	19,6	88,9
	AM VIERTWICHTIGSTEN	353	10,9	11,1	100,0
	Gesamt	3170	98,0	100,0	
Fehlend	WEISS NICHT	52	1,6		
	KA	12	,4		
	Gesamt	64	2,0		
Gesamt		3234	100,0		

Variable berechnen

Wichtigstes Ziel	Zweitwichtigstes Ziel	Inglehart-Index
Ruhe und Ordnung (ro)	Bekämpfung Inflation (bi)	Reiner Materialist (Code 1)
Bekämpfung Inflation (bi)	Ruhe und Ordnung (ro)	Reiner Materialist (1)
Bürgereinfluss (be)	Meinungsäußerung (me)	Reiner Postmaterialist (4)
Meinungsäußerung (me)	Bürgereinfluss (be)	Reiner Postmaterialist (4)
Ruhe und Ordnung (ro)	Meinungsäußerung (me)	Materialistischer Mischtyp (2)
Ruhe und Ordnung (ro)	Bürgereinfluss (be)	Materialistischer Mischtyp (2)
Bekämpfung Inflation (bi)	Meinungsäußerung (me)	Materialistischer Mischtyp (2)
Bekämpfung Inflation (bi)	Bürgereinfluss (be)	Materialistischer Mischtyp (2)
Meinungsäußerung (me)	Ruhe und Ordnung (ro)	Postmaterialistischer Mischtyp (3)
Meinungsäußerung (me)	Bekämpfung Inflation (bi)	Postmaterialistischer Mischtyp (3)
Bürgereinfluss (be)	Ruhe und Ordnung (ro)	Postmaterialistischer Mischtyp (3)
Bürgereinfluss (be)	Bekämpfung Inflation (bi)	Postmaterialistischer Mischtyp (3)

Variable berechnen

Prozedur	Syntax

Berechnung des Inglehart-Index nach dem zuvor dargestellten Schema	IF (ro = 1 AND bi = 2) OR (bi = 1 AND ro = 2) ingle = 1.
	IF (be = 1 AND me = 2) OR (me = 1 AND be = 2) ingle = 4.
	IF (ro = 1 AND me = 2) OR (ro = 1 AND be = 2) ingle = 2.
	IF (bi = 1 AND me = 2) OR (bi = 1 AND be = 2) ingle = 2.
	IF (me = 1 AND ro = 2) OR (me = 1 AND bi=2) ingle = 3.
	IF (be = 1 AND ro = 2) OR (be = 1 AND bi = 2) ingle = 3.
	EXECUTE.

Nicht vergessen: Jede Befehlszeile mit einem Punkt abschließen.
Ausführung erfolgt erst durch „EXECUTE.“

Variable berechnen

- Während man mit dem IF-Befehl immer nur eine Berechnung unter eine Bedingung stellen kann, kann man mit **DO IF** mehrere Berechnungen unter dieselbe Bedingung stellen oder für verschiedene Bedingungen verschiedene Berechnungen anfordern
- Letzteres wird nun anhand eines Beispiels demonstriert: In der Tabelle auf der nächsten Folie sind fünf Ereignisse dargestellt, die eine zunehmende Instabilität der Ehe anzeigen
- Die Ereignisse sind Skalenniveaus zugeordnet; wir nehmen also z.B. an, dass die Scheidungsberatung (Stufe 5) ein gravierenderes Anzeichen für Instabilität ist als Gedanken über Schwierigkeiten in der Partnerschaft (Stufe 1)

Variable berechnen

Variable (1=ja)	Ereignis	Stufe
ps1	Gedanken, dass Partnerschaft in Schwierigkeiten ist	1
ps2	Ernsthafte Gedanken an eine Trennung	2
ps3	Mit Freund/in über Trennung gesprochen	3
ps4	Partner eine Trennung vorgeschlagen	4
ps5	Scheidungsberatung aufgesucht	5

- Die nachfolgend dargestellte Syntax produziert eine Guttman-Skala (mit 6 Ausprägungen) und ist wie folgt zu lesen: Wenn man schon eine Scheidungsberatung aufgesucht hat (ps5 = 1), dann ist die Partnerschaft sehr instabil (instabil = 5); wenn dies nicht der Fall ist, man dem Partner aber eine Trennung vorgeschlagen hat (ps4 = 1), dann ist die Partnerschaft instabil (instabil = 4) usw…

Variable berechnen

Prozedur	Syntax
Berechnung einer Instabilitätsskala nach dem zuvor dargestellten Schema	DO IF ps5 = 1. COMPUTE instabil = 5. ELSE IF ps4 = 1. COMPUTE instabil = 4. ELSE IF ps3 = 1. COMPUTE instabil = 3. ELSE IF ps2 = 1. COMPUTE instabil = 2. ELSE IF ps1 = 1. COMPUTE instabil = 1. ELSE IF (ps 5 = 0 AND ps4 = 0 AND ps3 = 0 AND ps2 = 0 AND ps1 = 0). COMPUTE instabil = 0. END IF. EXECUTE.

Nicht vergessen: Jede Befehlszeile mit einem Punkt abschließen.
Ausführung erfolgt erst durch „EXECUTE.“

Variable berechnen

- Ein weiterer nützlicher Befehl ist **COUNT**, mit dem man für eine Untersuchungseinheit (Zeile im Datensatz) die Anzahl von spezifischen Werten über verschiedene Variablen hinweg zählen kann
- Auf der nächsten Folie sind zwei sinnvolle Anwendungsmöglichkeiten dargestellt
- Im ersten Beispiel wird gezählt, wie viele systemdefiniert fehlende Werte eine Person (Zeile im Datensatz) bei den Variablen 1-5 aufweist; die entsprechende Anzahl wird als neue Variable „nmissing“ im Datensatz abgespeichert (man kann mit ihrer Hilfe z.B. nur Personen auswählen, die nmissing = 0 haben)
- Im zweiten Beispiel zählen wir die Anzahl der Kinder einer Person; zugrunde liegen 5 Dummy-Variablen (kind1 – kind5), die jeweils erfassen, ob ein erstes Kind geboren wurde (kind1 = 1), ein zweites usw.

Variable berechnen

Prozedur	Syntax
Berechnung einer Variablen „nmissing“, die die Anzahl von SYSMIS bei den Variablen 1-5 erfasst	COUNT nmissing = var1 var2 var3 var4 var5 (SYSMIS). EXECUTE.
Berechnung einer Variablen „nkids“, die die Anzahl von 1er-Codes bei den Variablen kind 1-5 erfasst	COUNT nkids = kind1 kind2 kind3 kind4 kind5 (1). EXECUTE.

Nicht vergessen: Jede Befehlszeile mit einem Punkt abschließen.
Ausführung erfolgt erst durch „EXECUTE.“

Z-Transformation

- Zu den häufig angewendeten Variablentransformationen gehört auch die *Z-Standardisierung* von Variablenwerten
- Man erzeugt Z-transformierte Werte, indem man von jedem Messwert (x_i) das arithmetische Mittel ($\bar{x}$) subtrahiert und die Differenz (xi- $\bar{x}$) durch die Standartabweichung (s) dividiert:

$$z_i = \frac{x_i - \bar{x}}{s}$$

- Die Z-Transformation ermöglicht es auch, zwei Variablen, die in unterschiedlicher Metrik gemessen sind (z.B. Alter und Einkommen), direkt miteinander zu vergleichen

Z-Transformation

Prozedur	Syntax
Z-Standardisierung einer Variablen X	DESCRIPTIVES VARIABLES = X /SAVE /STATISTICS = MEAN STDDEV MIN MAX . EXECUTE.

- Seltsamerweise findet sich die Z-Transformation in SPSS im Rahmen der Auswertungsprozedur „deskriptive Statistiken“; der entsprechende zusätzliche Befehl lautet „/SAVE“
- SPSS führt dann die Z-Transformation durch und speichert eine entsprechende neue Variable im Datensatz ab (der alte Name wird übernommen und ein Z davor gesetzt)

Aggregieren von Daten

- Manchmal wird es erforderlich, basierend auf den Werten von einer oder mehreren Gruppenvariablen (*Break-Variablen*) Fälle zusammenzufassen bzw. zu aggreggieren; die neue Datendatei enthält dann für jede Break-Gruppe einen Fall
- Verstehen tut man das nur anhand eines Beispiels: Betrachten wir den frei erfundenen Datensatz auf der nächsten Folie, der 15 Personen und 4 Variablen enthält
- Anhand der Personen- und Haushalts-ID erkennen wir, dass einige Personen zusammen in einem Haushalt wohnen, z.B. die Personen mit der ID 2, 3 und 4 im Haushalt 2

Aggregieren von Daten

personid	haushaltid	stellunghv	alter
1.00	1.00	.00	45.00
2.00	2.00	.00	36.00
3.00	2.00	1.00	34.00
4.00	2.00	3.00	12.00
5.00	3.00	.00	48.00
6.00	4.00	.00	29.00
7.00	4.00	1.00	35.00
8.00	4.00	3.00	5.00
9.00	4.00	3.00	3.00
10.00	4.00	3.00	1.00
11.00	6.00	.00	66.00
12.00	7.00	.00	50.00
13.00	7.00	1.00	45.00
14.00	7.00	3.00	17.00
15.00	7.00	3.00	22.00

Aggregieren von Daten

- Die Variable „stellunghv“ (Stellung zum Haushaltsvorstand) sagt uns, bei welcher Person es sich um den Haushaltsvorstand handelt (Code 0), um den Partner des Haushaltsvorstandes (Code 1) und um Kinder des Haushaltsvorstandes oder dessen Partner (Code 3)
- Außerdem wissen wir für jede Person das Alter
- Wir wollen nun folgendes wissen: Wie viele Kinder leben im Haushalt des jeweiligen Haushaltsvorstandes? Wir können das im Datensatz zwar sehen, wie aber können wir eine entsprechende Variable im Datensatz abspeichern?
- Hier können wir uns mit der Aggregierung der Daten helfen; ein entsprechendes Syntaxbeispiel ist weiter unten dargestellt (hier macht es auch Sinn, das Menü anzuschauen: *Daten – aggregieren*)

Aggregieren von Daten

- Wir haben folgende Idee: Wir aggregieren unseren Datensatz auf Haushaltsebene und speichern für jeden Haushalt ab, wie viele Kinder darin leben
- Anschließend spielen wir diese Kinderzahl wieder dem Individualdatensatz zu
- Der erste Schritt besteht darin, eine Dummy-Variable zu erstellen die für jede Person angibt, um es sich um ein Kind handelt (1 = ja, 0 = nein)
- Syntax: RECODE stellunghv (3 = 1) (ELSE = 0) INTO kind.
  EXECUTE.
- Der Datensatz sieht dann so aus:

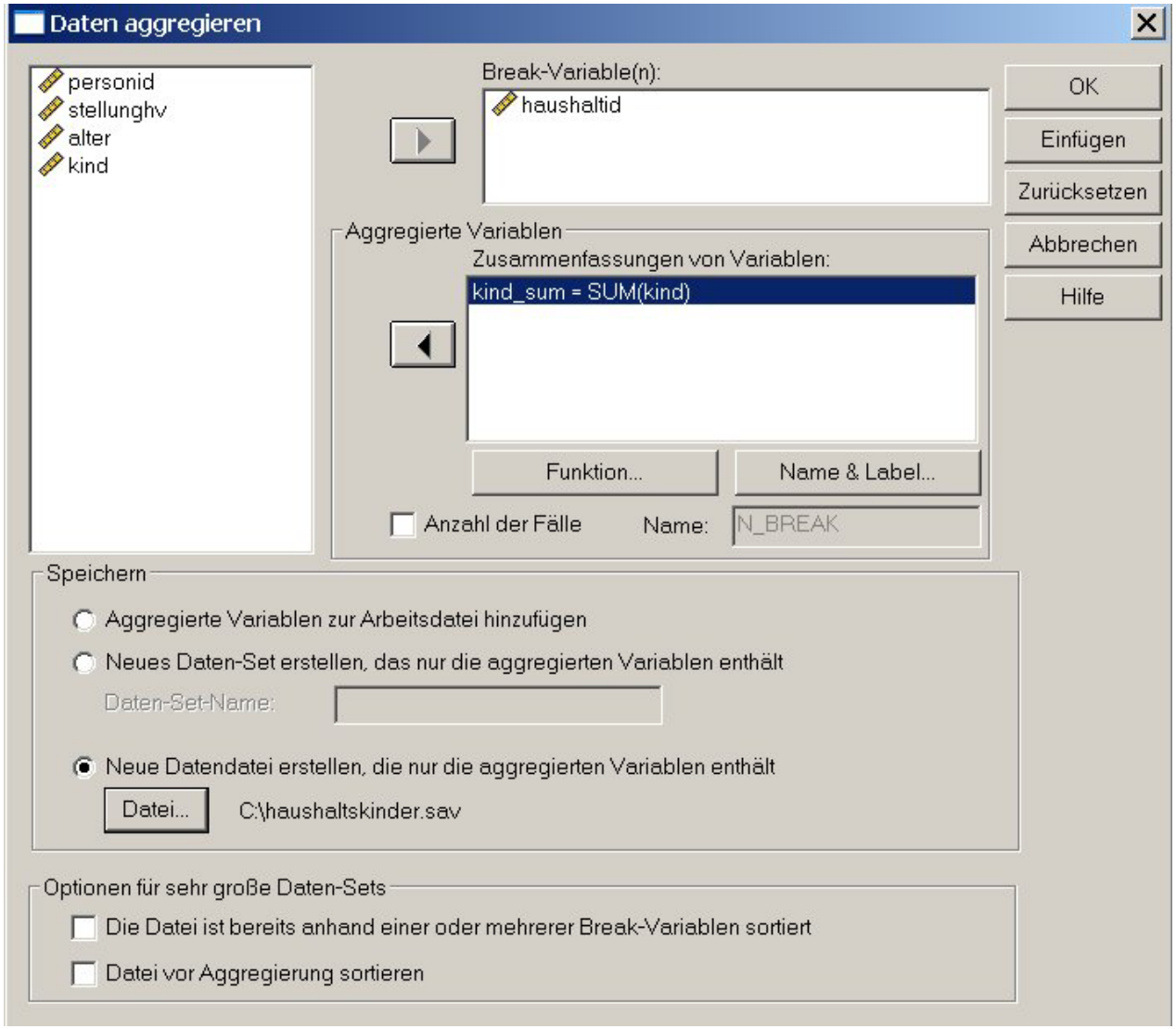
Aggregieren von Daten

personid	haushaltid	stellunghv	alter	kind
4.00	2.00	3.00	12.00	1.00
5.00	3.00	.00	48.00	.00
6.00	4.00	.00	29.00	.00
7.00	4.00	1.00	35.00	.00
8.00	4.00	3.00	5.00	1.00
9.00	4.00	3.00	3.00	1.00
10.00	4.00	3.00	1.00	1.00
11.00	6.00	.00	66.00	.00
12.00	7.00	.00	50.00	.00
13.00	7.00	1.00	45.00	.00
14.00	7.00	3.00	17.00	1.00
15.00	7.00	3.00	22.00	1.00

Aggregieren von Daten

- Nun geben wir (per Menü oder Syntax, siehe die nächsten Folien) an, dass wir den Individualdatensatz auf Haushaltsebene aggregieren wollen („/BREAK = haushaltsid“)
- Dies bedeutet, dass in dem neu abgespeicherten Datensatz „haushaltskinder.sav“ die Zeilen nicht mehr Personen entsprechen, sondern Haushalten
- Wir aggregieren die Variable „kind“, wobei hier verschiedene Aggregierungsfunktionen zur Verfügung stehen (Summe, Mittelwert, usw., siehe das Menü)
- In unserem Fall brauchen wir die Summen-Funktionen, da wir ja die Anzahl von Kindern pro Haushalt wissen wollen

Aggregieren von Daten



Aggregieren von Daten

Prozedur	Syntax

Aggregierung der Summe der Kinder auf Haushaltsebene	AGGREGATE
	  /OUTFILE= 'C:\haushaltskinder.sav'
	  /BREAK= haushaltid
	  /kind_sum = SUM(kind).
	EXECUTE.

- Rechts sehen wir das Ergebnis der Aggregierung: Der neue Datensatz hat 7 Zeilen, die 7 Haushalten entsprechen
- Für jeden Haushalt ist die Anzahl der Kinder in „kind_sum“ abgespeichert

haushaltid	kind_sum
1.00	.00
2.00	1.00
3.00	.00
4.00	3.00
6.00	.00
7.00	2.00

Verschmelzen von Datensätzen

- Nun stehen wir vor dem nächsten Problem: Wie bekommen wir die Anzahl der Kinder (kind_sum) zurück in unseren Individualdatensatz?
- Nun geht es also um die Verschmelzung von Datensätzen
- Man kann in einen Arbeitsdatensatz andere Variablen aus anderen Datensätzen einfügen oder neue Fälle
- Im Folgenden wird demonstriert, wie man eine Variable aus einer externen Datei in einen Arbeitsdatensatz zuspielt
- Zunächst muss man sich klar machen, in welcher Beziehung die beiden Quelldateien stehen. Es gibt drei Möglichkeiten:

Verschmelzen von Datensätzen

Quelldatei 1

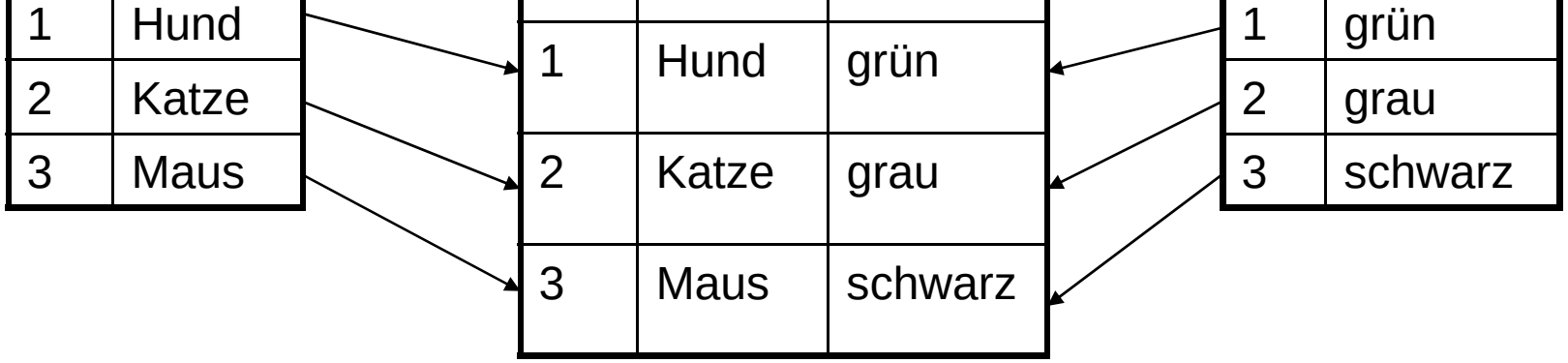
ID	Tier
1	Hund
2	Katze
3	Maus

Zieldatei

ID	Tier	Farbe
1	Hund	grün
2	Katze	grau
3	Maus	schwarz

Quelldatei 2

ID	Farbe
1	grün
2	grau
3	schwarz



- *1:1-Beziehung*: Jedem Fall der einen Datei entspricht genau ein Fall der anderen Datei. Die Werte sind also lediglich auf zwei Dateien aufgeteilt und können ohne weiteres in einer Datei nebeneinander stehen

Verschmelzen von Datensätzen

Quelldatei 1

ID	Tier
1	Hund
2	Katze
3	Maus

Zieldatei

ID	Tier	Farbe
1	Hund	grün
2	Katze	grau
3	Maus	
4		schwarz

Quelldatei 2

ID	Farbe
1	grün
2	grau
4	schwarz

- *Lose Beziehung*: Jedem Fall der einen Datei entspricht höchstens ein Fall der anderen. Im Gegensatz zu einer 1 zu 1 Beziehung ist es hier möglich, dass bei einzelnen Fällen kein Äquivalent in der jeweils anderen Datei existiert

Verschmelzen von Datensätzen

Schlüsseltabelle

ID	Tier
1	Hund
2	Katze
3	Maus
6	Esel

Zieldatei

ID	Tier	Farbe
1	Hund	grün
2	Katze	grau
3	Maus	schwarz
4		rot

abhängige Tabelle

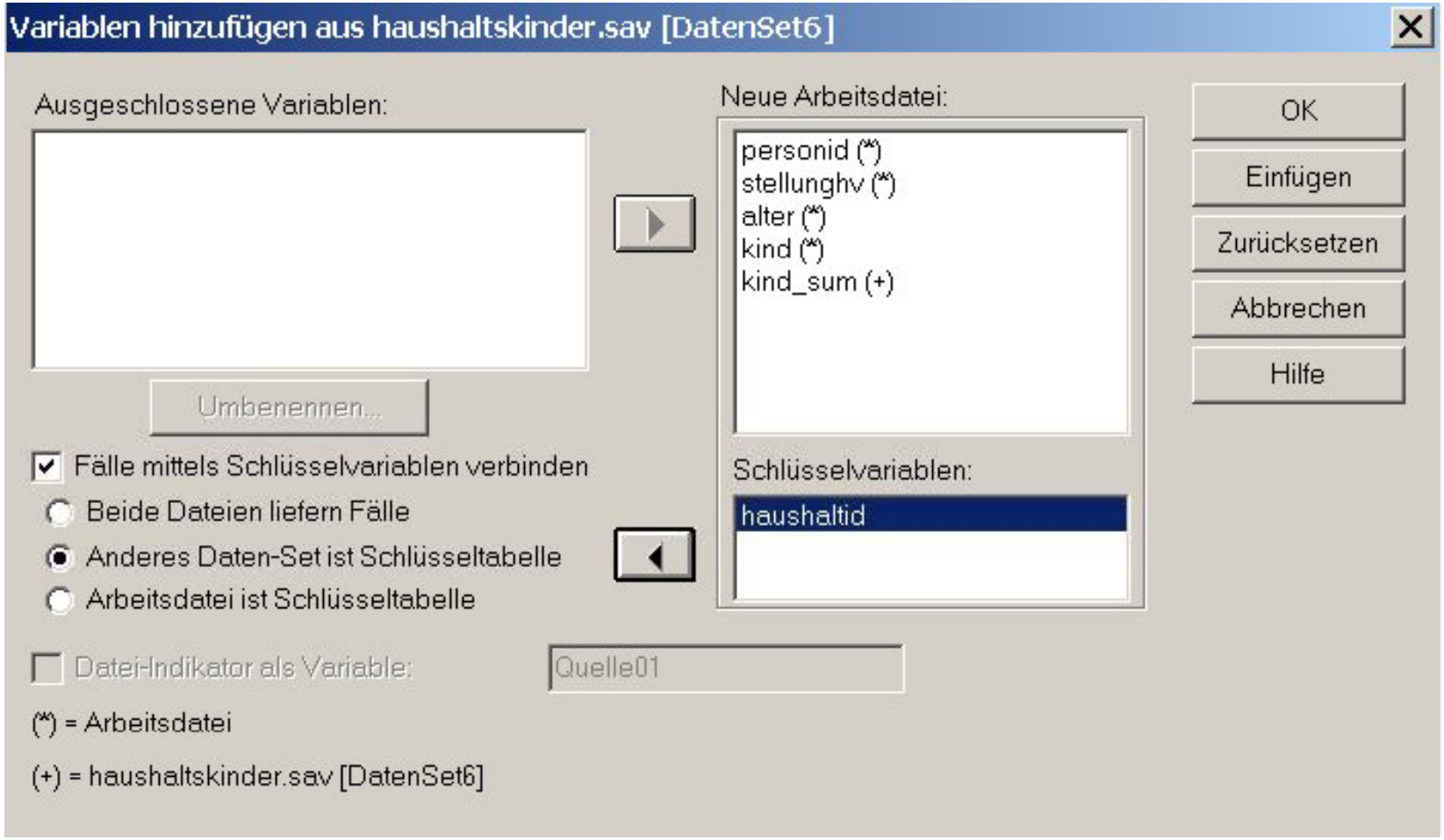
ID	Farbe
1	grün
2	grau
3	schwarz
4	rot

- *1:n-Beziehung*: Eine der beiden Dateien ist übergeordnet und damit die Schlüsseltabelle. Jedem Fall der Schlüsseltabelle können mehrere Fälle der anderen (abhängigen) Tabelle entsprechen. Jedem Fall der abhängigen Tabelle entspricht jedoch höchstens ein Fall der Schlüsseltabelle.

Verschmelzen von Datensätzen

- In unserem Beispiel handelt es sich um eine 1:n-Beziehung. Wir wollen unserer Arbeitsdatei auf Personenebene eine Haushaltsinformation (Anzahl der Kinder) zuspielen
- Die nächste Folie zeigt, wie die entsprechenden Einstellungen im Menü *Daten – Dateien zusammenfügen – Variablen hinzufügen* vorzunehmen sind (die Syntax ist wenig eingängig und wir hier unterschlagen; siehe das Log-File)
- Als Schlüsselvariable kommt nur die Haushalts-ID in Frage, die in beiden Datensätzen vorhanden sein muss; außerdem müssen beide Datensätze aufsteigend nach der Schlüsselvariable sortiert sein
- Unter „Neue Arbeitsdatei“ symbolisiert das „+“ hinter „kind_sum“, dass es sich hier um die Variable handelt, die neu hinzukommt. Die übernächste Folie zeigt das Ergebnis

Verschmelzen von Datensätzen



Verschmelzen von Datensätzen

personid	haushaltid	stellunghv	alter	kind	kind_sum
1.00	1.00	.00	45.00	.00	.00
2.00	2.00	.00	36.00	.00	1.00
3.00	2.00	1.00	34.00	.00	1.00
4.00	2.00	3.00	12.00	1.00	1.00
5.00	3.00	.00	48.00	.00	.00
6.00	4.00	.00	29.00	.00	3.00
7.00	4.00	1.00	35.00	.00	3.00
8.00	4.00	3.00	5.00	1.00	3.00
9.00	4.00	3.00	3.00	1.00	3.00
10.00	4.00	3.00	1.00	1.00	3.00
11.00	6.00	.00	66.00	.00	.00
12.00	7.00	.00	50.00	.00	2.00
13.00	7.00	1.00	45.00	.00	2.00
14.00	7.00	3.00	17.00	1.00	2.00
15.00	7.00	3.00	22.00	1.00	2.00

Verschmelzen von Datensätzen

- Zum Verschmelzen von Datensätzen (Variablen hinzufügen) hier eine Übersicht zu den Menüeinstellungen:

Verhältnis der Quelldateien	**Menüeinstellungen**
1:1-Beziehung	Es muss lediglich die andere Quelldatei angegeben werden und beide Datensätze müssen aufsteigend nach der ID sortiert sein (keine weiteren Menüeinstellungen notwendig)
Lose Beziehung	Es muss eine Schlüsselvariable angegeben und zusätzlich die Option „Beide Dateien liefern Fälle“ angewählt werden
1:n Beziehung	Es muss eine Schlüsselvariable angegeben werden; zusätzlich ist festzulegen, welche Datei abhängig ist (deren Fallzahl wird nicht verändert) und welche die Schlüsseltabelle

Umstrukturierung ins Long-Format

- Insbesondere für jede Form der Panel-Analyse ist es notwendig, die Daten vom sog. Wide- ins Long-Format umzustrukturieren (siehe auch die Skripte zur Panel- und Ereignisdatenanalyse)
- Die nächste Folie zeigt das wide-Format. Die Zeilen im Datensatz stehen hier wie gewohnt für Personen; zeitveränderliche Variablen werden in separaten Spalten nebeneinander abgespeichert
- Dargestellt ist hier u.a. die abhängige Variable („heirat1-heirat4“). Hier ist für insgesamt 4 Jahre erfasst, ob die entsprechende Person entweder unverheiratet ist (= 0) oder geheiratet hat (= 1)

Umstrukturierung ins Long-Format

	id	heirat1	heirat2	heirat3	heirat4	kdau1	kdau2	kdau3	kdau4
1	1	0	0	0	0	1	2	3	4
2	2	0	1	.	.	1	2	.	.
3	3	0	.	.	.	1	.	.	.
4	4	0	0	0	1	1	2	3	4
5	5	0	0	0	1	1	2	3	4
6	6	0	0	0	0	1	2	3	4
7	7	0	0	0	1	1	2	3	4
8	8	0	1	.	.	.	.	.	.
9	9	0	0	0	1	1	2	3	4
10	10	0	1	.	.	1	2	.	.
11	11	0	0	0	1	1	2	3	4
12	12	0	1	.	.	1	2	.	.
13	13	0	0	0	.	1	2	3	.
14	14	0	1	.	.	1	2	.	.
15	15	0	0	0	0	1	2	3	4
16	16	0	0	0	.	1	2	3	.
17	17	0	0	1	.	1	2	3	.
18	18	0	0	.	.	1	2	.	.
19	19	0	0	0	0	1	2	3	4
20	20	0	0	1	.	1	2	3	.

Umstrukturierung ins Long-Format

- Die nächste Folie zeigt den SPSS-Syntax-Befehl zur Umstrukturierung des Datensatzes ins long-Format (Menü: *Daten – Umstrukturieren – ausgewählte Variablen in Fälle*)
- Es ist ersichtlich, dass, z.B. bei der abhängigen Variablen, die Informationen aus den 4 einzelnen Spalten („heirat1-heirat4“) zu einer Variablen (einer Spalte) im long-Format zusammengefasst werden
- Damit SPSS weiß, welche personenspezifischen Beobachtungen zusammengehören, wird unter „KEEP“ die Personen-ID angegeben; zeitkonstante Variablen (im Menü: „mit festem Format“) können ebenfalls hier angegeben werden
- „Index = Index1(4)“ bedeutet, dass SPSS eine neue Variable „Index1“ erstellt, die hier der laufenden Nummer der Panelwelle entspricht

Umstrukturierung ins Long-Format

Prozedur	Syntax
Umstrukturierung vom Wide- ins Long-Format	VARSTOCASES /MAKE heirat FROM heirat1 heirat2 heirat3 heirat4 /MAKE kdau FROM kdau1 kdau2 kdau3 kdau4 /MAKE alter FROM alter1 alter2 alter3 alter4 /INDEX = Index1(4) /KEEP = id /NULL = KEEP. EXECUTE.

Umstrukturierung ins Long-Format

	id	Index1	heirat	kdau	alter
1	1	1	0	1	33
2	1	2	0	2	34
3	1	3	0	3	35
4	1	4	0	4	36
5	2	1	0	1	35
6	2	2	1	2	36
7	2	3	.	.	.
8	2	4	.	.	.
9	3	1	0	1	31
10	3	2	.	.	.
11	3	3	.	.	.
12	3	4	.	.	.
13	4	1	0	1	29
14	4	2	0	2	30
15	4	3	0	3	31
16	4	4	1	4	32
17	5	1	0	1	30
18	5	2	0	2	31
19	5	3	0	3	32
20	5	4	1	4	33
21	6	1	0	1	31
22	6	2	0	2	32
23	6	3	0	3	33
24	6	4	0	4	34

- Das Ergebnis der Umstrukturierung: Personen fließen im long-Format nun in der Form von Zeilen mehrfach in den Datensatz ein (hier jeweils 4-mal)
- Pro Variable gibt es im long-Format jedoch nur noch eine Spalte
- Die ID identifiziert die personenspezifischen Beobachtungen

Datensatz per Syntax laden und abspeichern

Prozedur	Syntax
Öffnen eines Datensatzes über Syntax	GET FILE = ‘ C:\Eigene Dateien\Beispiel.sav‘. EXECUTE.
Abspeichern eines Datensatzes über Syntax	SAVE OUTFILE = ‘C:\Eigene Dateien\Beispiel.sav‘ /KEEP = var1 var2 var3 var4. EXECUTE.

- Die letzte Folie zeigt, wie man einen Datensatz in SPSS per Syntax laden und abspeichern kann
- Der Save-Outfile-Befehl ist sehr nützlich, um einen neuen reduzierten Datensatz abzuspeichern, in dem nicht alle Variable des ursprünglichen Datensatzes enthalten sind; außerdem kann die Reihenfolge der Variablen verändert werden