SPSS Grundlagen online-Kurs

Teil 1

Ablauf

- Tag 1: Dateneingabe und Datenbereinigung
- Tag 2: Datenbeschreibung, statistische Tests
- Tag 3: statistische Tests, Analysen

1 Dateneingabe

Es gibt mehrere Möglichkeiten, Daten in SPSS einzugeben:

- Direkte Eingabe der Daten über die Tastatur, es muss auch die Variablenansicht vervollständigt werden
- Übertragung aus anderem Dateiformat, z.B. Excel. Hier wird ein Teil der Variablenbeschreibung automatisch eingetragen.
  Es können problemlos auch Dateiformate wie z.B. .txt, .por, .csv… übertragen werden. Bei anderen Formaten empfiehlt sich ein Zwischenschritt über eins der vorab genannten Formate.
  Wie sollten die Datensätze aussehen? → s. nächste Folie

2 Dateneingabe

- In den **Spalten** stehen die einzelnen gemessenen Merkmale (Variablen), d.h. alle Messungen für z.B. das Gewicht stehen in einer Spalte
- In einer **Zeile** stehen alle stehen alle Messungen einer Beobachtungseinheit (z.B. Patient, Pflanze)
- Keine freien Zeilen zur besseren Übersichtlichkeit
- Werteeingaben immer einheitlich, also entweder nur Text oder nur Zahl (Empfehlung: immer als Zahl kodieren)
- Bei fehlenden Werten Zellen frei lassen
- Variablenname muss mit Buchstaben beginnen, keine Leer- und Sonderzeichen
- SPSS ist nicht „case-sensitiv“, d.h. Alter = alter = ALTER

3 Dateneingabe

falsch

	Männer		Frauen	
	Größe	Gewicht	Größe	Gewicht
1	180	78	161	53
2	166	86	161	58
3	186	80	157	59
4	191	88	170	75
5	179	85	166	57
6	188	95	168	k.A.
7	175	70	166	65
8	186	77	175	61
9	180	86	168	62
10	190	90	170	61

richtig

Geschlecht	Größe	Gewicht
m	180	78
m	166	86
m	186	80
m	191	88
m	179	85
m	188	95
m	175	70
m	186	77
m	180	86
m	190	90
w	161	53
w	161	58
w	157	59
w	170	75
w	166	57
w	168	
w	166	65
w	175	61
w	168	62
w	170	61

4 Dateneingabe

Einlesen eines Excel-Datensatzes: atemwege.xls (Quelle: Open Data LMU München https://doi.org/10.5282/ubm/data.13)

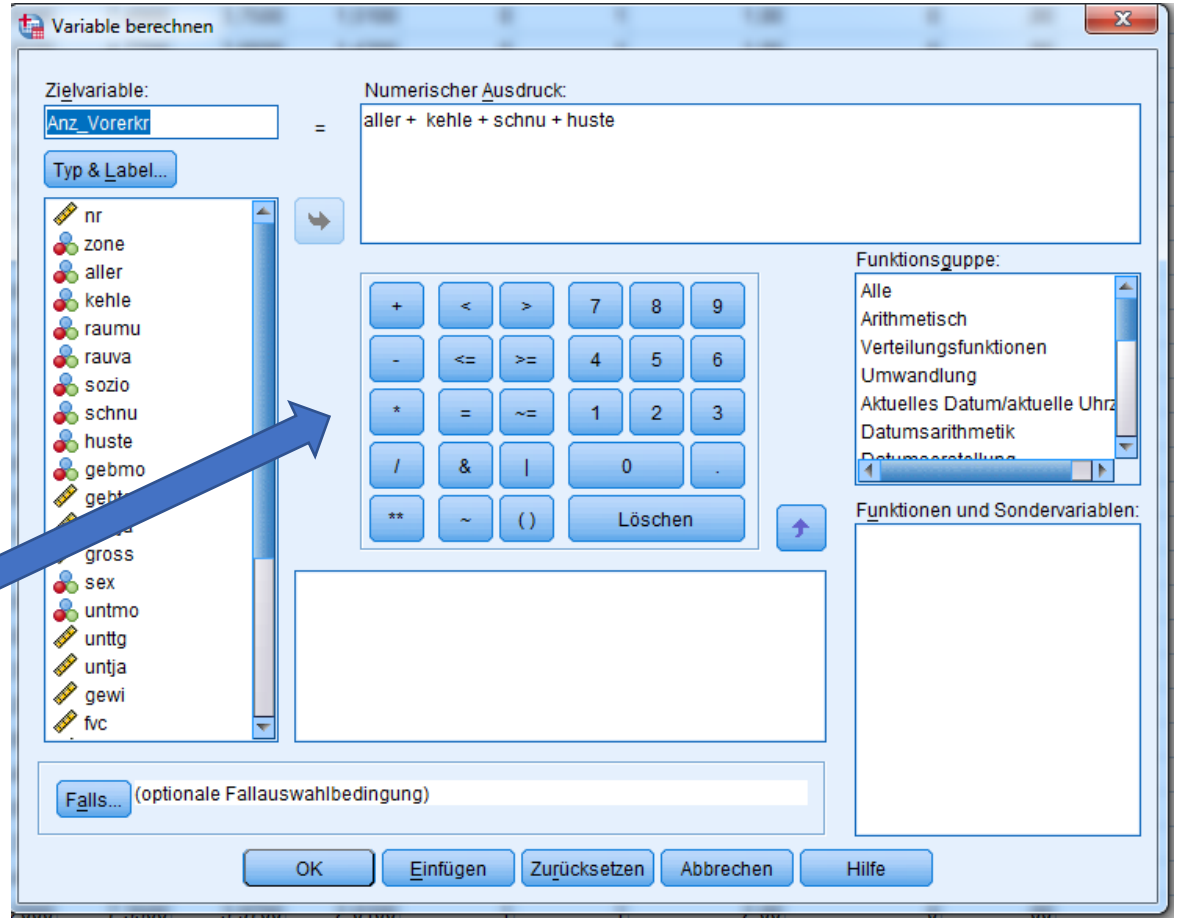
Datei→ Öffnen → Daten → Auswahl des **Dateityps**, hier .xls, die gewünschte Datei atemwege.xls kann ausgewählt werden.
Dann Abfrage, ob in erster Zeile die Variablennamen stehen (in unserem Fall ist das so).
Empfehlung: kurz visuell überprüfen, ob Daten sinnig in der Datenansicht angeordnet sind.

Dann vervollständigt man die Variablen- und Wertelabels gemäß den Angaben auf dem Beiblatt „atemwegeZettel.pdf“.

5 Datenbereinigung

- Menüpunkt **Transformieren** :
  →**Variable berechnen:** Man möchte eine Variable erhalten, die über die Anzahl der Vorerkrankungen (Allergie, Kehlkopfentzündung, Schnupfen, Husten) Auskunft gibt.

Verwendung der Tastatur zur Berechnung, hier Addition durch „+“ Zeichen

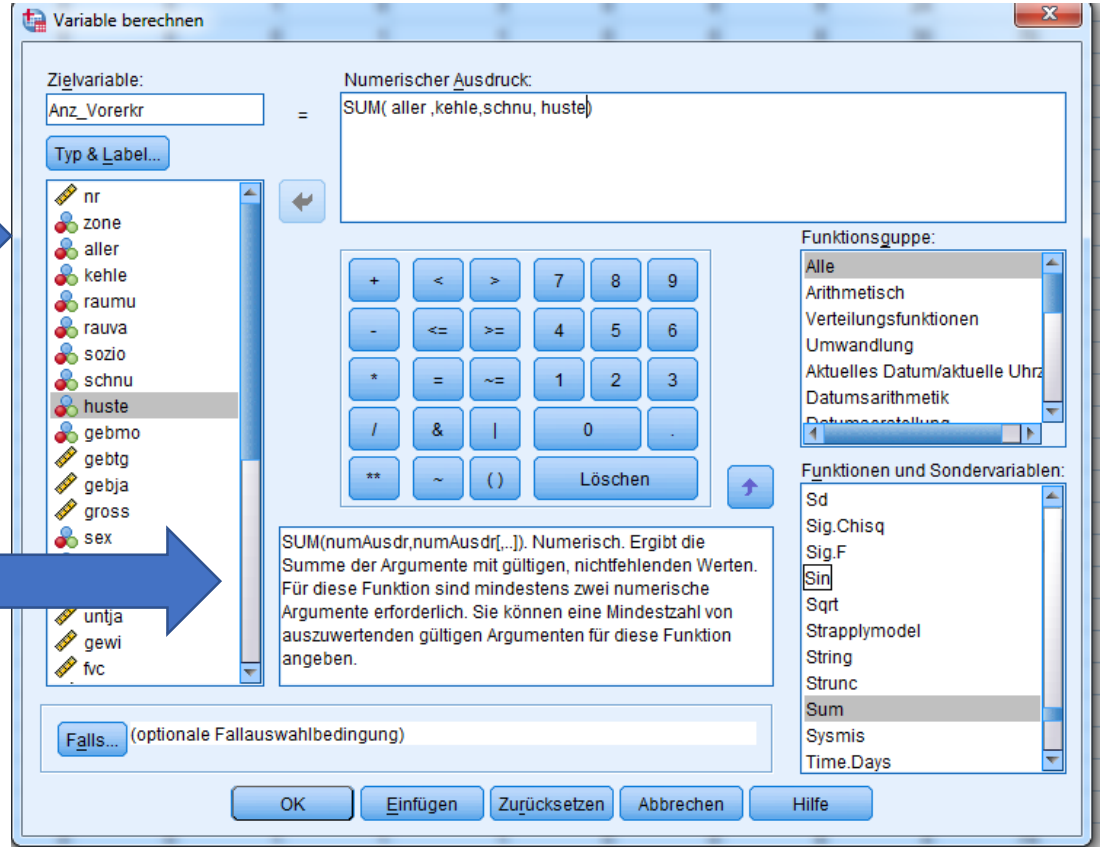


6 Datenbereinigung

Transformieren → Variable berechnen, auch Anzahl an Vorerkrankungen, nur etwas anderer Weg

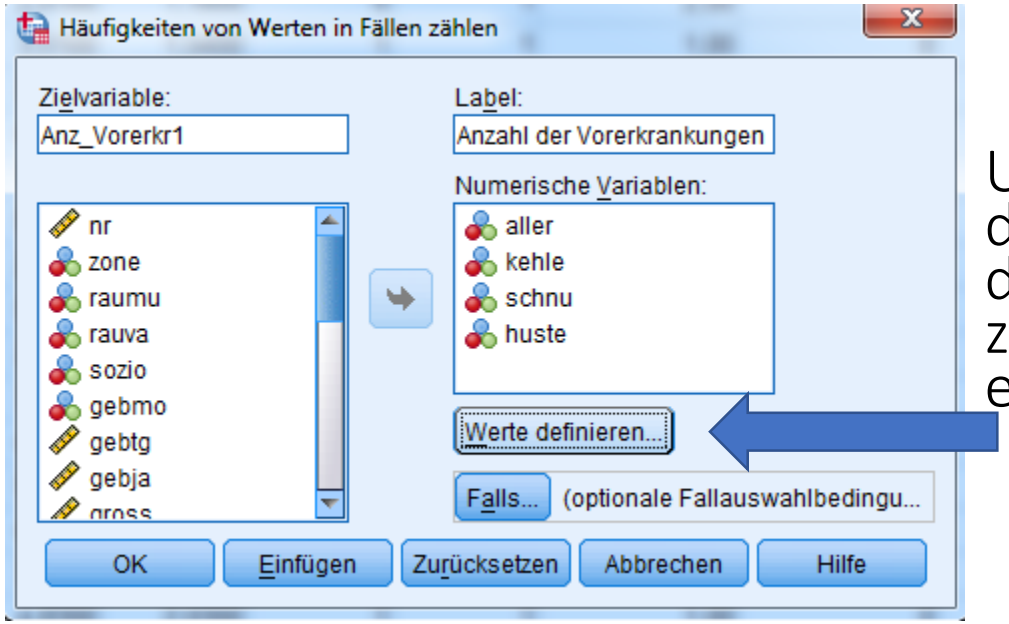
Exkurs: Umstellen von Variablennamen auf Beschriftung geht über rechte Maustaste

Verwendung einer Funktion (hier SUM)



7 Datenbereinigung

Immer noch Anzahl an Vorerkrankungen, aber über:
Transformieren → Werte in Fällen zählen



Unter dem Punkt „Werte definieren“ muss man dann noch den zu zählenden Wert „1“ eingeben.

Mit dieser Funktion kann man sich zeilenweise fehlende Werte oder jeden gewünschten Wert, auch Bereiche, durchzählen lassen!

8 Datenbereinigung

Klassifizieren von Werten in Gruppen geht über **Transformieren**

→ **Umcodieren in dieselben Variablen** (Ursprungsvariable wird überschrieben!)
→ **Umcodieren in andere Variablen** (neue, umcodierte Variable wird erzeugt)
→ **Automatisch Umcodieren** (SPSS teilt Daten automatisch in Gruppen ein)

→ **Visuelle Klassierung** (vorab Übersicht über Werte, Festlegung Anzahl Klassen, Auswahl, ob gleiche Wertintervalle oder ein balanciertes Design (gleiche Gruppengröße) gewünscht sind)

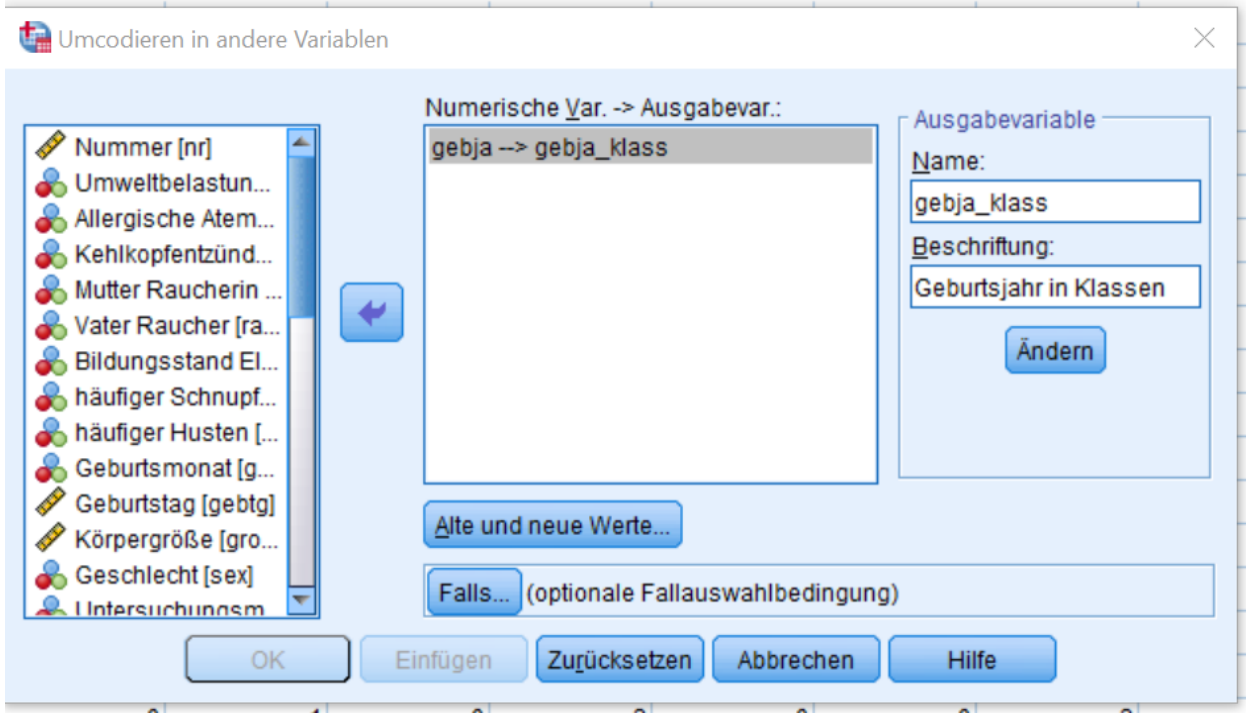
→ **Optimale Klassierung** (metrische Variablen werden unter Berücksichtigung einer nominalen Optimierungsvariable klassifiziert)

9 Datenbereinigung

Beispiel: Man möchte ausgehend vom Geburtsjahrgang (gebja) Analysen für zwei Altersklassen (alt(1) und jung(2)) durchführen:
1: Jahrgänge 73 – 77; 2: Jahrgänge 78 – 82

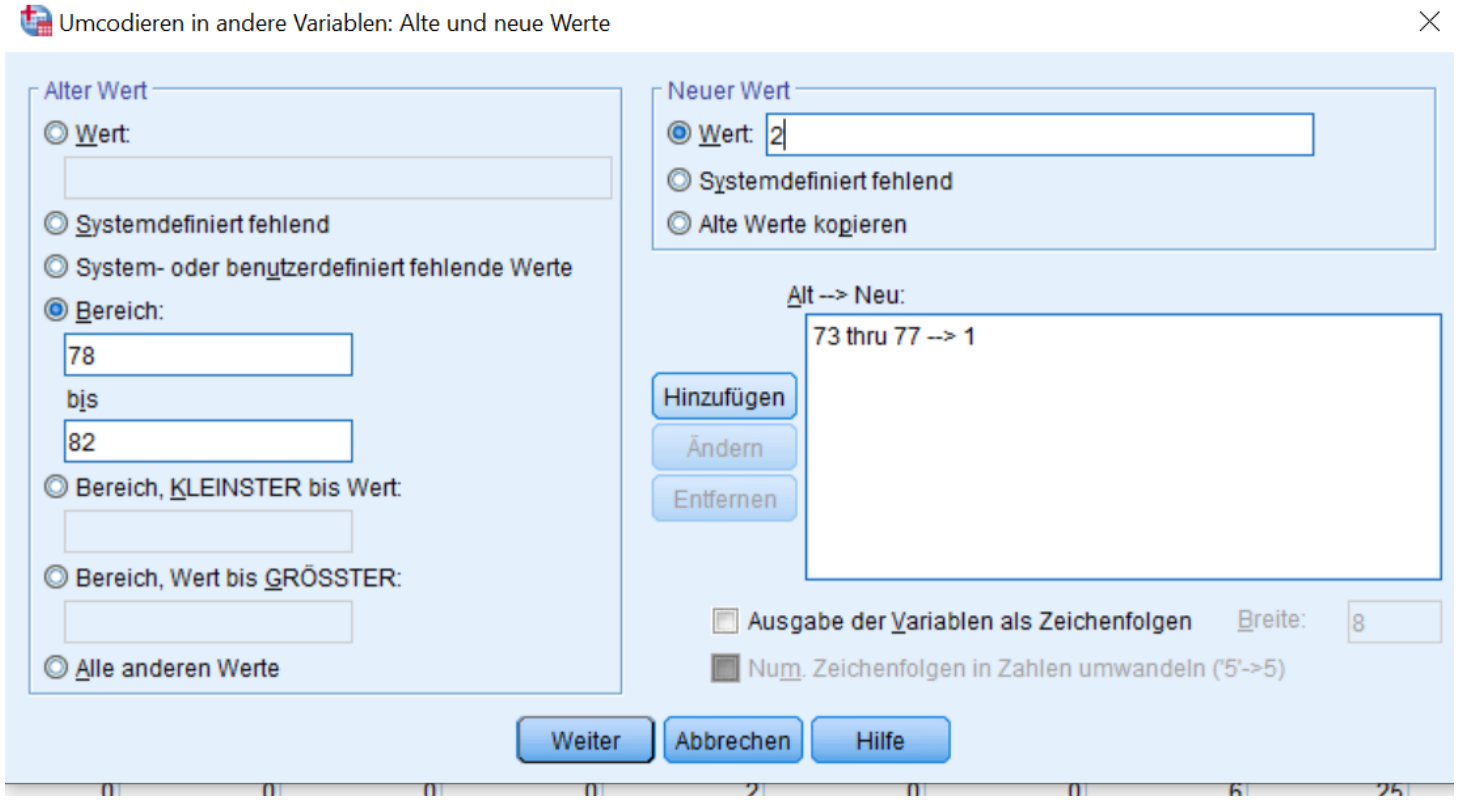
→ Umcodieren in andere Variablen

Angabe der Ausgangsvariable „gebja“ und der Ausgabevariable „gebja_klass“.
Dann Schaltfläche „Alte und neue Werte“.



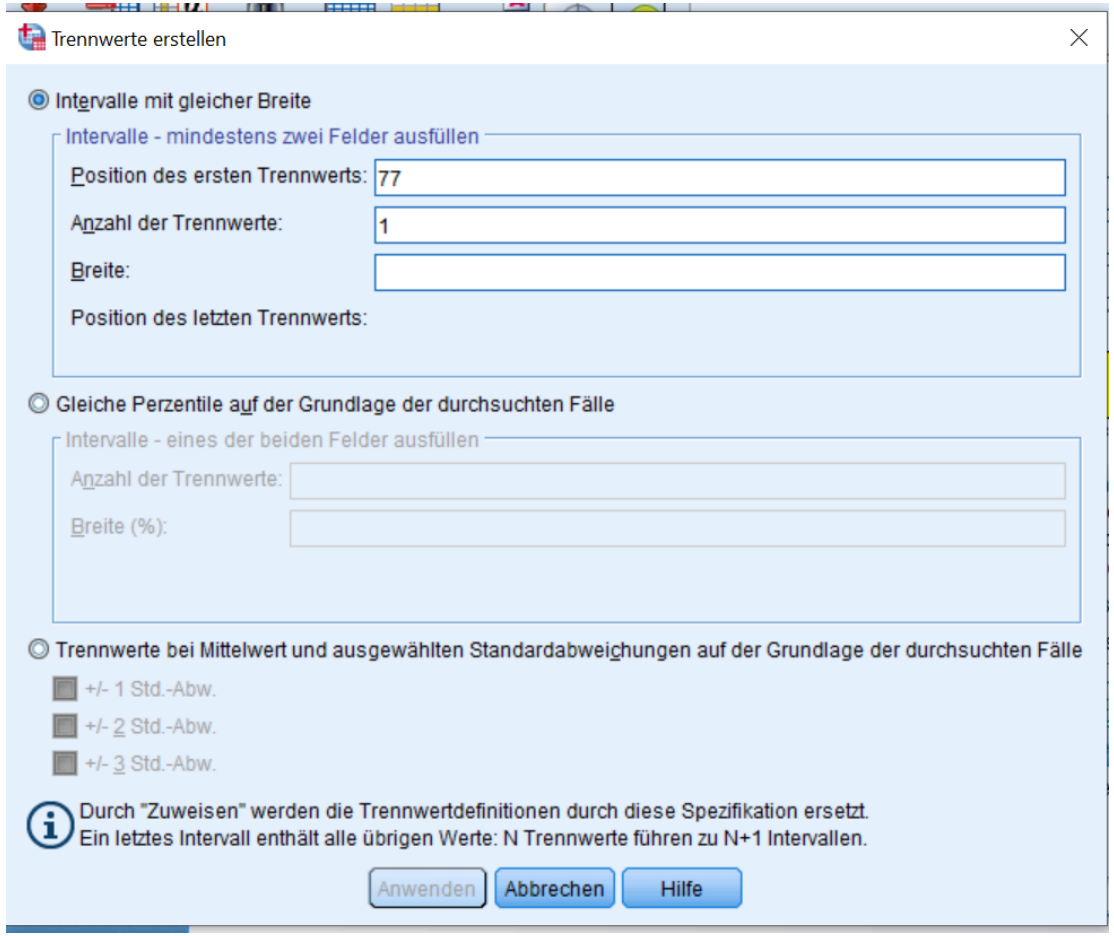
10 Datenbereinigung

Dann Angabe der Werte-bereiche wie rechts darge-stellt.
Durch „Hinzufügen“ werden die eingegebenen Zahlen in den großen weißen Kasten übertragen.
Dann auf „Weiter“.



11 Datenbereinigung

Das gleiche Ergebnis hätte man auch mit → **visuelle Klassierung** erreichen können:
Variable „gebja“ für die Klassierung aussuchen, auf „Weiter“.
Im nächsten Fenster unten rechts auf „Trennwerte erstellen“.
Wie rechts angegeben Werte eintragen.



12 Datenauswahl: Daten

Menüpunkt Daten:

- Man möchte in einer Analyse nur die Kinder früherer Jahrgänge („alt“) untersuchen. Von der selbst erstellten Variable gebja_klass sollen daher nur die Fälle „1“ ausgewählt werden.
  Daten → Fälle auswählen
  Beachten: dieser „Filter“ bleibt auch nach Deaktivierung als eigene Variable erhalten
- Die nicht ausgewählten Fälle „verschwinden“ (ab Version 28). Möchte man sie eingeblendet haben: **Bearbeiten → Ausgeschlossene Fälle ausblenden**, hier Häkchen entfernen. Sie sind dann nach dem Filter sortiert. Möchte man die alte Sortierung wieder haben:
  Daten → Fälle sortieren, Sortieren nach „Beobachtungsnummer“

Hilfe!

- Menüpunkt **Hilfe**

  → **Themen** öffnet online-Hilfe sortiert nach Themen

  → **SPSS Support** verbindet zu IBM support Seite

  → **SPSS Foren** verbindet zu IBM Seite mit verschiedenen Foren

  → **Dokumentation im PDF Format** Benutzerhandbuch

  → **Befehlssyntaxreferenz** Beschreibung der Syntaxsprache (PDF)

  → **Kompatibilitätsberichtstool** IBM Seite mit Produktsuche

- Hilfe zu Prozeduren: im Fenster findet sich eine Schaltfläche „Hilfe“, es öffnet sich online eine Beschreibung mit Beispielen.

Aufgaben/ Übung 1

1. Laden Sie sich den Datensatz atemwege.xls in SPSS hoch (Folie 4), vervollständigen Sie die Variablenbeschreibungen
2. *Menüpunkt Transformieren → Variable berechnen*: Erstellen Sie eine Variable „Anz_Vorerkr“ (Folien 5,6 oder 7)
3. Erzeugen Sie eine Variable, die die Differenz zwischen fef50 und fef75 angibt, Name: diff_fef, Berechnung: fef50 – fef75
4. *Menüpunkt Transformieren → Umcodieren in andere Variablen*: Erstellen Sie die Variable „gebja_klass“ (Folien 9-11)
5. Erzeugen Sie eine Variable geb_jz (Jahreszeit Geburt) ausgehend von der Variable gebmo:
   12, 1, 2 → Winter; 3-5 → Frühjahr; 6-8 → Sommer
   9-11 → Herbst
6. Setzen Sie einen Filter: wählen Sie die Fälle von männlichen Kindern aus, die als Anzahl an Vorerkrankungen (Anz_Vorerkr) mindestens 2 haben

14 Exkurs Syntax

Man möchte eine Variable über die örtliche Belastung des Kindes einführen, dabei wird berücksichtigt

- der Wohnort des Kindes (Variable **zone**, vorhandene Werte: 1= stark belastet, 2=eher weniger belastet, 3=hohe Ozonbelastung)
- ob die Mutter raucht (**raumu**, vorhandene Werte: 1=ja, 0=nein)
- ob der Vater raucht (**rauva**, vorhandene Werte: 1=ja, 0=nein)

Die neue Variable soll „belast“ heißen und soll folgende Werte annehmen können:

1: geringe Belastung (bei zone 2 raumu 0 rauva 0, 2 0 1, 2 1 0, 1 0 0, 3 0 0)

2: hohe Belastung (bei 2 1 1, 1 0 1, 1 1 0, 1 1 1, 3 0 1, 3 1 0, 3 1 1)

15 Exkurs Syntax

Es besteht die Möglichkeit, diese Variable über das Menü zu erzeugen (**Transformieren → Variable berechnen**), dies ist allerdings etwas umständlich! Empfehlenswert ist der Weg über den Syntax (**Datei → Neu → Syntax**):

```
IF (zone = 2 & raumu = 0 & rauva = 0)|(zone = 2 & raumu = 0 & rauva = 1)|
(zone = 2 & raumu = 1 & rauva = 0)|(zone = 1 & raumu = 0 & rauva = 0)| (zone
= 3 & raumu = 0 & rauva = 0)  belast=1.

IF (zone = 2 & raumu = 1 & rauva = 1)|(zone = 1 & raumu = 0 & rauva = 1)|
(zone = 1 & raumu = 1 & rauva = 0)|(zone = 1 & raumu = 1 & rauva=1)| (zone
= 3 & raumu = 1 & rauva = 0)|(zone = 3 & raumu = 0 & rauva = 1)| (zone = 3 &
raumu = 1 & rauva = 1)  belast=2.

EXECUTE.
```

Das logische „oder“ erhält man über die Tasten „alt gr“ und „<“
MAC: ALT +7

16 Exkurs Syntax

Vorteile

- zur Dokumentation der Analysen, z.B. bei Abschlussarbeiten
- Bei sich regelmäßig wiederholenden Analysen (Empfehlung: Kommentare schreiben)
- Implementierung von Python oder R-Code möglich

Nachteil

- es hat den Charakter einer Programmiersprache

Merke: Öffnen des Syntax-Editors über **Datei → Neu → Syntax**
Bei Prozeduren Drücken der Schaltfläche „Einfügen“
Kommentare werden mit einem * gekennzeichnet.

Datenbeschreibung: *Explorative Datenanalyse*

- Menüpunkt **Analysieren → Deskriptive Statistik → Explorative Datenanalyse**

Gibt Kennzahlen für eine Variable aus: Minimum, Maximum, Mittelwert, Median, Standardabweichung, KI,…

Gruppierung der Variable: Körpergröße, unterteilt nach Geschlecht: Variable „gross“ als abhängige Variable, Variable „sex“ in Faktorenliste. Im Fenster „Statistiken“ Häkchen bei den Perzentilen setzen.

Perzentile

			Perzentile						
		Geschlecht	5	10	25	50	75	90	95
Gewichtetes Mittel (Definition 1)	Körpergröße	männlich	120,25	124,00	131,25	142,00	153,00	166,00	172,00
		weiblich	118,00	121,00	128,00	138,00	151,00	160,40	164,00
Tukey-Angelpunkte	Körpergröße	männlich			131,50	142,00	153,00		
		weiblich			128,00	138,00	151,00		

18 Datenbeschreibung: *Pivot-Tabellen*

Pivot-Tabellen

- Die Formatvorlage kann verändert werden unter **Bearbeiten→ Optionen** „Pivot-Tabellen“.
- Die Pivot-Tabelle an sich kann durch Doppelklicken im Pivot-Tabellen-Editor bearbeitet werden, z.B. reduziert werden zu

Perzentile

		Perzentile		
	Geschlecht	25	50	75
Körpergröße	männlich	131,25	142,00	153,00
	weiblich	128,00	138,00	151,00

Beachten: beim Löschen von Spalten oder Reihen aus einer Tabelle immer die Werte markieren und löschen, nicht die Spalten- oder Reihenbeschriftungen!

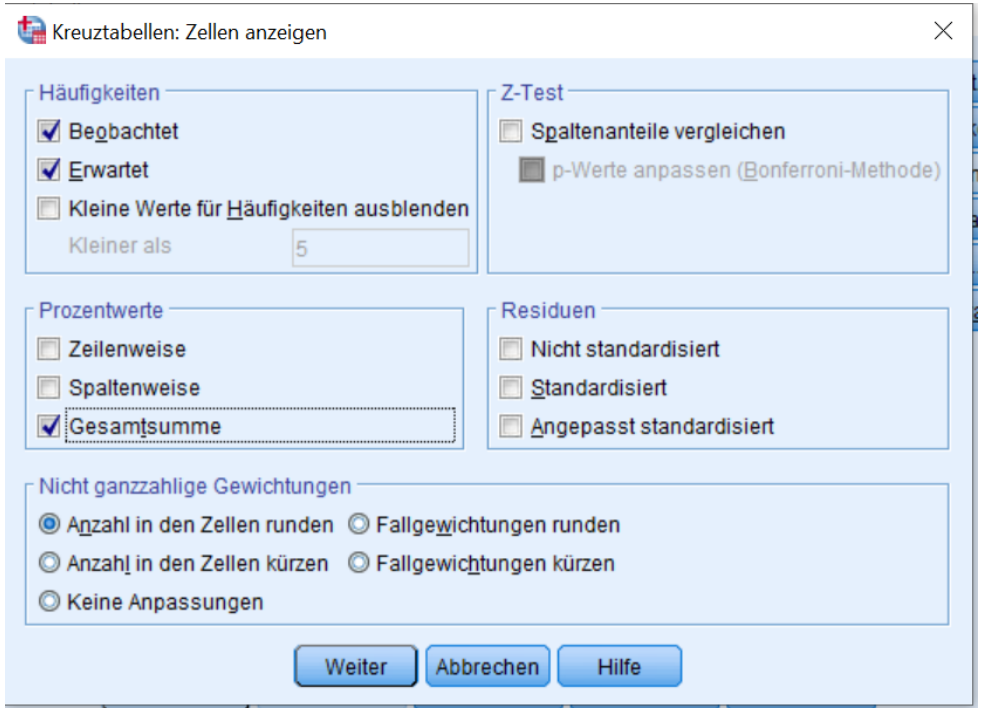
19 Datenbeschreibung: *Kreuztabellen*

Beispiel: Besteht ein Einfluss der Umweltbelastung am Wohnort auf die Anzahl an Vorerkrankungen?

Analysieren → Deskriptive Statistik → Kreuztabellen

Eingabe: Zeile: zone
Spalte: Anz_Vorerkr

Schaltfläche „Zellen“ , s. Bild
→ Häufigkeiten „Beobachtet“ und „Erwartet“
→ Prozentwerte „Gesamtsumme“



Datenbeschreibung: *Kreuztabellen*

Umweltbelastung am Wohnort * Anz_Vorerkr Kreuztabelle

			Anz_Vorerkr					
			,00	1,00	2,00	3,00	4,00	Gesamt
Umweltbelastung am Wohnort	stark belastet	Anzahl	112	45	17	9	3	186
		Erwartete Anzahl	120,8	42,4	16,1	5,8	1,0	186,0
		% der Gesamtzahl	7,2%	2,9%	1,1%	0,6%	0,2%	12,0%
	eher weniger belastet	Anzahl	389	150	60	21	0	620
		Erwartete Anzahl	402,7	141,3	53,6	19,2	3,2	620,0
		% der Gesamtzahl	25,1%	9,7%	3,9%	1,4%	0,0%	40,0%
	erhöhte Ozonbelastung durch Höhenlage	Anzahl	505	158	57	18	5	743
		Erwartete Anzahl	482,5	169,3	64,3	23,0	3,8	743,0
		% der Gesamtzahl	32,6%	10,2%	3,7%	1,2%	0,3%	48,0%
Gesamt		Anzahl	1006	353	134	48	8	1549
		Erwartete Anzahl	1006,0	353,0	134,0	48,0	8,0	1549,0
		% der Gesamtzahl	64,9%	22,8%	8,7%	3,1%	0,5%	100,0%

21 Datenbeschreibung: *Kreuztabellen*

Chi-Quadrat-Test: Schaltfläche „Statistiken“, Häkchen bei „Chi-Quadrat“ setzen

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (zweiseitig)
Chi-Quadrat nach Pearson	16,198[a]	8	,040
Likelihood-Quotient	17,623	8	,024
Zusammenhang linear-mit-linear	7,019	1	,008
Anzahl der gültigen Fälle	1549		

a. 3 Zellen (20,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist ,96.

Der p-Wert ist 0,04, also <0,05, d.h. es besteht ein signifikanter Einfluss der Umweltbelastung am Wohnort auf die Anzahl der Vorerkrankungen.

Beachten: die Meldung, dass 20% der Zellen eine erwartete Häufigkeit <5 haben. Dieses kann (muss aber nicht) zu Verfälschungen führen und sollte vermieden werden!

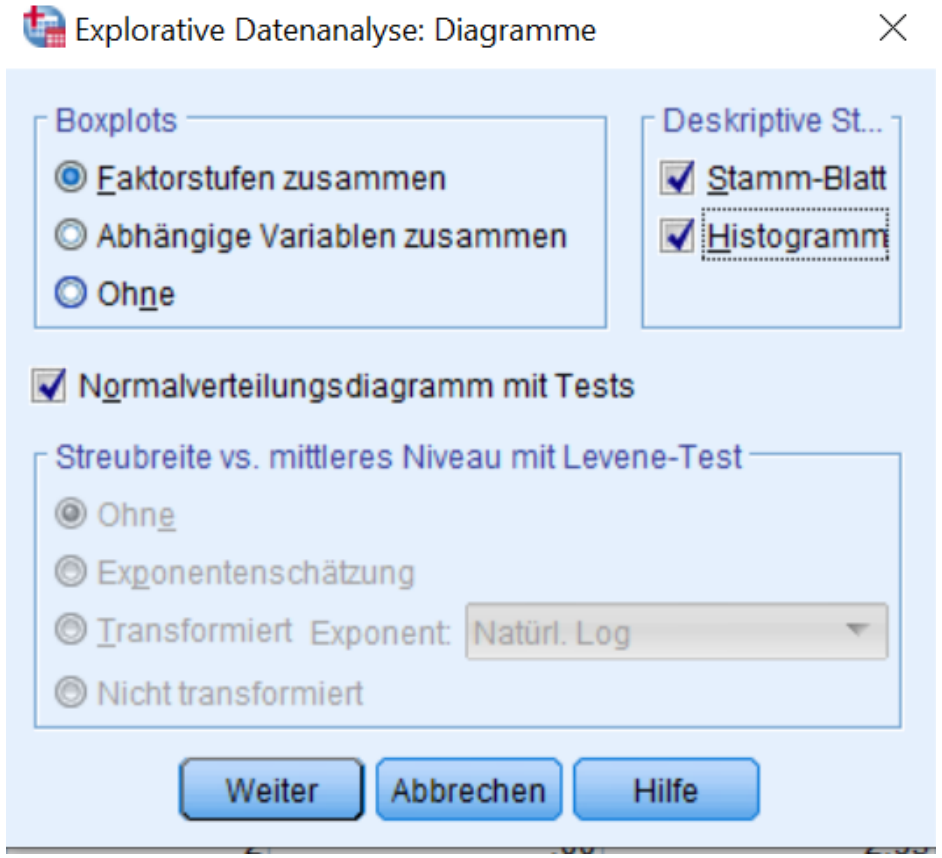
Datenbeschreibung: *Untersuchung der Normalverteilung*

Teil 1

Analysieren → Deskriptive Statistik → Explorative Datenanalyse

Macht nur für metrische Variablen Sinn!

Schaltfläche „Diagramme“



Datenbeschreibung: *Untersuchung der Normalverteilung*

Visuelle Beurteilung der Diagramme auf NV (s. auch nächste Folie)
Ggf. auch visuelle Beurteilung der Varianzhomogenität (Boxplots)

Hinzuziehen des Testergebnisses (Kolmogorov-Smirnov bzw. Shapiro-Wilk (geeigneter bei kleinerem n)). Ein p-Wert <0.2 (bei KS) bzw. <0.05 (SW) weist auf Abweichung von der NV hin.

Ggf. noch den Levene-Test mit auswählen, wenn man die Varianzhomogenität zwischen Gruppen untersuchen möchte, hier weist ein p-Wert <0.2 auf Varianzheterogenität hin.

Datenbeschreibung: *Untersuchung der Normalverteilung*

Faustregeln:

Bei n<10 sollte man generell nicht von einer Normalverteilung ausgehen.

Bei sehr großen n (>50) darf man parametrische Verfahren anwenden ohne Untersuchung der NV (wenn Daten unabhängig voneinander und identisch verteilt sind) → zentraler Grenzwertsatz.

Visuelle Beurteilung: bei höherer Fallzahl kritischer sein (Testverfahren berücksichtigen das!)

Ausreißer: überprüfen (Tippfehler?), bei vielen Ausreißern eher nicht-parametrisches Verfahren wählen.

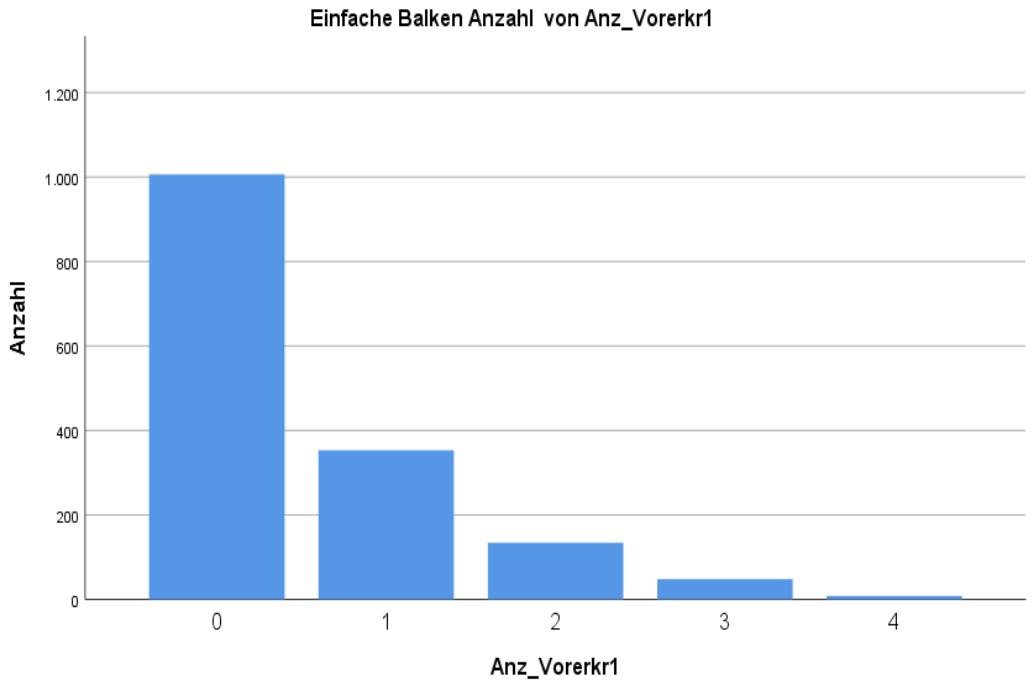
25 Diagrammerstellung

Die Anzahl an Vorerkrankungen soll als Balkendiagramm dargestellt werden.

Grafik→ Diagrammerstellung

Per Drag and Drop gewünschte Diagrammart (Einfache Balken) und die Variable (Anz_Vorerkr1 auf die x-Achse) in das Vorschaufeld ziehen.

<u>Bearbeiten</u> des Diagramms
→ doppelt klicken



Übung 2

1. Zeigen Sie grafisch und über eine Kreuztabelle, dass kein deutlicher Unterschied zwischen den alten und jungen Kindern (Variable gebja_klass) hinsichtlich der Anzahl an Vorerkrankungen besteht.
2. Geben Sie an, wie häufig allergische Atemwegserkrankungen auftreten, wie häufig Kehlkopfentzündung und wie häufig beides zusammen auftritt.
3. Die Kinder sollen in drei Größengruppen eingeteilt werden. Die Grenzen setzt man aber je nach Geschlecht unterschiedlich an! Neuer Name der Variable: gross_klass, s. nächste Seite

Übung 2

3ff. 1 = „klein“ = Größe unterhalb des 25% Quartils

2 = „mittel“ = Größe zwischen 25% und 75% Quartil

3 = „groß“ = Größe oberhalb des 75% Quartils

Die Quartile für die Größe der Mädchen sind
25: 128,00 75: 151,00
für die der Jungen
25: 131,25 75: 153,00

4. Stellen Sie den Bildungsstand der Eltern als Balkendiagramm dar. Verändern Sie das Muster und die Dicke der Balken, verringern Sie den y-Achsenabschnitt. Fügen Sie eine Anmerkung ein. Versuchen Sie, dieses Diagramm mit „rauva“ als Unterteilungsvariable zu erstellen.