Zu Abschnitt 4.1.3: Häufigkeitsveränderungen auf Populationsebene

Beispiel 2: Ist der Anstieg der HIV-Inzidenz zwischen 2012 und 2013 signifikant?

Das folgende Beispiel basiert nicht auf echten, sondern auf fiktiven Daten. Im Gesundheitsamt der Stadt K. wurden im vergangenen Jahr 2.000 AIDS-Tests durchgeführt, im Jahr darauf waren es 2.500. Tabelle 4.1 zeigt die absolute Verteilung der Diagnosen (positiv oder negativ) in beiden Jahren.

	2012	2013	Randsumme
HIV-negativ	1.950 *(97,5 %)*	2400 *(96 %)*	4.350
HIV-positiv	50 *(2 %)*	100 *(4 %)*	150
Randsumme	2.000	2.500	4.500

Auf den ersten Blick sieht man, dass die Anzahl der „positiv"-Diagnosen zwischen den beiden Jahren zugenommen hat, und zwar sowohl absolut (von 50 auf 100) als auch relativ (von 2,5% auf 4%). Ist diese Zunahme statistisch bedeutsam?

Diese Frage kann mit Hilfe eines Vierfelder-χ^2-Tests (s. Eid et al., 2013; Abschn. 11.4.1) beantwortet werden. Der Wert der Prüfgröße berechnet sich wie folgt:

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{k} \frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}}$$

mit

- $i \in \{1,...,p\}$ Kategorien in den Zeilen der Tabelle (hier gibt es $p = 2$ Kategorien: $i = 1$ für HIV-negativ, $i = 2$ für HIV-positiv).
- $j \in \{1,...,k\}$ Kategorien in den Spalten der Tabelle (hier gibt $k = 2$ Kategorien: $j = 1$ für 2012, $j = 2$ für 2013).
- n_{ij} = Beobachtete Häufigkeit in einer Zelle ij,
- e_{ij} = Erwartete Häufigkeit in einer Zelle ij unter der Nullhypothese

Die erwarteten Häufigkeiten unter der Nullhypothese (Unabhängigkeit der beiden Merkmale „Jahr“ und „Diagnose“) berechnen sich wie folgt:

$$e_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$$

wobei $n_{i\bullet}$ für die Spaltensumme innerhalb einer Zeile *i* und $n_{\bullet j}$ für die Zeilensumme innerhalb einer Spalte *j* steht.
Die Nullhypothese besagt, dass es keine Zunahme der HIV-positiv-Diagnosen zwischen den beiden Jahren gab. Das bedeutet aber nicht, dass die absoluten Häufigkeiten der beiden Diagnosen (positiv und negativ) sich zwischen den beiden Jahren nicht unterscheiden! Das ist schon deshalb unplausibel, da im Jahre 2013 ja 500 Tests mehr durchgeführt wurden als im Jahre 2012. Lediglich das Verhältnis von Positiv- zu Negativdiagnose soll sich nicht verändern. Demnach lauten die erwarteten Häufigkeiten unter der Nullhypothese wie folgt:

	2012	2013	Randsumme
HIV-negativ	1.933,33 *(96,67 %)*	2416,67 *(96,67 %)*	4.350
HIV-positiv	66,67 *(3,33 %)*	83,33 *(3,33 %)*	150
Randsumme	2.000	2.500	4.500

Berechnet man nun aus den beobachteten und den erwarteten Zellhäufigkeiten den Wert der Prüfgröße χ^2, so erhält man $\chi^2 = 7{,}76$. Die Quantile unter der χ^2-Verteilung hängen von der Anzahl der Freiheitsgrade (*df*) ab. Diese berechnen sich zu

$$df = (p - 1) \cdot (k - 1)$$

In unserem Fall hat die χ^2-Verteilung also $df = 1$ Freiheitsgrad. Die Wahrscheinlichkeit für diesen (oder jeden größeren) χ^2-Wert beträgt etwa $p = 0{,}006$. Dies bedeutet: Die Abweichung der empirischen von der erwarteten Verteilung ist signifikant. Man kann sagen, dass im Jahre 2013 die Anzahl der HIV-positiv-Diagnosen tatsächlich signifikant gegenüber 2012 angestiegen ist.

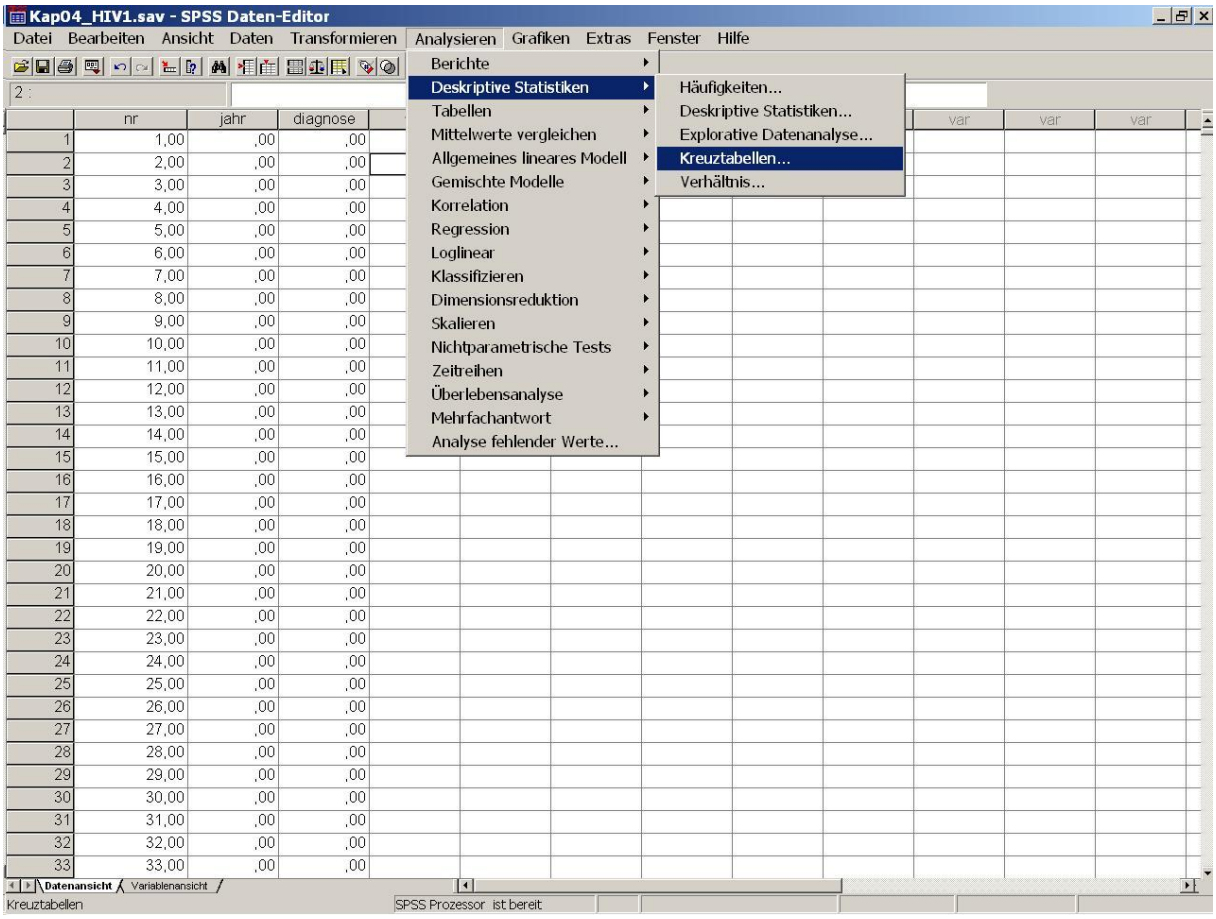
Falls Sie dieses Beispiel selbst mit SPSS ausrechnen wollen:
Für dieses Beispiel liegen auch Originaldaten vor (`KAP04_HIV1.SAV`). In dieser Datei finden Sie drei Variablen:

- eine Variable „nr“: Laufende Nummer für jede getestete Person (von 1-4500)
- eine Variable „jahr“, welche das Jahr indiziert, in dem der Test gemacht wurde (codiert mit 1 = 2012 und 2 = 2013) und
- eine Variable „diagnose“, welche das Ergebnis des HIV-Tests indiziert (codiert mit 0 = negativ und 1 = positiv).

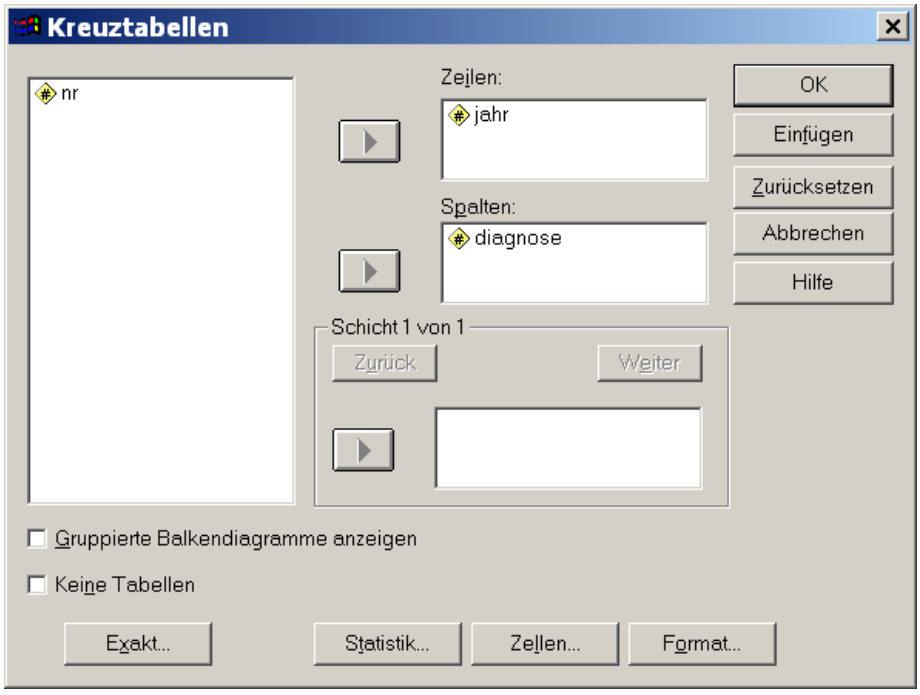
Die Syntax für den χ^2-Test lautet:

```
CROSSTABS
 /TABLES=jahr  BY diagnose
 /STATISTIC=CHISQ.
```

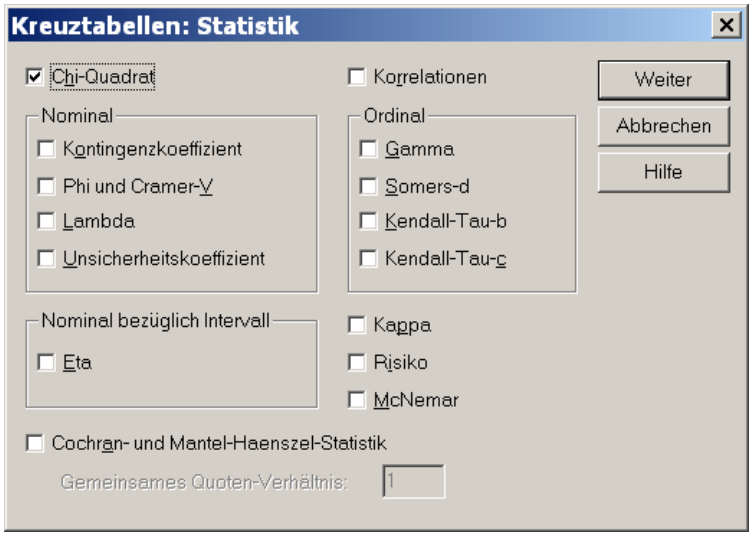
Wenn Sie den Test über die Menüführung rechnen wollen, klicken Sie im Menü „Analysieren“ auf „Deskriptive Statistiken“ und anschließend „Kreuztabellen“ (s. Screenshot):



Es öffnet sich ein Dialogfenster mit dem Titel „Kreuztabellen“. Schieben Sie die Variable „jahr“ (mit Hilfe der Maus oder mit Hilfe der Pfeilspitzen) in das Feld mit dem Titel „Zeilen“; schieben Sie die Variable „diagnose“ in das Feld mit dem Titel „Spalten“.



Klicken Sie dann auf den Button „Statistik ...“. Es öffnet sich ein neues Dialogfenster. Aktivieren Sie dort das Klickkästchen „Chi-Quadrat“.



Klicken Sie nun auf „Weiter“ und anschließend auf „OK“.
Nun öffnet sich die SPSS-Ausgabe. Es werden drei Tabellen ausgegeben.

Verarbeitete Fälle

	Fälle					
	Gültig		Fehlend		Gesamt	
	N	Prozent	N	Prozent	N	Prozent
jahr * diagnose	4500	100,0%	0	,0%	4500	100,0%

In Tabelle 1 sind lediglich die Anzahl der verarbeiteten Fälle abgetragen.

jahr * diagnose Kreuztabelle

Anzahl

		diagnose		
		,00	1,00	Gesamt
jahr	,00	1950	50	2000
	1,00	2400	100	2500
Gesamt		4350	150	4500

In Tabelle 2 sind die beobachteten Häufigkeiten abgetragen. Sie entsprechen den Werten aus Tabelle 4.1 im Buch.

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)	Exakte Signifikanz (2-seitig)	Exakte Signifikanz (1-seitig)
Chi-Quadrat nach Pearson	7,759[b]	1	,005		
Kontinuitätskorrektur[a]	7,300	1	,007		
Likelihood-Quotient	7,955	1	,005		
Exakter Test nach Fisher				,006	,003
Zusammenhang linear-mit-linear	7,757	1	,005		
Anzahl der gültigen Fälle	4500				

a. Wird nur für eine 2x2-Tabelle berechnet

b. 0 Zellen (,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 66,67.

In Tabelle 3 finden Sie (unter anderem) den Wert der Prüfgröße („Chi-Quadrat nach Pearson“, hier: $\chi^2 = 7{,}76$), die Freiheitsgrade ($df = 1$) und den p-Wert ($p = 0{,}006$).

Falls Sie dieses Beispiel mit einem anderen Programm ausrechnen wollen:
Die Originaldaten für dieses Beispiel sind auch in einem allgemein lesbaren ASCII-Format abgespeichert `(KAP04_HIV1.DAT)`.