



Universität Trier

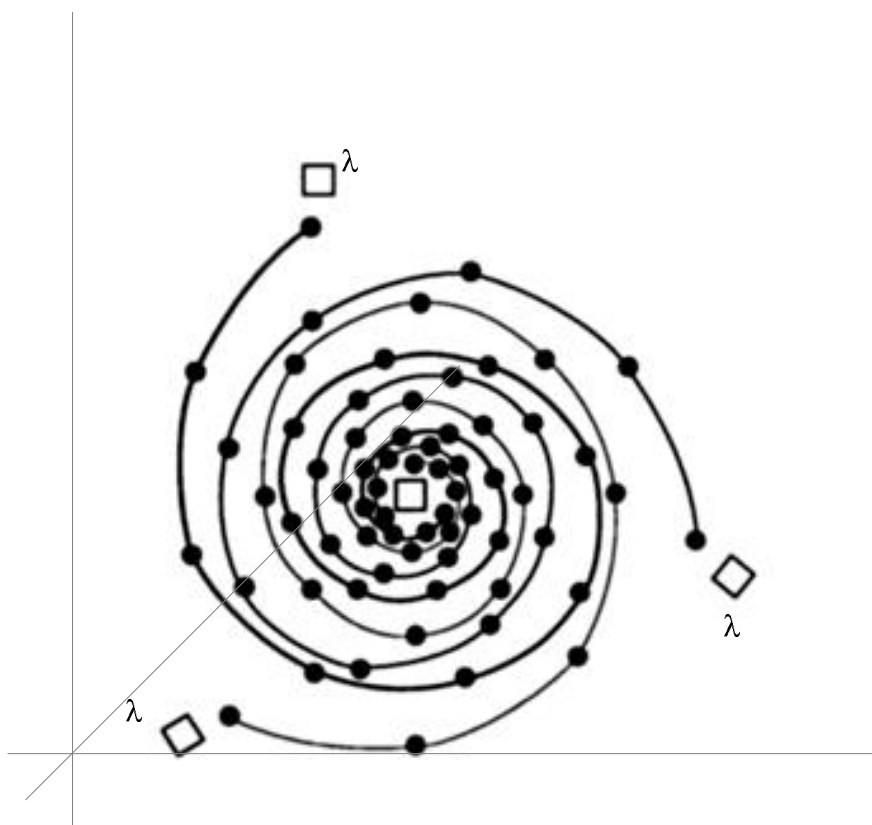
15.06.2002

Fachbereich I Psychologie

Skript zur *Multivariaten Statistik*

Thema: *Log-lineare Modelle*

LOG-LINEARE MODELLE



Inhaltsverzeichnis

1. Einführung.....	4
2. χ^2 -Tests.....	5
3. Herleitung des log-linearen Modells und dessen Parameter.....	7
3.1 Das multiplikative Ursprungs-Modell.....	7
3.2 Das 'additive' log-lineare Modell.....	8
3.3 Restriktionen.....	9
3.4 Submodelle.....	10
3.5 Konditionale und partielle Effekte.....	11
4. Schätzung der Erwartungshäufigkeiten.....	13
5. Signifikanztests.....	13
5.1 Testen eines Modells: Goodness of fit.....	14
5.2 Hierarchische Modelltests.....	15
5.3 Testen eines Modells gegen ein anderes.....	16
5.4 Einzeltests.....	17
6. Interpretation der Effekte.....	18
7. Probleme & Fallen im Umgang mit log-linearen Modellen.....	19
7.1 Alpha-Kumulierung.....	19
7.2 Stichprobengröße.....	19
7.3 Post-hoc Teststärke.....	19
7.4 Leere Zellen.....	20
7.5 Zu kleine Erwartungshäufigkeiten.....	21
7.6 Nichtlineare Effekte.....	21
7.7 Vorgehensweise bei der Erhebung.....	21
8. Logit-Modelle.....	22
9. Odds und Prozentsatzdifferenzen.....	24
9.1 Odds.....	24
9.2 Prozentsatzdifferenzen.....	26
9.3 Vergleich: Odds, Prozentsatzdifferenzen & λ	27
10. Aufgaben.....	27
10.1 Aufgabe 1.....	27
10.1.1 Tipps.....	28
10.1.2 Lösungen.....	29
10.1.3 Abschließende Bemerkung zu Aufgabe 1.....	29
10.2 Aufgaben 2.....	30
10.2.1 Lösungen.....	30
11. Literatur.....	31

Vorbemerkungen

Dieses Skript kann eine vertiefende Lektüre der Originalliteratur leider nicht vollständig ersetzen. Abschnitt 3.5 beispielsweise hat klare Schwächen, die auszumerzen ich (noch) nicht in der Lage bin. Leider tut man sich mit Literatur-Empfehlungen recht schwer. Der ANDREß, HAGENAARS & KÜHNEL (1997) ist zwar recht verständlich geschrieben, geht aber schon ein wenig ins Eingemachte, und alles Übrige mag ich wirklich nicht empfehlen.

Die Abschnitte 5.2 und 7.7 beziehen sich noch spezifisch auf Prüfungsliteratur von Eid (LANGEHEINE 1986). Auch die mathematische Notation orientiert sich sehr stark an LANGEHEINE. Kapitel 9 ist dagegen speziell aus ANDREß, HAGENAARS & KÜHNEL (1997) entnommen.

Und was die Abbildung auf der Titelseiten angeht, die hat mit dem Thema insofern etwas zu tun, als deren Spiral-Arme auf einer logarithmischen Funktion beruhen. Vielleicht war's keine gute Idee, aber ich wollte, dass es ein wenig freundlicher aussieht.

Falls Fehler im Skript auftauchen mögen, bin ich um Korrekturvorschläge zwar dankbar, ansonsten entziehe ich mich aber jeglicher Verantwortung. Dazu kann ich gleich anmerken, dass alle Versionen, die vor dem 12. Juli erstellt wurden, noch einen ganz tiefgreifenden Fehler enthielten in Kapitel 3. Und sogar Versionen vor dem 8. August würde ich heute nicht mehr empfehlen.

1. Einführung

Häufig tritt in den Sozialwissenschaften der Fall ein, dass Variablen nur im nominalen Skalenniveau vorliegen: Geschlecht (m, w), Familienstand (ledig, verheiratet, ...), Hautfarbe (schwarz, weiß, ...), EU-Bürger (Ja, Nein). Um Haupteffekte und Wechselwirkungen zwischen nominalen Variablen zu untersuchen bedient man sich der log-linearen Modelle.

Die log-linearen Modelle könnte man beschreiben als Allgemeines lineares Modell (ALM, BORTZ 1993 Kap. 14) für kategoriale Daten. Allerdings gibt es auch wesentliche Unterschiede zwischen den beiden Modellen. So konnte STEYER (1983) zeigen, dass die Randomisierung im log-linearen Modell eine ganz andere Funktion hat, als im allgemeinen linearen. Die Zielsetzung ist aber bei beiden doch dieselbe: Die Beschreibung von Haupt- und Wechselwirkungseffekten.

Um vorerst einige Grundbegriffe und bereits einige Variablennamen einzuführen, dient die folgenden *Kontingenztafel* (auch *Kreuztabelle* genannt).

Tabelle 1

		Geschlecht		
		w=1	m=2	
Uni-Karriere	Ja=1	n_{11}	n_{12}	$n_{1.}$
	Nein=2	n_{21}	n_{22}	$n_{2.}$
		$n_{.1}$	$n_{.2}$	$n_{..}$

Diagrammische Erläuterungen der Tabelle:

- „Rand“-verteilung (12) zeigt auf die Spaltenüberschriften w=1 und m=2.
- Randverteilung (1) zeigt auf die Zeilenüberschriften Ja=1 und Nein=2.
- Gesamt-n zeigt auf die Summe $n_{..}$.
- Randverteilung (2): z.B. $n_{.1} = n_{11} + n_{21}$ zeigt auf die Spaltensummen $n_{.1}$ und $n_{.2}$.

In den Zellen stehen die *beobachteten Häufigkeiten* n_{ij} . Davon zu unterscheiden sind die *Erwartungshäufigkeiten* e_{ij} . Erwartungshäufigkeiten sind diejenigen Häufigkeiten, die dort stehen müssten, wenn die Nullhypothese erfüllt wäre. Dazu gibt es noch die Wahrscheinlichkeiten p_{ij} , die ebenfalls bei Annahme der Nullhypothese gelten. Sie geben an, wie groß die Wahrscheinlichkeit ist, dass ein zufälliger Proband in die entsprechende Zelle aufgenommen wird. Die Indizierung ist für alle drei Angaben gleich.

Die Variablennamen sind bei verschiedenen Autoren leider recht unterschiedlich. Man kann nicht davon sprechen, dass es eine Konvention gäbe, an die sich die meisten Autoren hielten. Wenn jemand daher auf die Originalliteratur zurückgreift, sollte er sich zuerst vergewissern, wie die Variablen im Einzelnen benannt sind. Darüber hinaus unterscheiden die Autoren (korrekterweise) Populationsvariablen und geschätzte Variablen. Da man eine Erwartungshäufigkeit e_{ij} nicht direkt beobachten kann, muss man sie schätzen. Diesen Schätzer würde man mit dem Bezeichner \hat{e}_{ij} vom Populationsparameter e_{ij} unterscheiden können. Da diese Darstellung aber mehr Wert auf ein

Verständnis der Zusammenhänge legt und nicht so sehr an mathematischen Details interessiert ist, verzichte ich auf diese Unterscheidung.

2. χ^2 -Tests

Um das Prinzip log-linearer Modelle noch ein wenig mehr zu veranschaulichen, möchte ich an die χ^2 -Tests zur Analyse von kategorialen Daten (BORTZ 1993, S. 147ff) erinnern. Sie wurden bereits im 1. oder 2. Semester angesprochen und sollen hier aus didaktischen Gründen am Beispiel eines 4-Felder- χ^2 noch einmal wiederholt werden.

Tabelle 2 – fiktives Beispiel

		Geschlecht		
		w=1	m=2	
Uni-Karriere	Ja=1	20	15	35
	Nein=2	50	15	65
		70	30	100

Wir untersuchen den Zusammenhang zwischen dem Geschlecht und der Frage, ob sich jemand vorstellen kann, an der Universität Karriere zu machen. Die Frage lautet also, gibt es einen Zusammenhang zwischen dem Geschlecht und der Absicht einer Uni-Karriere. Die Haupteffekte (Unterschiede in den beiden Randverteilungen) sind dabei nicht weiter interessant. Dass es beispielsweise mehr Frauen als Männer in der Kreuztabelle gibt, lässt lediglich darauf schließen, dass die Befragung bei Psychologiestudenten durchgeführt wurde. Was wir suchen, ist die Interaktion¹ zwischen den beiden Variablen Geschlecht und Uni-Karriere.

Die Randverteilung können wir daher als gegeben betrachten. Wir können auf sie zurückgreifen, um die nötigen Wahrscheinlichkeiten zu schätzen. Die Wahrscheinlichkeit, dass ein Proband nur aufgrund der Haupteffekte in einer bestimmten Zelle landet. Wir schätzen also die Wahrscheinlichkeiten, die im Falle der Nullhypothese (keine Assoziation) gelten müssten.

Die Wahrscheinlichkeit weiblich zu sein beträgt

$$p_{i1} = \frac{n_{.1}}{n_{..}} = \frac{70}{100} = 0.70 \quad (\text{Formel 1})$$

und diejenige männlich zu sein

$$p_{i2} = \frac{n_{.2}}{n_{..}} = 1 - p_{.1} = 1 - 0.70 = 0.30 \quad (\text{Formel 2})$$

Entsprechendes gilt für die Wahrscheinlichkeiten, sich eine Uni-Karriere (nicht) zuzutrauen.

1 Bei kategorialen Daten spricht man bei Interaktionen zwischen *genau zwei* Variablen auch häufig von Assoziation.

$$p_{1.} = \frac{n_{1.}}{n} = \frac{35}{100} = 0.35 \quad (\text{Formel 3})$$

$$p_{2.} = \frac{n_{2.}}{n} = \frac{65}{100} = 0.65 \quad (\text{Formel 4})$$

Aus diesen Wahrscheinlichkeiten können nun die Wahrscheinlichkeiten für die einzelnen Zellen berechnet werden (Formel 5). Das Ergebnis ist in Tabelle 3 festgehalten, zusammen mit den Erwartungshäufigkeiten, die sich leicht aus den Wahrscheinlichkeiten ergeben (Formel 6). Die Erwartungshäufigkeiten geben an, welche Zahlen in den einzelnen Zellen stehen müssten, wenn die Nullhypothese gilt.

$$p_{ij} = p_{i.} \cdot p_{.j} \quad (\text{Formel 5})$$

$$e_{ij} = p_{ij} \cdot n = p_{i.} \cdot p_{.j} \cdot n \quad (\text{Formel 6})$$

Tabelle 3 - Wahrscheinlichkeiten

		Geschlecht		
		w=1	m=2	
Uni-Karriere	Ja=1	0,245	0,105	0,350
	Nein=2	0,455	0,195	0,650
		0,700	0,300	1,000

Tabelle 4 Erwartungshäufigkeiten

		Geschlecht		
		w=1	m=2	
Uni-Karriere	Ja=1	24,5	10,5	35
	Nein=2	45,5	19,5	65
		70	30	100

Damit hätten wir nun Alles. Die Gleichung

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \quad (\text{Formel 7})$$

gibt uns einen χ^2 -verteilten Koeffizienten an die Hand, mit dessen Hilfe wir über eine Tabelle (BORTZ 1993, S. 699) leicht auf (Nicht-)Signifikanz schließen können. Der χ^2 -Wert wird ermittelt, indem die *quadrierte und gewichtete Abweichung* vom Erwartungswert bestimmt wird. Also je größer die Abweichung, desto größer auch der χ^2 -Wert. Wird daher der kritische χ^2 aus der Tabelle überschritten, dann sind die empirischen Daten nicht durch die Haupteffekte hinreichend erklärt, das heißt die Assoziation ist signifikant. In unserem Fall ist $\chi^2 = 6,81$. Für den kritischen Wert gilt laut

Tabelle $\chi^2_{(1,95\%)} = 3,93 \cdot 10^{-8}$. Das heißt, die Nullhypothese wird verworfen, die Assoziation ist erfüllt auf dem 5%-Niveau. Was schließen wir also daraus? Alles Männergeklüngel!

In Spss gibt es zwei Möglichkeiten log-lineare Modelle auf Signifikanz zu testen (beide werden automatisch ausgegeben). Eine davon ist der eben beschriebene Pearson- χ^2 -Test. Man sieht also, log-lineare Modelle sind nicht wirklich etwas Neues. Das Bekannte wird lediglich verallgemeinert. Dadurch wird der Umgang mit kategorialen Daten flexibler.

3. Herleitung des log-linearen Modells und dessen Parameter

Aus dem ALM will ich folgende Darstellung in Erinnerung rufen. Der einzelne Wert wurde hier aufgefasst als Linearkombination dargestellt von dem Zellenmittel, den Haupteffekten α und β , und der Wechselwirkung $\alpha\beta$ (plus einem Fehler):

$$x_{mij} = \mu_{..} + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{mij} \quad (\text{Formel 8})$$

Was wir im ALM haben, übertragen wir nun auf die kategoriale Datenanalyse, hier dargestellt an einem vier-faktoriellen Beispiel (λ - sprich Lambda).

$$\ln(e_{ijkl}) = \lambda_0 + \lambda_i^{(1)} + \dots + \lambda_r^{(4)} + \lambda_{ij}^{(12)} + \dots + \lambda_{kr}^{(34)} + \lambda_{ijk}^{(123)} + \dots + \lambda_{jkr}^{(234)} + \lambda_{ijkr}^{(1234)} \quad (\text{Formel 9})$$

Also auch hier werden beobachtete Daten durch Haupt- und Wechselwirkungseffekte beschrieben. Denn λ_0 bezeichnet den *Gesamteffekt (overall effect)* λ_i bis λ_r bezeichnen die Haupteffekte, usw.

Warum die Erwartungshäufigkeit ($\ln(e_{ijkl})$) hier logarithmiert ist, hat einen bestimmten Grund. Denn eigentlich ist das log-lineare Modell ein multiplikatives Modell. Am 4-Felder- χ^2 zur Uni-Karriere (s.o.) will ich veranschaulichen, warum das mathematisch aussieht und wie man schließlich wieder auf ein additives Modell kommt.

3.1 Das multiplikative Ursprungs-Modell

In den log-linearen Modellen wird eigentlich davon ausgegangen, dass sich die Effekte durch Multiplikation zusammen setzen. Also nicht durch Addition, wie das im Allgemeinen linearen Modell der Fall ist. Das log-lineare Modell wird im nächsten Abschnitt durch einen mathematischen 'Trick' in ein additives Modell überführt. Wer mehr über den Zusammenhang zwischen additiven und multiplikativen Modellen wissen möchte, dem sei das informelle Skript von Eid (1997) empfohlen.

Formel 10 zeigt dieses multiplikative Modell (γ - sprich Gamma).

$$e_{ij} = \gamma_0 \cdot \gamma_i^{(1)} \cdot \gamma_j^{(2)} \cdot \gamma_{ij}^{(12)} \quad (\text{Formel 10})$$

Der γ_0 Gesamteffekt berechnet sich als Durchschnitt aller Zellen der Kreuztabelle. Dieser Durchschnitt ist aber nicht das arithmetische Mittel, wie man das im ersten Augenblick erwarten würde. Da sich die Effekte multiplikativ zusammensetzen, muss man das geometrische Mittel verwenden. Dieses geometrische Mittel ist sozusagen die Entsprechung zum arithmetischen (im Falle multiplikativer Zusammenhänge). Es ist folgendermaßen definiert:

$$\bar{x}_{geom} = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n} = \left(\prod x_i \right)^{1/n} \quad (\text{Formel 11})$$

Der Gesamteffekt berechnet sich entsprechend:

$$\gamma_0 = \bar{n}_{geom} = \sqrt[4]{n_{11} \cdot n_{12} \cdot n_{21} \cdot n_{22}} = \sqrt[4]{20 \cdot 15 \cdot 50 \cdot 15} = \sqrt[4]{225000} = 21,78 \quad (\text{Formel 12})$$

Abbildung 1 – fiktives Beispiel mit Effektgrößen

		Geschlecht	
		w=1	m=2
Uni-Karriere	Ja=1	20	15
	Nein=2	50	15

$$\gamma_1^{(1)} = \bar{n}_1 = \frac{\sqrt[2]{n_{11} \cdot n_{12}}}{\gamma_0} = \frac{\sqrt[2]{20 \cdot 15}}{21,78} = \frac{\sqrt[2]{300}}{21,78} = 0,80$$

$$\gamma_2^{(1)} = \bar{n}_2 = \frac{\sqrt[2]{n_{21} \cdot n_{22}}}{\gamma_0} = \frac{\sqrt[2]{50 \cdot 15}}{21,78} = \frac{\sqrt[2]{750}}{21,78} = 1,26$$

$$\gamma_2^{(2)} = \bar{n}_{12} = \frac{\sqrt[2]{n_{12} \cdot n_{22}}}{\gamma_0} = \frac{\sqrt[2]{15 \cdot 15}}{21,78} = \frac{\sqrt[2]{225}}{21,78} = 0,69$$

$$\gamma_1^{(2)} = \bar{n}_{1.} = \frac{\sqrt[2]{n_{11} \cdot n_{21}}}{\gamma_0} = \frac{\sqrt[2]{20 \cdot 50}}{21,78} = \frac{\sqrt[2]{1000}}{21,78} = 1,45$$

In Abbildung 1 sind die Formeln inklusive Beispiel für die Haupteffekte dargestellt. Man sieht, dass das geometrische Mittel der jeweiligen Zeile/Spalte relativiert wird am Gesamteffekt. Vergleichbar geht man vor bei der Bestimmung des Interaktionseffektes. Die Häufigkeit in einer Zelle wird relativiert an den entsprechenden Haupteffekten und am Gesamteffekt.

$$\gamma_{ij}^{(12)} = \frac{n_{ij}}{\gamma_0 \cdot \gamma_i^A \cdot \gamma_j^B} \quad (\text{Formel 13})$$

Für Zelle {1,1} unseres Beispiels also folgende Rechnung:

$$\gamma_{11}^{(12)} = \frac{n_{11}}{\gamma_0 \cdot \gamma_1^{(1)} \cdot \gamma_1^{(2)}} = \frac{20}{21,78 \cdot 0,80 \cdot 1,45} = 0,79 \quad (\text{Formel 14})$$

3.2 Das 'additive' log-lineare Modell

Da die additive Form eines Effektmodells leichter interpretierbar ist, und weil wir diese Form auch bereits aus dem ALM kennen, verwenden die meisten Autoren die additive Form. Mathematisch gesehen ist es völlig gleichgültig, ob multiplikativ oder additiv, die Ergebnisse bleiben dieselben.

Formel 10 wird einfach auf beiden Seiten logarithmiert, und weil mathematisch gilt $\ln(ab) = \ln(a) + \ln(b)$, gelangt man über die folgenden Umformungen zum additiven Modell.

$$\begin{aligned}\ln(e_{ij}) &= \ln(\gamma_0 \cdot \gamma_i^{(1)} \cdot \gamma_j^{(2)} \cdot \gamma_{ij}^{(12)}) \\ &= \ln(\gamma_0) + \ln(\gamma_i^{(1)}) + \ln(\gamma_j^{(2)}) + \ln(\gamma_{ij}^{(12)})\end{aligned}\quad (\text{Formel 15})$$

Man bedient sich praktisch eines Tricks, um das multiplikative in das bequemere additive Modell umzuwandeln. Und weil Methodiker auch ein wenig faul sind schreiben sie für jedes $\ln(\gamma) = \lambda$ (λ - sprich Lambda).

$$\ln(e_{ij}) = \lambda_0 + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_{ij}^{(12)} \quad (\text{Formel 16})$$

Wenn man mit den logarithmierten Häufigkeiten Effektgrößen berechnen will, dann stellt man erfreulicherweise fest, dass das erforderliche geometrische Mittel sich in ein arithmetisches ‘verwandelt’ hat. Der additive Gesamteffekt λ_0 ist also der Mittelwert der logarithmierten Zellhäufigkeiten:

$$\begin{aligned}\lambda_0 = \bar{n}_{arith} &= \frac{1}{4} \cdot (\ln(n_{11}) + \ln(n_{12}) + \ln(n_{21}) + \ln(n_{22})) = \frac{1}{4} (3,00 + 2,71 + 3,91 + 2,71) \\ &= \frac{12,33}{4} = 3,08 = \ln(21,78)\end{aligned}\quad (\text{Formel 17})$$

Das Prinzip läßt sich wieder leicht übertragen auf alle anderen Effektgrößen, wobei man aber beachten muss, dass aus der Division beim logarithmieren eine Subtraktion wird $\ln(a/b) = \ln(a) - \ln(b)$. Ein Haupteffekt-Parameter und der Interaktionsparameter sollen als Beispiel ausformuliert werden.

$$\lambda_1^{(1)} = \left(\frac{1}{2} \cdot (\ln(n_{11}) + \ln(n_{12})) \right) - \lambda_0 = \frac{1}{2} (3,00 + 2,71) - 3,08 = \frac{5,71}{2} - 3,08 = -0,23 = \ln(0,80) \quad (\text{Formel 18})$$

$$\lambda_{11}^{(12)} = \ln(n_{11}) - \lambda_0 - \lambda_1^{(1)} - \lambda_1^{(2)} = 3,00 - 3,08 - (-0,23) - (-0,37) = 0,52 = \ln(1,66) \quad (\text{Formel 19})$$

3.3 Restriktionen

Damit das log-lineare Modell mathematisch handhabbar ist und die λ -Parameter geschätzt werden können, muss den λ s noch eine Beschränkung auferlegt werden. In den aller-(aller-)meisten Fällen sehen diese Einschränkungen vor, dass die Summe der λ -Parameter einer Zeile/Spalte Null ergibt (was ANDREß, HAGENAARS & KÜHNEL (1997, S.71ff) als Effektkodierung oder auch zentrierte Effekte bezeichnen). Wird eine solche Einschränkung nicht vorgenommen, ist man nicht in der Lage, die Parameter zu errechnen (das Modell zu identifizieren). Eben darum nennt man diese Restriktionen *Identifikationsrestriktionen*.

$$\begin{aligned}
 \sum_{i=1}^I \lambda_i^{(1)} &= 0 \\
 \sum_{j=1}^J \lambda_j^{(2)} &= 0 \\
 \sum_{i=1}^I \lambda_{ij}^{(12)} &= \sum_{j=1}^J \lambda_{ij}^{(12)} = 0
 \end{aligned}
 \tag{Formel 20}$$

Wenn ich also alle Effekte eine Zeile oder Spalte aufsummiere, dann muss das Ergebnis Null sein.

Wenn eine Variable (1) zwei Stufen hat und $\lambda_{i=1}^{(1)} = -0,5$, dann muss $\lambda_{i=2}^{(1)} = +0,5$ sein, damit $\lambda_1^{(1)} + \lambda_2^{(1)} = -0,5 + 0,5 = 0$ gilt.

Im Prinzip könnte man sich noch die verschiedensten Restriktionen für ein solches Modell ausdenken. Je mehr Restriktionen man einführt, desto mehr Freiheitsgrade erhält man. Man könnte einen (oder mehrere) Effektparameter auf Null setzen. In diesem Fall spricht man von einem Submodell und diese werden im folgenden Abschnitt behandelt (Kap. 3.4).

Oder man könnte beispielsweise festlegen, dass bestimmte Parameter gleich groß sein müssen. In einigen Spezialfällen der log-linearen Modelle ist das tatsächlich üblich. Da es aber für den allgemeinen Fall (wie er hier besprochen wird) nicht so wichtig ist, soll das hier nicht weiter vertieft werden. So etwas wird interessant, wenn man mit log-linearen Modellen Pfad- oder Latent-Class Modelle konstruieren möchte. Die log-linearen Modelle werden durch die oben genannte Restriktion (Formel 20) identifizierbar (also berechenbar); das ist erst einmal das Wesentliche und darauf lasse ich es beruhen.

3.4 Submodelle

Das Modell in Formel 16 wird als *saturiertes (gesättigtes)* Modell bezeichnet. Hier sind alle Haupt- und Wechselwirkungseffekte berücksichtigt. Lässt man dagegen einige Parameter weg, setzt man sie ‘gewaltsam’ auf Null, so erhält man eine große Vielfalt von *ungesättigten* Modellen oder *Submodellen*. Ändert sich die Zusammensetzung eines Zellenwertes, so hat das Auswirkungen auf die Berechnung der erwarteten Häufigkeiten. Diese müssen entsprechend dem Modell der Nullhypothese geschätzt werden.

An einem Beispiel einer 4-Felder-Kontingenztafel sollen verschiedene Submodelle besprochen werden.

1. $\ln(e_{ij}) = \lambda_0$: Bedeutet, dass die Verteilung der beobachteten Häufigkeiten in den Zellen (im wesentlichen) konstant sind, d.h. die Variation innerhalb der Tafel und innerhalb der Randsummen

ist irrelevant. In einer 2x2-Tafel ergibt sich für jede Zelle ein Schätzwert $e_{ij} = n_{..}/4$. Bei einem Vergleich zwischen n_{ij} mit diesen e_{ij} sind dann 3 Freiheitsgrade zu berücksichtigen!

2. $\ln(e_{ij}) = \lambda_0 + \lambda_i$ bedeutet, dass die Verteilung der beobachteten Häufigkeiten (im wesentlichen) durch die Randsummen in A beschrieben ist. Die Randverteilung von B wird als irrelevant angenommen. In einer 2x2-Tafel ergibt sich für jede Zelle ein Schätzwert $e_{ij} = n_{i.}/2$. Bei einem Vergleich zwischen n_{ij} mit diesen e_{ij} dann 2 Freiheitsgrade zu berücksichtigen!
3. $\ln(e_{ij}) = \lambda_0 + \lambda_j$ ist wie (2) zu interpretieren, nur dass die Verteilung der Randsumme von A als irrelevant angenommen wird.
4. $\ln(e_{ij}) = \lambda_0 + \lambda_i + \lambda_j$ ist das "Standardmodell" mit einem Freiheitsgrad (in einer 2x2-Tafel). Es repräsentiert die typische Nullhypothese, die nur die Assoziation zwischen den beiden Variablen ignoriert. Die Erwartungshäufigkeiten werden hier geschätzt durch $e_{ij} = n_{i.}n_{.j}/n_{..}$.
5. $\ln(e_{ij}) = \lambda_0 + \lambda_i + \lambda_j + \lambda_{ij}$: Dies ist das vollständige sog. saturierte Modell, das 0 Freiheitsgrade aufweist. Ohne Freiheitsgrade ist das Modell nicht direkt statistisch testbar, was aber auch nicht weiter schlimm ist. Die Vorgehensweise bei der statistischen Analyse ist derart beschaffen, dass so etwas gar nicht nötig ist (siehe Kap. 5).

3.5 Konditionale und partielle Effekte

Aus dem Allgemeinen linearen Modell kennen wir die einfachen Haupteffekte und die einfachen ... Bei den log-linearen Modellen gibt es etwas Ähnliches. Allerdings sind Partial- und Konditionaleffekte erst dann wirklich interessant, wenn mehr als zwei Faktoren vorliegen. Tabelle 5 zeigt ein drei-faktorielles Beispiel. Es ist aus dem ANDREß, HAGENAARS & KÜHNEL (1997) übernommen. Sie äußern die theoretische Vermutung, dass Ältere (aufgrund ihrer eher konservativen Einstellung) stärker in das politische System integriert seien und deshalb öfter wählen gehen. Diese konservative Haltung müsste sich auch in einer stärkeren konfessionellen Bindung niederschlagen.

Tabelle 5

		Konfession (2)			
		ja		nein	
		Alter (3)			
		alt	jung	alt	jung
Wahl? (1)	ja				
	nein				

Um das zu untersuchen betrachten wir die konditionalen Effekte. Nach der theoretischen Überlegung müsste ein Zusammenhang zwischen dem Alter und dem Wählerverhalten verschwinden, wenn die konfessionelle Bindung kontrolliert wird. Im Prinzip isoliert man einfach einen Teil der Kreuztabelle vom anderen, und rechnet dann weiter als hätte man von Anfang an nur eine zweidimensionale Tabelle gehabt.

Gemäß ANDREß, HAGENAARS & KÜHNEL (1997) ergibt sich eine konditionale Assoziation ($\lambda_{ijk}^{(13|2)}$) bei den Kirchgängern: hier wählen Jüngere weniger. Für die nicht Kirchgänger verschwindet dieser Zusammenhang.

Partielle Effekte ergeben sich als Durchschnitt der konditionalen Effekte.

$$\lambda_{ik}^{(1\sim 3)} = \frac{1}{2} \cdot (\lambda_{ik1}^{(13|2)} + \lambda_{ik2}^{(13|2)}) \quad (\text{Formel 21})$$

Ist der partielle Effekt von dem marginalen Effekt verschieden, so deutet das auf folgenden Zusammenhang hin. Wenn man eine Variable kontrolliert (hier Konfession), dann verändert sich der Effekt einer anderen Variable (hier Alter), weil ein Teil der Alterseffekte auf die Konfession zurückzuführen ist. Der Partialeffekt zeigt also an, wie die Beziehung zwischen Alter und Wahlverhalten aussieht, wenn der Effekt der Konfession ‘herauspartialisiert’ wurde. Er kann also analog zur Partialkorrelation intervallskalierter Merkmale interpretiert werden.

Zusätzlich sei noch folgender Zusammenhang besprochen: Eine Wechselwirkung 2. Ordnung gibt an, ob die Daten Variationen zeigen, die nicht durch die Wechselwirkung 1. Ordnung erklärbar ist. Man kann das auch so formulieren. Die Wechselwirkung 2. Ordnung gibt an, wie sehr die beiden konditionalen Wechselwirkungen von ihrem Mittelwert (dem Partialeffekt) abweichen. Aus dieser Logik lässt sich folgende Formel erstellen:

$$\lambda_{ijk}^{(123)} = \lambda_{ikj}^{(13|2)} - \lambda_{ik}^{(1\sim 3)} = \lambda_{ijk}^{(12|3)} - \lambda_{ij}^{(12\sim)} = \lambda_{jki}^{(23|1)} - \lambda_{jk}^{(2\sim 3)} \quad (\text{Formel 22})$$

Die konditionalen und partiellen Effekte stehen mathematisch im Zusammenhang mit den Odds, die unten (Kapitel 9.1) noch angesprochen werden (genauer hierzu, siehe ANDREß, HAGENAARS & KÜHNEL 1997, S.155).

4. Schätzung der Erwartungshäufigkeiten

Die Erwartungshäufigkeiten werden bei kategorialen Daten so geschätzt, dass sie die Zusammenhänge widerspiegeln, die in der Nullhypothese angenommen werden. Lautet die Nullhypothese $\lambda_0 + \lambda_i^{(1)} + \lambda_j^{(2)}$, so werden die Erwartungshäufigkeiten auf Grundlagen der Randverteilungen (1) und (2) geschätzt. Man bezeichnet die Randverteilungen (1) und (2) als die angepassten Randverteilungen. Wie das genau vor sich geht ist bereits in Abschnitt 2 beschrieben.

Lautet die Nullhypothese $\lambda_0 + \lambda_i^{(1)}$, so werden die Erwartungshäufigkeiten allein auf Grundlage der angepassten Randverteilung (1) geschätzt. Für den Faktor (2) wird nun dagegen angenommen, die Daten seien gleichverteilt. Anhand der Daten aus Abschnitt 2 soll das kurz verdeutlicht werden. Faktor (1) ist der Faktor Uni-Karriere. Die entsprechenden Werte in der Randverteilung (1) sind also 35 und 65. Also gilt für $e_{11} = e_{12} = 35/100 \cdot 50/100 \cdot 100 = 35/2 = 17,5$, und entsprechendes natürlich für $e_{21} = e_{22} = 65/100 \cdot 50/100 \cdot 100 = 65/2 = 32,5$.

Würde die Nullhypothese dagegen heißen $\lambda_0 + \lambda_{ij}^{(12)}$, müsste die angepasste Verteilung also (12) lauten. Nun ist die Verteilung (12) im unserem zwei-faktoriellen die Kreuztabelle selbst (siehe Tabelle 1). Das bedeutet, die Erwartungshäufigkeiten sind identisch mit den beobachteten Häufigkeiten. Dasselbe gilt, wenn wir das saturierte Modell der Nullhypothese zugrunde legen.

In der Regel rechnet man seine Statistik nicht selbst, sondern benutzt dazu ein Computerprogramm. Deshalb soll noch kurz erwähnt werden, dass solche Programme die Erwartungshäufigkeiten meist nicht auf die Weise berechnen, wie sie hier dargestellt ist. Spss zum Beispiel verwendet dazu einen Algorithmus, der als Newton-Raphson-Algorithmus bezeichnet wird. Was es mit diesen Verfahren genauer auf sich hat, kann man bei ANDREB, HAGENAARS & KÜHNEL (1997) nachlesen. Allerdings ergeben diese Algorithmen exakt dieselben Ergebnisse. Nur in einigen (nicht-hierarchischen) Fällen lassen sich die Erwartungshäufigkeiten nicht in der dargestellten Weise berechnen. Dann sind diese Schätzverfahren unabdingbar.

5. Signifikanztests

Betrachten wir noch einmal das Beispiel aus Kapitel 2. Daran läßt sich die Vorgehensweise recht leicht veranschaulichen. Wir schätzen dort die Erwartungshäufigkeiten e_{ij} aufgrund der Annahme, dass keine Interaktion vorliegt. Wir nehmen also das Modell $\ln(e_{ij}) = \lambda_0 + \lambda_i + \lambda_j$ und betrachten das als Nullhypothese. Wenn wir nun einen Signifikanztest durchführen, dann testen wir, ob das Modell der Nullhypothese die Daten gut erklären kann. Ist das nicht der Fall wird der statistische Test signifikant und wir können die Alternativhypothese als gegeben betrachten.

Daraus wird nun ersichtlich, dass es nicht wichtig ist, das saturierte Modell einem Signifikanztest zu unterziehen. Denn das saturierte Modell würde als das *Modell der Alternativhypothese* betrachtet werden. Das Modell der Nullhypothese wäre das Submodell $\ln(e_{ij}) = \lambda_0$. Ich will also wissen, ob *irgendeiner* der verbleibenden Parameter (Haupteffekt λ_i , λ_j oder Interaktion λ_{ij}) einen bedeutsamen Beitrag zur Erklärung meiner Datenstruktur beitragen kann.

5.1 Testen eines Modells: Goodness of fit

Es gibt zwei Möglichkeiten ein log-lineares Modell auf Signifikanz zu testen (auch Spss™ beherrscht beide Alternativen). Die erste Möglichkeit wurde bereits oben angesprochen. Es handelt sich um das Pearson- χ^2 (hier an einem 4-faktoriellen Beispiel):

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \sum_{r=1}^R \frac{(n_{ijkl} - e_{ijkl})^2}{e_{ijkl}} \quad (\text{Formel 23})$$

Ebenso kann ein Log-Likelihood-Quotiententest mit der Testgröße

$$L^2 = 2 \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \sum_{r=1}^R n_{ijkl} (\ln(n_{ijkl}) - \ln(e_{ijkl})) \quad (\text{Formel 24})$$

durchgeführt werden. Beide Größen sind *annähernd* χ^2 -verteilt und beide erbringen in der Regel ganz ähnliche Ergebnisse. Nur wenn $n_{ijkl} < 5$ erbringt das L^2 oft schlechtere Leistungen (AGRESTI 1990 S.49).

Ein (Sub-)Modell wird nun beibehalten, wenn der entsprechende χ^2 - oder L^2 -Wert kleiner ist als der kritische Wert aus der χ^2 -Tabelle. Die Nullhypothese ist erfüllt. Ansonsten muss das Modell verworfen werden, die Signifikanz p ist *kleiner* als die α -Fehlerwahrscheinlichkeit.

Die Anzahl der Freiheitsgrade entspricht (nach LANGEHEINE 1986) der Anzahl der Zellen einer Tabelle minus Eins (für λ_0) minus der Anzahl derjenigen λ -Parameter, die voneinander unabhängig sind. Für eine Reihe von 3-dimensionalen Modellen lassen sich die Freiheitsgrade in Tabelle 6 nachsehen.

Wie man sich nach dieser Regel ausrechnen kann, hat das saturierte Modell 0 Freiheitsgrade. Es ist also gar nicht direkt statistisch testbar. Soll das saturierte Modell als Grundlage dienen, dann müssen die λ -Koeffizienten einzeln auf Signifikanz geprüft werden (siehe 5.4).

5.2 Hierarchische Modelltests

Im wissenschaftlichen Sinne ist die optimale Strategie sehr restriktiv: aus bekannten Modellvorstellungen werden klare Hypothesen aufgestellt darüber, welche Haupteffekte und Interaktionen bedeutsam sein müssen. Diese werden also gezielt der statistischen Prüfung unterzogen. Und kann Signifikanz nicht bestätigt werden, so muss das Modell verworfen werden.

Oftmals ist diese Strategie jedoch kaum angebracht, denn sie ist nicht besonders konstruktiv. Oft sind noch gar keine Modelle und theoretische Variablenbeziehungen bekannt und es wäre oberstes Ziel der Statistik mögliche Hypothesen überhaupt erst zu explorieren, anstatt bestehende Hypothesen zu bestätigen. Beide Strategien sind im richtigen Zusammenhang angewandt richtig und wichtig.

Hierarchische Modelltests bieten eine Möglichkeit verschiedene Effekte explorativ zu suchen und Licht ins theoretische Dunkel zu bringen. Zuerst aber muss erläutert werden, was ein hierarchisches Modell überhaupt ist.

Wenn unter einer Hypothese angenommen wird, ein bestimmter λ -Koeffizient sei Null, dann müssen folglich auch alle übrigen (höher interagierenden) λ 's Null sein, von denen unser λ eine Teilmenge darstellt, wenn also

$$\lambda_j^{(2)}=0 \quad , \text{ dann gilt auch } \lambda_{ij}^{(12)}=0 \quad , \quad \lambda_{jk}^{(23)}=0 \quad , \quad \lambda_{jkr}^{(234)}=0 \quad \text{ usw.}$$

oder

$\lambda_{ij}^{(12)}=0 \quad , \text{ dann gilt auch } \lambda_{ijk}^{(123)}=0 \quad , \quad \lambda_{jkr}^{(124)}=0 \quad , \quad \lambda_{ijkr}^{(1234)}=0 \quad \text{ usw.}$ (Tabelle 6 zeigt alle hierarchischen Modelle für eine 3-faktorielle Kreuztabelle).

Mit Hilfe der *Goodness-of-fit* Statistik läßt sich ein jedes mögliches hierarchisches Modell auf seine Güte testen. Ist Signifikanz gegeben, ist es zur Erklärung der Daten nicht ausreichend. In diesem Fall geht die Suche nach dem 'besten' Modell weiter.

Bei der hierarchischen Modellsuche geht Spss folgendermaßen vor (Menüpunkt Modellauswahl). Die verwendete Strategie wird dort als Rückwärtselimination bezeichnet. Man geht vom saturierten Modell aus und eliminiert Schritt für Schritt denjenigen Effekt, der am kleinsten ist.

Und es existieren darüber hinaus auch andere Suchstrategien, wobei diese Strategien nicht alle hierarchisch vorgehen hierarchisch vorgehen müssen. Man findet diese beschrieben bei RECKE (1989, S.112ff).

Manchmal kann es dabei geschehen, dass mehrere Modelle signifikant werden und sich nicht per Augenmaß entscheiden läßt, welches davon das Beste ist. Mit der Lösung dieses Problems beschäftigt sich der folgende Abschnitt.

Tabelle 6 - Alle hierarchischen Modelle einer I x J x K - Tabelle

Modell	Angepasste Randverteilung	df
$\ln(e_{ijk}) = \lambda_0 + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)} + \lambda_{ik}^{(13)} + \lambda_{jk}^{(23)} + \lambda_{ijk}^{(123)}$	(123)	0
$\ln(e_{ijk}) = \lambda_0 + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)} + \lambda_{ik}^{(13)} + \lambda_{jk}^{(23)}$	(12),(13),(23)	(I-1)(J-1)(K-1)
$\ln(e_{ijk}) = \lambda_0 + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)} + \lambda_{ik}^{(13)}$	(12),(13)	I(J-1)(K-1)
$\ln(e_{ijk}) = \lambda_0 + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)} + \lambda_{jk}^{(23)}$	(12),(23)	(I-1)J(K-1)
$\ln(e_{ijk}) = \lambda_0 + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ik}^{(13)} + \lambda_{jk}^{(23)}$	(13),(23)	(I-1)(J-1)K
$\ln(e_{ijk}) = \lambda_0 + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)}$	(12),(3)	(IJ-1)(K-1)
$\ln(e_{ijk}) = \lambda_0 + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ik}^{(13)}$	(13),(2)	(IK-1)(J-1)
$\ln(e_{ijk}) = \lambda_0 + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{jk}^{(23)}$	(23),(1)	(JK-1)(I-1)
$\ln(e_{ijk}) = \lambda_0 + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_{ij}^{(12)}$	(12)	IJ(K-1)
$\ln(e_{ijk}) = \lambda_0 + \lambda_i^{(1)} + \lambda_k^{(3)} + \lambda_{ik}^{(13)}$	(13)	IK(J-1)
$\ln(e_{ijk}) = \lambda_0 + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{jk}^{(23)}$	(23)	JK(I-1)
$\ln(e_{ijk}) = \lambda_0 + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)}$	(1),(2),(3)	IJK-I-J-K+2
$\ln(e_{ijk}) = \lambda_0 + \lambda_i^{(1)} + \lambda_j^{(2)}$	(1),(2)	IJK-I-J+1
$\ln(e_{ijk}) = \lambda_0 + \lambda_i^{(1)} + \lambda_k^{(3)}$	(1),(3)	IJK-I-K+1
$\ln(e_{ijk}) = \lambda_0 + \lambda_j^{(2)} + \lambda_k^{(3)}$	(2),(3)	IJK-J-K+1
$\ln(e_{ijk}) = \lambda_0 + \lambda_i^{(1)}$	(1)	I(JK-1)
$\ln(e_{ijk}) = \lambda_0 + \lambda_j^{(2)}$	(2)	J(IK-1)
$\ln(e_{ijk}) = \lambda_0 + \lambda_k^{(3)}$	(3)	K(IJ-1)
$\ln(e_{ijk}) = \lambda_0$	N	IJK-1

5.3 Testen eines Modells gegen ein anderes

Es gibt die Möglichkeit zu testen, ob ein Submodell die empirischen Daten signifikant besser erklären kann als ein anderes. Hierbei kann der Likelihood-Ratio Test L^2 wertvolle Dienste leisten, was ihn gegenüber dem Pearson- χ^2 auszeichnet. Voraussetzung dafür ist allerdings, dass die beiden Modelle zueinander hierarchisch sind (nicht zu verwechseln mit Abschnitt 5.2); das bedeutet, dass das eine Modell aus dem anderen hervorgeht, indem man Parameter herausstreicht (aber keine neuen hinzufügt) (HAGENAARS 1990). In diesem Sinne wären die Modelle $\lambda_0 + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{jk}^{(23)}$ und

$\lambda_0 + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)}$ nicht zueinander hierarchisch; die Modelle $\lambda_0 + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{jk}^{(23)}$ und $\lambda_0 + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ik}^{(13)} + \lambda_{jk}^{(23)}$ dagegen sehr wohl.

Die L^2 -Werte der beiden Modelle können nun einfach subtrahiert werden ($L^2_{\text{neu}} = L^2_{\text{Modell2}} - L^2_{\text{Modell1}}$). Diese neue Größe ist wiederum (annähernd) χ^2 -verteilt und kann in der Tabelle nachgesehen werden. Die Anzahl der Freiheitsgrade entspricht der Differenz zwischen den Freiheitsgraden beider Submodelle (siehe ANDERSEN 1996).

Allerdings sollten nach LANGEHEINE (1986) bei der Modellauswahl auch andere Kriterien zum Tragen kommen. Seiner Ansicht nach sollte das Modell nicht nur den größtmöglichen Varianzanteil aufklären, sondern auch mit möglichst wenigen Parametern auskommen. Oder anders formuliert, ein Modell sollte nicht nur effektiv sondern auch ökonomisch sein.

5.4 Einzeltests

Wenn Null- und Alternativhypothese sich durch mehr als einen λ -Parameter unterscheiden, und wenn mein statistischer Gesamttest (Kap. 5.1) signifikant wird, dann ist unklar, welcher der Parameter nun einen zusätzlichen Erklärungswert liefert. Lautete die Nullhypothese beispielsweise $\ln(e_{ij}) = \lambda_0$, dann könnte einer der beiden Haupteffekte die zusätzliche Bedeutsamkeit ausmachen; oder auch beide; oder die Interaktion; oder alle drei; oder...

In einem anderem Fall sind Einzeltests ebenfalls notwendig. Nämlich wenn dem Test das saturierte Modell zugrunde liegt. Jeder einzelne λ -Parameter kann für sich überprüft werden.

Dazu benötigt man die Varianz der Kreuztabelle

$$s_{\lambda}^2 = \frac{\sum_{i=1}^{I-2} \sum_{j=1}^{J-2} \sum_{k=1}^{K-2} \sum_{r=1}^{R-2} \frac{1}{n_{ijk r}}}{C^2}, \quad (\text{Formel 25})$$

wobei C die Anzahl der Zellen ist.

Der Quotient zwischen einem λ -Parameter und der Streuung λ/s_{λ} ermöglicht den Test, ob sich λ signifikant von Null unterscheidet. Denn unter zwei Voraussetzungen ist dieser Quotient annähernd normalverteilt. Zum ersten müssen die Stichproben hinreichend groß sein, und zum anderen muss für λ erwartet werden, dass es gleich Null sei. Dann ist λ/s_{λ} also normalverteilt mit dem Mittelwert $m=0$ und der Streuung $s=1$. So bleibt nur zu prüfen, ob der Quotient größer/kleiner $\pm 1,96$ (5%-Grenze) ist. Denn dann unterscheidet sich das betreffende λ signifikant von Null, der Effekt ist gegeben.

Allerdings gilt Formel 25 nur für (beliebig viele) *dichotome* Variablen. Für polytome Variablen ist die Berechnung komplizierter (also für Variablen mit mehr als zwei Stufen) (vgl. LANGEHEINE 1986, S.129).

6. Interpretation der Effekte

Um die Größe eines Effektes interpretieren zu können, müssen wir erst einmal dessen Wertebereich kennen (Tabelle 7). Diese Grenzen werden praktisch aber zwar nie erreicht, aber sie stecken den absoluten Rahmen ab. (ANDREß, HAGENAARS & KÜHNEL 1997, S. 158).

Tabelle 7

	<i>Minimum</i>	<i>Kein Effekt</i>	<i>Maximum</i>
γ	0	1	$+\infty$
λ	$-\infty$	0	$+\infty$

Hier läßt sich eine weitere Stärke der λ -Parameter ausmachen. Sie verteilen sich symmetrisch um die Null herum, wodurch positive und negative Effekte direkt miteinander vergleichbar sind. Bei dem asymmetrisch um die Eins verteilten γ müssen die Parameter erst umgerechnet werden (indem man den Kehrwert bildet). Bleibt nun nur noch zu beachten, dass das Vorzeichen eines Effektes nur bei ordinalen Daten sinnvoll interpretierbar ist.

Ein weiterer Aspekt einer Effektgröße betrifft nicht ihre absolute Ausprägung, sondern ihr Ausmaß an *praktischer Bedeutsamkeit*. Auch kleine Effekte können in bestimmtem Zusammenhängen große Wirkungen hervor rufen: Eine Steigerung der Früherkennung von Brustkrebs um auch nur ein Prozent würde den Krankenkassen hierzulande hohe Kosten ersparen.

Auch für Kontingenztafeln hat COHEN (1988) versucht eine standardisierte Einteilung zu ermitteln. Er verwendet den Gesamteffekt w , der sich als Maß für die Abweichung der empirischen von der vorhergesagten Tabelle verstehen läßt.

Tabelle 8

$w = \sqrt{\frac{\sum_{i=1}^I (p_i - \hat{p}_i)^2}{\hat{p}_i}}$	<i>Klein</i>	<i>Mittel</i>	<i>Groß</i>
	0.10	0.30	0.50

Weitere bedeutsame Effektgrößen sind der ϕ -(phi) und der Kontingenz-Koeffizient.

$$\phi^2 = \frac{X^2_{Pearson}}{n} \approx \frac{L^2}{n} \quad (\text{Formel 26})$$

Das ϕ^2 ist allerdings kein standardisiertes Maß und daher sind ϕ^2 zwischen Untersuchungen nicht vergleichbar. Darüber hinaus ist es bei künstlicher Nominalisierung von intervall-/ ordinalskalierten Variablen davon abhängig, wo man die Kategoriengrenzen zieht.

$$CC = \sqrt{\frac{\chi^2}{\chi^2 + n}}, \text{ wobei } 0 \leq CC \leq CC_{\max} \leq 1 \quad (\text{Formel 27})$$

$$CC_{\max} = \sqrt{(R-1)/R}, \text{ R ist die kleinste Anzahl an Stufen}$$

Leider ist auch dieses Maß nicht vollständig standardisiert. Der Kontingenzkoeffizient kann nie größer werden als CC_{\max} . Diese Eigenschaft macht man sich zu nutze, CC wird an CC_{\max} relativiert und variiert nun zwischen 0 und 1.

$$CC_{\text{relativ}} = \sqrt{CC^2 / CC_{\max}^2} \quad (\text{Formel 28})$$

7. Probleme & Fallen im Umgang mit log-linearen Modellen

7.1 Alpha-Kumulierung

Werden mehrere Haupteffekte und Interaktionen einzeln auf Signifikanz geprüft (Kap. 5.4) begegnet uns wieder das alte Problem der α -Kumulierung. Aber auch beim Test alternativer Submodelle (wie im hierarchischen Fall) stellt sich dieses Problem. LANGEHEINE (1980) verweist auf die (altbekannte) Bonferroni-Korrektur: Wenn r simultane Tests durchgeführt werden so muss $\alpha' = \alpha/r$ kritisches Niveau verwendet werden.

7.2 Stichprobengröße

STELZ (2000) schlägt als Minimum die fünffache Anzahl der Zellen vor, denn dies stellt die Untergrenze für eine passable Teststärke dar (bei großem Effekt). Darüber hinaus konnte sie zeigen, dass sich das Pearson- χ^2 unter dieser Voraussetzung als zuverlässiger Schätzer erweist. Das L^2 zeigt dagegen einige Schwächen, vor allem bei Interaktionen 3. Ordnung und höher.

Darüber hinaus sollte die Stichprobengröße individuell an die Untersuchungsbedingungen angepasst werden aufgrund einer Festlegung von Effekt, α - und β -Fehler.

7.3 Post-hoc Teststärke

Für das Pearson- χ^2 errechnet sich der Nonzentralitätsparameter λ folgendermaßen (AGRESTI 1990):

$$\lambda = n \sum_{i=1}^I \sum_{j=1}^J \frac{(\pi_{ij} - \hat{\pi}_{ij})^2}{\hat{\pi}_{ij}} \quad (\text{Formel 29})$$

Beim Likelihood-Ratio gilt:

$$\lambda = 2 \cdot n \cdot \sum_{i=1}^I \sum_{j=1}^J \pi_{ij} - \ln \left(\frac{\pi_{ij}}{\hat{\pi}_{ij}} \right) \quad (\text{Formel 30})$$

In beiden Fällen ist $\hat{\pi}_{ij}$ gleich Null, wenn $n_{ij} = e_{ij}$. Je größer deren Abweichung, desto größer wird $\hat{\pi}_{ij}$ und desto größer wird damit die Power. Der genaue Betrag muss in der χ^2 nachgesehen werden.

Tabelle 9 - Power des Chi-Quadrat Tests bei $\alpha = 0.05$ (Auszug aus Tabelle 7.12 in AGRESTI 1990)

Nonzentralitätsparameter										
df	0	0,4	1	2	3	4	5	8	10	15
1	0,500	0,097	0,170	0,293	0,410	0,516	0,609	0,807	0,885	0,972
2	0,500	0,081	0,133	0,226	0,322	0,415	0,504	0,718	0,815	0,944
3	0,500	0,075	0,116	0,192	0,275	0,358	0,440	0,654	0,761	0,917
4	0,500	0,071	0,106	0,172	0,244	0,320	0,396	0,605	0,716	0,891
6	0,500	0,066	0,094	0,146	0,206	0,270	0,336	0,531	0,644	0,843
8	0,500	0,064	0,087	0,131	0,182	0,238	0,296	0,477	0,588	0,799
10	0,500	0,062	0,082	0,121	0,166	0,215	0,268	0,435	0,542	0,760
20	0,500	0,056	0,066	0,096	0,125	0,158	0,193	0,315	0,402	0,611
50	0,500	0,054	0,061	0,076	0,092	0,110	0,129	0,198	0,250	0,398

7.4 Leere Zellen

Es kann ein Problem sein, wenn nicht alle Zellen in einer Kontingenztafel besetzt sind. Man unterscheidet dabei den Fall, dass Zellen zufällig leer geblieben sind und den Fall, dass bestimmte Merkmalskombinationen logisch nicht möglich sind.

Im ersten Fall (*sampling zeros*) sind Leerzellen auf Zufälle zurück zu führen. Das Problem kann praktisch gelöst werden, indem man noch einige Probanden mehr erhebt, bis sich für alle Zellen Werte finden lassen. Mathematisch kann man in Ausnahmefällen auch zu jeder Zelle eine additive Konstante addieren. Goodman empfiehlt die Addition von 0,5, da dies zu einer besseren Schätzung der λ -Parameter führt. Inzwischen weisen zwar einige Autoren darauf hin, dass eine Erhöhung um 0,5 zu strikt sei und zu einer unerwünschten Begünstigung der Nullhypothese führt (LANGEHEINE 1986, STEVENS 1996), trotzdem führt Spss jedoch diese Korrektur automatisch durch.

In den meisten Fällen bereiten *sampling zeros* aber keine Schwierigkeiten, weil sich die Erwartungshäufigkeiten e_{ij} trotzdem korrekt schätzen lassen (LANGEHEINE 1980).

Im zweiten Fall (*structural zeros*) treten Leerzellen auf, weil bestimmte Merkmalskombinationen einfach nicht realistisch oder logisch sind. Wird etwa die Teilnahme an Krebsfrüherkennungsprogrammen untersucht in Abhängigkeit von den Variablen Einstellung,

Geschlecht und Alter (< 44 , ≥ 45), so werden einige Zellen unbesetzt bleiben, weil das Gesetz eine regelmäßige Teilnahme an Vorsorgeuntersuchungen für Männer erst ab 45 Jahren vorsieht.

Auch dieses Problem läßt sich lösen, indem man die 0,5 hinzu addiert. Es bringt allerdings den Nachteil, das dann die Freiheitsgrade nicht mehr korrekt berechnet werden. Leider kann ich euch nicht sagen, ob Spss in der Lage ist, die Freiheitsgrade entsprechend zu korrigieren.

7.5 Zu kleine Erwartungshäufigkeiten

BORTZ (1993) weist darauf hin, dass die –Techniken zur Bestimmung von Signifikanzen relativ robust sind. Dessen ungeachtet müsse jedoch darauf geachtet werden, dass die Anzahl kleiner erwarteter Häufigkeiten nicht überhand nimmt. Maximal $1/5$ (20%) aller Zellen dürfen Erwartungshäufigkeiten kleiner als 5 aufweisen. Spss ist da mit 30% liberaler.

7.6 Nichtlineare Effekte

Für den Fall, dass ‘echte’ nominale Daten erhoben werden, stellt sich die Frage nach nichtlinearen Effekten nicht. Ist eine Variable aber künstlich nominalisiert und hat sie eine natürliche Rangfolge, dann können nichtlineare Effekte durchaus ein Problem darstellen. Beispiel für eine solche Variable wären Altersgruppen.

LANGEHEINE (1986) skizziert folgende zwei Lösungsansätze, verweist aber für Details seinerseits auf GOODMAN (1971) oder HABERMAN (1979):²

1. In einem explorativen Ansatz lassen sich die Effekte und Interaktionen in lineare, quadratische, kubische (usw.) Komponenten zerlegen. Diese Komponenten können wie in der Varianzanalyse auf Signifikanz getestet werden.
2. Vermutet man a priori lineare oder nichtlineare Beziehungen, so lassen sich solche Hypothesen bei der Modellspezifikation durch Einführung entsprechender Restriktionen berücksichtigen. Dabei kann sich herausstellen, dass ein Variablensystem bei deutlicherer Herausarbeitung der Struktur durch weniger Parameter beschreibbar und einfacher interpretierbar ist.

7.7 Vorgehensweise bei der Erhebung

Die Mathematiker können beweisen, dass die Vorgehensweise bei der Datenerhebung die wahrscheinlichkeitstheoretische Verteilung der Daten beeinflusst, die der Kreuztabelle zugrunde liegt. So bietet Spss die Unterscheidung zwischen der Poisson- und der Multinomialverteilung an. Es ist nicht wichtig, genau zu wissen, was das heißt; Aber es ist wichtig folgendes zu berücksichtigen (FIENBERG 1980):

2 Ich zitiere hier Langeheine lediglich, weil ich gestehen muß, daß seine Darstellung in dieser Kürze nicht wirklich verständlich ist.

1. *Poisson*: In diesem Fall werden einfach Daten erhoben für eine bestimmte Zeit. Es existieren keine genaueren Vorstellungen über die Anzahl der Probanden, die letztlich erhoben werden sollen. In diesem Fall ist die Poisson-Verteilung als Basisverteilung zu wählen. Beispiel: Umfrage vor einem Kaufhaus von 14 bis 16 Uhr.
2. *Multinomial*: Wenn den Versuchspersonenzahlen von vorne herein Restriktionen auferlegt werden, dann ist die multinomiale Verteilung zu wählen. Wenn also die gesamte Probandenzahl (n .) a priori festgelegt wird.
3. *Produkt-multinomial*: Es werden Häufigkeiten für verschiedene Gruppen von vorne herein festgelegt, das heißt bestimmte Randverteilungen werden fixiert. Dieser Fall unterscheidet sich von den beiden oberen doch ziemlich. Denn ab hier beginnt man von abhängigen und unabhängigen Variablen zu sprechen. Die unabhängigen Variablen sind diejenigen, deren Randverteilung festgelegt wird. Die Häufigkeiten auf den Stufen der abhängigen Variablen können dann als multinomial-verteilt betrachtet werden.

Wie man bei log-linearen Modellen mit abhängigen und unabhängigen Variablen umgeht, wird in Kapitel 8 angesprochen.

8. Logit-Modelle

Logit-Modelle stellen einen Spezialfall der log-linearen Modelle dar. Man könnte sie als Entsprechung zur Varianzanalyse auffassen. Die Logit-Modelle finden dort Anwendung, wo die Variablen von vorne herein definiert werden als *abhängige* (hier auch *response*) oder als *unabhängige Variable* (*explanatory*).

Tabelle 10

Krebsprävention (UV)					
	<i>Ja</i>	<i>Nein</i>		<i>Ja</i>	<i>Nein</i>
Alter (AV)	58	29	< 45	3	48
			≥ 45	37	12
				40	60
Wenn ein Zusammenhang besteht zwischen dem Alter und der Frage, ob eine Person bereits an einer Untersuchung zur Krebsprävention teilgenommen hat, so drückt sich das in einem Haupteffekt aus.			Derselbe Zusammenhang drückt sich hier dagegen als Interaktion aus zwischen den kategorialen Variablen Alter und Krebsprävention.		

In der Gegenüberstellung von Tabelle 10 sieht man den Unterschied zwischen einer varianzanalytischen und einer log-linearen Darstellungsweise. Was bei der ANOVA Haupteffekt heißt, heißt dagegen bei den log-linearen Modellen Interaktion.

Im Logit-Modell wird das log-lineare Modell nun so umformuliert, dass aus dieser Interaktion auch hier ein Haupteffekt wird. Der Rest ist Mathematik.

Mathematik, die der unbegabtere Leser bestimmt überspringen kann. Aber der Vollständigkeit halber, will ich doch darauf eingehen. Im Logit werden gar nicht die Zellhäufigkeiten selbst betrachtet, sondern die Quotienten n_{2j}/n_{1j} (bzw. e_{2j}/e_{1j} ; siehe Tabelle 11 und 12)³.

Tabelle 11 - beobachtete Häufigkeiten

	Prävention (UV)	
	<i>Ja</i>	<i>Nein</i>
Alter (AV)	$\frac{n_{21}}{n_{11}} = \frac{37}{3} = 12,33$	$\frac{n_{22}}{n_{12}} = \frac{12}{48} = 0,25$

Tabelle 12 - erwartete Häufigkeiten

	Prävention (UV)	
	<i>Ja</i>	<i>Nein</i>
Alter (AV)	$\frac{e_{21}}{e_{11}} = \frac{19,6}{20,4} = 0,96$	$\frac{e_{22}}{e_{12}} = \frac{29,4}{30,6} = 0,96$
	$\beta_1^{(1)} = \ln(e_{21}/e_{11})$	$\beta_2^{(1)} = \ln(e_{22}/e_{12})$

Nun werden auch diese Quotienten logarithmiert -

$$\ln\left(\frac{e_{2j}}{e_{1j}}\right) = \ln(e_{2j}) - \ln(e_{1j}) \quad (\text{Formel 31})$$

und mit ein bisschen Termumformung läßt sich zeigen, dass auch das logit nur ein Spezialfall eines log-linearen Modells ist.

$$\begin{aligned} \ln(e_{2j}) - \ln(e_{1j}) &= (\lambda_0 + \lambda_2^{(1)} + \lambda_j^{(2)} + \lambda_{2j}^{(12)}) - (\lambda_0 + \lambda_1^{(1)} + \lambda_j^{(2)} + \lambda_{1j}^{(12)}) \\ &= \lambda_2^{(1)} + \lambda_{2j}^{(12)} - \lambda_1^{(1)} - \lambda_{1j}^{(12)} \\ &= 2 \cdot \lambda_2^{(1)} + 2 \cdot \lambda_{2j}^{(12)} \end{aligned} \quad (\text{Formel 32})$$

Setzt man nun $\beta_0 = 2 \cdot \lambda_2^{(1)}$ und $\beta_j^{(1)} = 2 \cdot \lambda_{2j}^{(12)}$ dann erhält man wieder ein simples additives Modell

$$\ln\left(\frac{e_{2j}}{e_{1j}}\right) = \beta_0 + \beta_j^{(1)} \quad (\text{Formel 33})$$

mit einem allgemeinen Effekt und einem Haupteffekt. Und dieser Haupteffekt kann nun interpretiert werden, wie in der ANOVA. Bedeutsamer Haupteffekt bedeutet, zwischen dem Alter

3 Man bedient sich hier einem alternativen Modell, die sogenannten Odds, das eben diese Quotienten einer statistischen Prüfung unterzieht. Aus der Kombination von *log* und *odd* wurde schließlich *logit* (LANGEHEINE 1986).

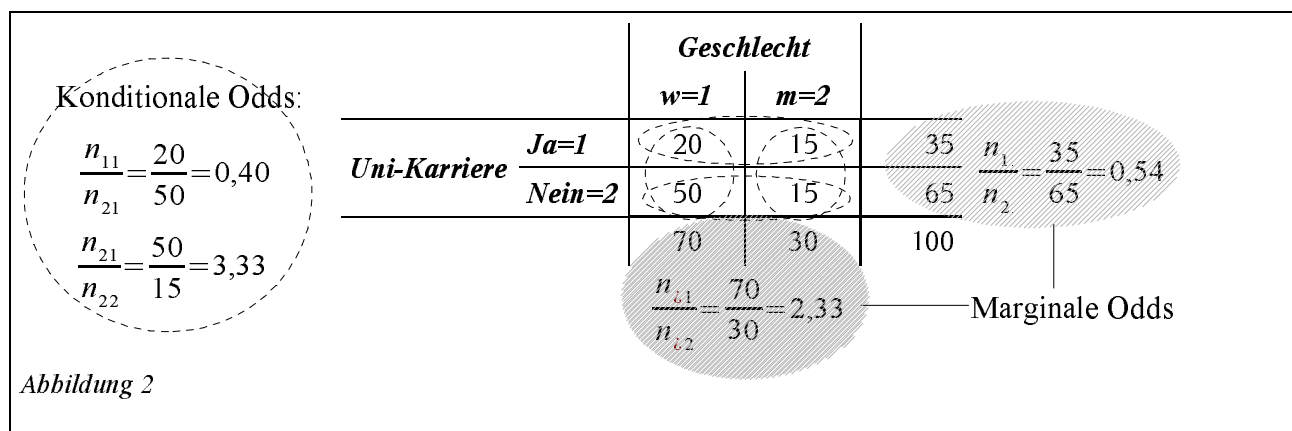
und der Tatsache, ob eine Person schon an einer Krebsvorsorge-Untersuchung teilgenommen hat, zwischen den beiden Variablen existiert also ein Zusammenhang.

Vielleicht sollte dazu noch erläutert werden, dass im logit-Modell nicht wirklich etwas anderes geschieht, als im log-linearen. Die Daten und die statistischen Ergebnisse werden lediglich anders dargestellt, wobei die Art der Interpretation eines Zusammenhanges vorweggenommen wird. Diese Interpretation ist allerdings verschieden von den log-linearen Modellen, denn log-lineare Modelle betrachten keine gerichteten Zusammenhänge von einer unabhängigen Variablen auf eine Abhängige. Logits tun das. Damit sind die Logits in ihrer Logik vergleichbar mit dem varianzanalytischen Ansatz.

9. Odds und Prozentsatzdifferenzen

9.1 Odds

Die Odds gehören zu einem Ansatz, der parallel zu den log-linearen Modellen entwickelt wurde. Sie sind zwar eigentlich nicht notwendig zum Verständnis log-linearer Modelle, aber weil manche Autoren die Odds noch gerne verwenden (z.B. ANDREB, HAGENAARS & KÜHNEL 1997), sollen diese „Dinger“ hier eben auch kurz eingeführt werden.



Odds sind eine Alternative zu den bekannten λ -Parametern (bzw. γ). Sie sind also Effektparameter, die anzeigen, wie groß ein Haupteffekt oder eine Interaktion ist. Allerdings besitzen sie etwas andere Eigenschaften als die λ -Parameter. Denn die Größe von Effektunterschieden ist bei den Odds nicht so leicht interpretierbar. Für λ -Parameter ist ein Unterschied $0,8 - 0,7 = 0,1$ vergleichbar mit einem Unterschied $0,6 - 0,5 = 0,1$. Für Odds gilt das nicht, wegen der absoluten Größe der Zahlen (0,8 ist größer als 0,6).

Generell sind Odds definiert als Verhältnis zwischen verschiedenen Wahrscheinlichkeiten. Es gibt aber verschiedene solcher Odds, je nachdem, ob sie auf einen Haupteffekt oder eine Wechselwirkung hinweisen.

Die *marginalen Odds* errechnen sich aus der Randverteilung. Wären beide Werte der Randverteilung Uni-Karriere gleich, so ergäbe sich $50/50 = 1$, und das würden wir dahingehend interpretieren, dass da kein Haupteffekt existiert. Je unterschiedlicher die beiden Werte der Randverteilung sind, desto stärker unterscheiden sich die Odds von Eins. In unserem Fall wollen ungefähr halb so viele Leute eine Uni-Karriere anstreben ($0,54 \approx 1/2$). Man würde also auf einen Effekt schließen.

Wichtiger als die marginalen Odds sind meistens die *konditionalen Odds*. Die marginalen Odds beschreiben Haupteffekte, die konditionalen Odds beschreiben die einfachen Haupteffekte (siehe Abbildung 2). Bilden wir nun aus den beiden konditionalen Odds nun wiederum einen Quotienten so erhalten wir das sogenannte *Odds Ratio* oder *Kreuzproduktverhältnis*. Das Odds Ratio stellt einen Effektparameter für eine Interaktion dar.

$$\text{Odds Ratio} = \frac{n_{11}/n_{21}}{n_{12}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{20 \cdot 15}{50 \cdot 15} = 0,40 \quad (\text{Formel 34})$$

Je unterschiedlicher die beiden konditionalen Odds sind, desto mehr unterscheidet sich das Odds Ratio von Eins (d.h. Interpretation entsprechend der marginalen Odds). Wie sich diese Odds Ratio auf einen dreifaktoriellen Fall übertragen lässt zeigen ANDREB, HAGENNAARS & KÜHNEL (1997).

Das Odds Ratio hat enge Beziehung zu den log-linearen Modellen. Ich werde das hier nicht ausführen, aber man kann zeigen, dass der Interaktionsparameter in Formel 14 der vierten Wurzel des Odds Ratios entspricht (siehe EID 1997).

Als dritten Fall gibt es noch *partielle Odds*. Partielle Odds stellen den Mittelwert der konditionalen Odds dar. Nun bedeutet Mittelwert in diesem Fall aber nicht das arithmetische Mittel. Hier haben wir ebenfalls ein multiplikatives Modell und deshalb müssen wir die multiplikative Entsprechung des arithmetischen Mittels verwenden, nämlich das sogenannte geometrische Mittel⁴:

$$\bar{x}_{geom} = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n} = \left(\prod x_i \right)^{1/n} \quad (\text{Formel 35})$$

In unserem Beispiel haben wir konditionale Odds von 0,40 und 1,00. Diese beiden miteinander multipliziert und die zweite Wurzel daraus gezogen ergibt 0,63. *Im Durchschnitt* gibt es also mehr Leute, die nicht an der Universität Karriere machen wollen, was sich durch den marginalen Odd von

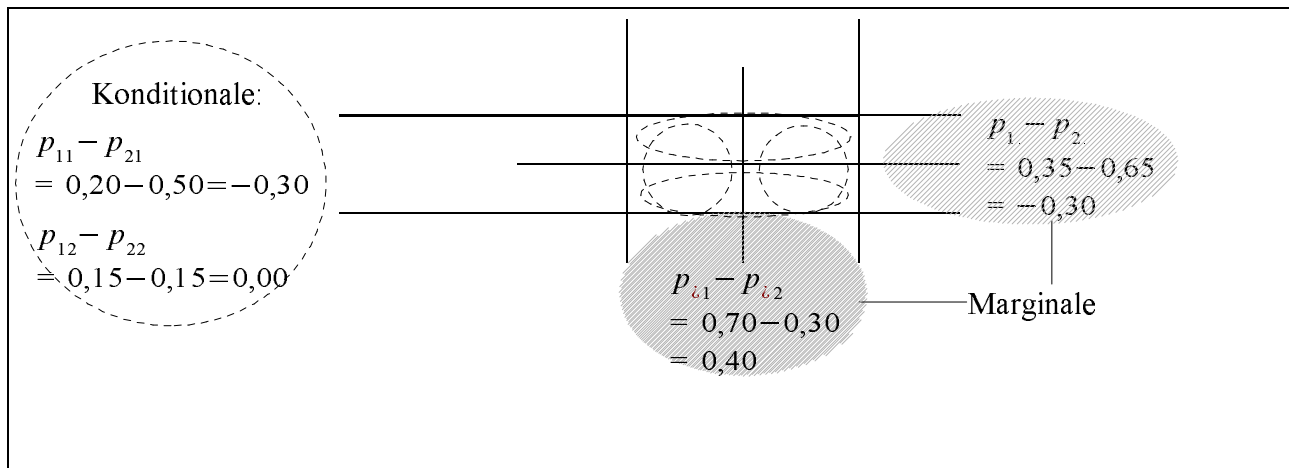
4 Wenn wir unseren alten Trick verwenden und x_{geom} logarithmieren, dann erhalten wir wieder ein additives Modell und aus dem geometrischen Mittel wird (Hokuspokus) das arithmetische Mittel:

$$\ln(\bar{x}_{geom}) = \frac{1}{n} \cdot \sum x_i$$

0,54 noch einmal bestätigt. Es gibt aber Fälle, in denen die partiellen Odds etwas anderes sagen, als die marginalen.

9.2 Prozentsatzdifferenzen

Mit Prozentsatzdifferenzen d% läßt sich das Gleiche ausdrücken, wie mit Odds. Entsprechend zu den Odds gibt es also auch marginale, konditionale und partielle Prozentsatzdifferenzen, die ebenfalls



als Effektmaße zu verstehen und daher ähnlich (aber nur ähnlich) zu interpretieren sind. Die Prozentsatzdifferenzen werden mit dem GSK-Ansatz in Verbindung gebracht. Dieser stellt eine weitere Alternative zu den log-linearen Modellen dar.

Prozentsatzdifferenzen beruhen auf den relativen Häufigkeiten p_{ij} in den einzelnen Zellen. Die relative Häufigkeit $p_{ij} = n_{ij}/n_{..}$ verwendet man anstatt der absoluten Häufigkeit vor allem aus einem Grund: die Effektgrößen würden sich mit der Stichprobengröße verändern. Da der Effekt aber davon unabhängig sein muss, wird die Häufigkeit einfach an Stichprobengröße relativiert.

Die *marginalen* Prozentsatzdifferenzen berechnen sich als Differenz der Werte in der Randverteilung, die konditionalen als Differenz *innerhalb* einer Merkmalsstufe (ganz analog zu den Odds; siehe Abbildung 3). Je kleiner die Effekte, desto näher liegen die Differenzen bei Null.

Ganz analog zu den Odds sind auch die partiellen Prozentsatzdifferenzen, die den Durchschnitt der konditionalen Prozentsatzdifferenzen bilden. Aber (!) da wir es hier nun mit einem additiven Modell zu tun haben, bedienen wir uns des arithmetischen Mittels:

$$\bar{x}_{arith} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \cdot \left(\sum x_i \right) \quad (\text{Formel 36})$$

Und wiederum analog zu den Odds läßt sich auch die Interaktion beschreiben durch die Differenz der beiden konditionalen Prozentsatzdifferenzen. In unserem Fall also $-0,30 - 0,00 = -0,30$.

- 26 **Geschlecht**

		<i>w=1</i>	<i>m=2</i>	
<i>Uni-Karriere</i>	<i>Ja=1</i>	0,2	0,15	0,35
	<i>Nein=2</i>	0,5	0,15	0,65

9.3 Vergleich: Odds, Prozentsatzdifferenzen & λ

An erster Stelle soll hier klargestellt werden, dass alle drei Ansätze zwar dieselben Zusammenhänge beschreiben wollen, dass sie aber unter Umständen zu unterschiedlichen Ergebnissen kommen. Vergleicht man die Effekte, die durch Odds, Prozentsatzdifferenzen und λ s angegeben werden, dann folgen manchmal ganz verschiedene Interpretationen. Eine klare Regel, in welcher Situation welcher Ansatz bevorzugt werden muss, scheint es nicht zu geben. Lediglich für den Fall, dass nur kleine Zellhäufigkeiten vorliegen, bevorzugen ANDREß, HAGENAARS & KÜHNEL (1997, S.25) die Odds ratios gegenüber den Prozentsatzdifferenzen.

Die Wertebereiche der drei Maße unterscheiden sich ebenfalls. Während Prozentsatzdifferenzen symmetrisch von -1 über 0 (kein Effekt) bis +1, und die Lambdas von minus unendlich über Null (ebenfalls kein Effekt) bis plus unendlich variieren, so verteilen sich die Odds asymmetrisch von 0 über 1 (kein Effekt) bis plus unendlich.

Ein wesentlicher Vorteil von Odds und Prozentsatzdifferenzen ist, dass sie recht intuitiv verständlich sind. Ein Odd von 2 bedeutet, dass der eine Wert doppelt so groß ist, wie der andere; eine Prozentsatzdifferenz von 0,10 bedeutet, dass die eine Zelle 10% häufiger besetzt ist, als die Vergleichszelle. Diesen Vorteil verlieren die beiden Ansätze jedoch bei mehrstufigen Merkmalen sofort. Habe ich eine Variable Parteipräferenz mit mehreren Stufen „CDU, SPD, FDP, Grüne, ...“, dann ist es vorbei mit der einfachen Interpretierbarkeit und auch mit der einfachen Berechenbarkeit.

10. Aufgaben

10.1 Aufgabe 1

Als ehrgeiziger Autor von Skripten zur multivariaten Statistik war ich natürlich heiß darauf zu erfahren, welche großartigen Lernvorteile dieses Skript meinen Kommilitonen verschafft. Also habe ich folgende Daten erhoben:

Tabelle 13 -

Variable A: Prüfung bestanden?

Variable B: Lernen mit/ohne Skript?

		B		
		mit	ohne	
A	Ja	18	45	63
	Nein	11	6	17
		29	51	80

- (g) Weil ich dieses Ergebnis (siehe (f)) immer noch nicht akzeptieren kann, hätte ich von Euch noch gerne einen $\lambda_{ij}^{(12)}$. Für diesen soll anschließend *einzel*n getestet werden, ob er von Null verschieden ist. Was sagt uns dieser Kennwert ‘schlussendlich’?

10.1.1 Tipps

Generell läßt sich für diese Aufgabe Einiges aus Abschnitt 2 abgucken. Etwas genauer hier die Angaben.

- (a) Siehe Abschnitt 9.1.
- (b) Siehe Abschnitt 9.1 und 9.3.
- (c) Siehe Kapitel 4 (und Formel 6).
- (d) Siehe Formel 23, 24.
- (e) Siehe Abschnitt 13.
- (f) Abschnitt 2, 13.
- (g) Um diese Aufgabe von Hand zu lösen muss vom saturierten Modell ausgegangen werden. Siehe Abschnitt 3 und 5.4.

10.1.2 Lösungen

- (a) 0,22
- (b) Dieses Skript scheint - nach dem Kreuzproduktverhältnis zu urteilen – eher kontraproduktiv zu sein; es schadet wohl mehr, als dass es nützt.
- (c) $E_{11} = 22,84$; $e_{12} = 40,16$; $e_{21} = 6,16$; $e_{22} = 10,84$
- (d) $\chi^2_{Pearson} = 7,57$; $L^2 = 7,33$
- (e) $df = 1$
- (f) Es gibt einen signifikanten Zusammenhang zwischen ‘Lernen mit Skript’ und ‘Prüfungserfolg’. Allerdings zeigt die Inspektion der Tabelle, dass dieser Effekt in die falsche Richtung geht und den bemühten Studenten eher völlig zu verwirren scheint.

(g) $\lambda_{11}^{(12)} = \ln \left(\frac{18 \cdot \sqrt[4]{18 \cdot 45 \cdot 11 \cdot 6} \cdot \sqrt[4]{18 \cdot 45 \cdot 11 \cdot 6}}{\sqrt[2]{18 \cdot 11} \sqrt[2]{18 \cdot 45} \sqrt[4]{18 \cdot 45 \cdot 11 \cdot 6}} \right) = -0,38$; $s_\lambda = 0,14$

also: $\frac{\lambda_{11}^{(12)}}{s_\lambda} = -2,62 < -1,96$, das ist also signifikant.

10.1.3 Abschließende Bemerkung zu Aufgabe 1

War das denn auch anders zu erwarten? Bei dieser Aufgabe ist nun wirklich so wenig Transferleistung gefragt, dass man bei einem solchen didaktischen Tiefflieger wohl kaum einen pädagogischen Effekt erwarten dürfte. So will ich mich bei der nächsten Aufgabe auf die Wiedergabe zweier kleiner Fragen von Michael Eid beschränken.

Aber keine Angst: die Daten sind frei erfunden, noch besteht Hoffnung.

Übrigens könnt ihr mit der Syntax in ‘Abbildung’ 4 das Beispiel in Spss ausprobieren. Einfach im Menü **Datei→Neu→Syntax (File→New→Syntax)** anklicken und die Zeilen eintippen. Dann auf **Ausführen→Alles (Run→Run All)**, und schon erscheinen die Daten im Tabellenblatt. Damit könnt ihr ganz normal log-lineare Modelle rechnen.

```
DATA LIST FREE /Skript, Bestandn, Anzahl.
BEGIN DATA.
1 1 18
1 2 45
2 1 11
2 2 6
END DATA.
VARIABLE LABELS Skript 'Skript benutzt?' Bestandn 'Prüfung bestanden?'.
VALUE LABELS Skript 1 'Ja' 2 'Nein' / Bestandn 1 'Ja' 2 'Nein'.
WEIGHT by Anzahl.
```

Abbildung 4

10.2 Aufgaben 2

(a) Welches der folgenden log-linearen Modelle ist hierarchisch?

1. $e_{ijk} = \lambda_0 + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)} + \lambda_{ik}^{(13)}$
2. $e_{ijk} = \lambda_0 + \lambda_i^{(1)} + \lambda_k^{(3)} + \lambda_{ik}^{(13)}$
3. $e_{ijk} = \lambda_0 + \lambda_i^{(2)} + \lambda_{jk}^{(23)}$

(b) Angenommen in unten stehender Tabelle gilt folgendes Modell

$e_{ijk} = \lambda_0 + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)} + \lambda_{ik}^{(13)} + \lambda_{jk}^{(23)}$, welcher Wert muss in der fehlenden Zellen stehen?

	<i>A1</i>		<i>A2</i>	
	<i>B1</i>	<i>B2</i>	<i>B1</i>	<i>B2</i>
<i>C1</i>	10	100	15	150
<i>C2</i>	100	20	150	

10.2.1 Lösungen

- (a) 1 und 2 sind hierarchisch; in 3. fehlt der Haupteffekt-Parameter für die Variable (3).
- (b) Im Falle des angegebenen Modells müssen die konditionalen Kreuzproduktverhältnisse für A1 und A2 identisch sein. Denn wäre das nicht der Fall, dann hätten die Variablen (BC) einen anderen Effekt je nachdem, welche Stufe von (A) man betrachtet. Dies würde wiederum bedeuten, dass eine Interaktion dritter Ordnung vorliegen muss, was aber nach dem Modell nicht sein kann. Also lässt sich folgendes berechnen:

$$\frac{10 \cdot 20}{100 \cdot 100} = \frac{15 \cdot ???}{150 \cdot 150}$$
$$??? = \frac{10 \cdot 20 \cdot 150 \cdot 150}{100 \cdot 100 \cdot 15} = \frac{4500000}{150000} = 30$$

11. Literatur

- AGRESTI A. (1990) „*Categorical data analysis*“; New York: Wiley
- ANDERSEN E.B. (1996) „*Log-lineare Modelle*“; in E. ERDFELDER, R. MAUSFELD, T. MEISER & G. RUDINGER „*Handbuch Quantitative Methoden*“; Beltz: Psychologische Verlags Union, Weinheim
- ANDREß H.-J., HAGENAARS J.A. & KÜHNEL S. (1997) „*Analyse von Tabellen und kategorialen Daten*“; Springer
- BORTZ J. (1993) „*Statistik für Sozialwissenschaftler*“; Springer, Berlin
- COHEN (1988) „*Statistical power analysis for the social sciences*“; Erlbaum NY
- EID M. (1997) „*Ergänzendes Skript zum Thema log-lineare Modelle – Vorlesung Forschungsmethoden SS1997*“; unveröffentlichtes Manuskript
- ERDFELDER E. (1984) „*Zur Bedeutung und Kontrolle des –Fehlers bei der inferenzstatistischen Prüfung log-linearer Modelle*“; Zeitschrift für Sozialpsychologie 15, S.18-32
- GEDIGA G. & KUHNT T. (2000) „*Praktische Methodenlehre (WS99/00)*“; <http://luce.psych.uni-osnabrueck.de/ggediga/www/pm98/pages/logreg.htm>
- GOODMAN L.A. (1971) „*The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications*“; Technometrics 13, S.33-61

- HABERMAN S.J. (1979) „*Analysis of qualitative data Vol. 2: New developments*“; New York: Academic Press
- HAGENAARS J.A. (1990) „*Categorical longitudinal data: log linear panel, trend, and cohort analysis*“; Newbury Park: Sage Publ.
- LANGEHEINE R. (1980) „*Log-lineare Modelle zur multivariaten Analyse qualitativer Daten*“; München: Oldenbourg
- LANGEHEINE R. (1986) „*Log-lineare Modelle*“; in J. VAN KOOLWIJK & M. WIEKEN-MAYSER (Hrsg.) „*Kausalanalyse*“ (Techniken der empirischen Sozialforschung Bd. 8); München: Oldenbourg
- RECKE C. (1989) „*Theorie und Anwendung log-linearer Modelle*“; Dissertation Freie Universität Berlin
- STELZ I. (2000) „*What sample sizes are needed to get correct significance levels for log-linear models? - A Monte Carlo Study using the SPSS-Procedure 'Hiloglinear'*“; Methods of Psychological Research Online 5 (2)
- STEVENS, J. (1996) „*Applied multivariate statistics for the social sciences*“; Erlbaum Associates Publ.
- STEYER R. (1983) „*Randomization in the reg-linear and logit-linear models: a farewell to a good old belief*“; Trier: Fachbereich I - Psychologie der Universität Trier (unveröffentlichtes Manuskript, zitiert nach Erdfelder 1984)