

Kapitel 2

Die drei Modellansätze zur Modellierung der Heterogenität

2.1 Allgemeine Modellierung der Heterogenität

Selten tritt in der epidemiologischen Praxis Homogenität auf. In den meisten Studien gibt es einige unberücksichtigte Kovariablen, oder Variablen die nicht erhoben werden können. Des Weiteren können auch bestimmte Unterpopulationen auftreten, die nicht durch Kovariablen erklärbar sind. Mit der Modellierung der unbeobachteten Heterogenität können Subpopulationen oder auch latente Kovariablen gefunden werden. Das Ziel jeder Studie besteht auch darin, die Ergebnisse möglichst auf die Gesamtpopulation auszudehnen. Dabei muss aber das keine-einige-alle Problem (vgl. Hutchinson [21]) beachtet werden. Dieses besagt, dass in einer klinischen Studie nicht alle Teilpopulationen enthalten sein müssen. Es könnte sogar vorkommen, dass nur eine Teilpopulation erfasst ist. Bestimmte Testverfahren können testen, ob in den Daten Heterogenität oder Homogenität vorliegt. Das Verfahren wird als Test der Hypothese $H_0 : m = 1$ (Homogenität) gegen $H_1 : m > 1$ formuliert. Dazu wird in Böhning [8, Seite 117] der χ^2 -Test vorgeschlagen, wobei

$$\chi_{(k-1)}^2 = \sum_{i=1}^k \frac{(y_i - \bar{y})^2}{\sigma_i^2}, \quad (2.1)$$

mit $Var(y_i) = \sigma_i^2$, k steht für die Anzahl der Zentren
und $\bar{y} = \frac{\sum w_i y_i}{\sum w_i}$, $w_i = \frac{1}{\sigma_i^2}$

Hierbei ist y_i das logarithmische Relative Risiko im i -ten Zentrum. Bei diesem Test wird die Varianz σ_i^2 in jedem Zentrum als bekannt vorausgesetzt. Nur im ND-Ansatz, der im

Tabelle 2.1: Chi-Quadrat-Test: H_0 : Homogenität gegen H_1 : keine Homogenität, Konfidenzniveau $\alpha = 0.05$

Daten	χ_{ND}^2	$\chi_{(fg,1-0.05)}^2$	Testergebnis
Tabelle 1.1	23.0396	32.67	nicht ablehnen
Tabelle 1.2	41.7122	32.67	ablehnen
Tabelle 1.3	28.6221	15.51	ablehnen
Tabelle 1.4	29.4897	31.41	nicht ablehnen
Tabelle 1.5	2.33632	12.59	nicht ablehnen
Tabelle 1.6	28.3952	43.77	nicht ablehnen
Tabelle 1.7	27.3812	22.36	ablehnen
Tabelle 1.8	25.5725	19.67	ablehnen
Tabelle 1.9	91.1916	46.19	ablehnen
Tabelle 1.10	24.8867	14.07	ablehnen
Tabelle 1.11	9.40139	14.07	nicht ablehnen
Tabelle 1.12	75.934	76.78	nicht ablehnen
Tabelle 1.13	78.2714	79.08	nicht ablehnen

nächsten Abschnitt beschrieben wird, ist die Varianz im Modell enthalten (2.4). Die Tabelle 2.1 beinhaltet die Ergebnisse des χ^2 -Tests, der an jeder hier vorgestellten multizentrischen Studie durchgeführt worden ist. Der Test zeigt, dass in 6 von den 13 multizentrischen Studien keine Homogenität in den Daten vorliegt. Die unbeobachtete Heterogenität spielt somit keine untergeordnete Rolle und muss adäquat modelliert werden. Dazu wird angenommen, dass der Parameter des Behandlungseffektes eine unbekannte, nicht weiter spezifizierte Verteilung P aufweist. Dies impliziert, dass der Effektparameter eine Rand- (marginal) Dichte besitzt, die anhand einer endlichen Mischverteilung modelliert wird. Diese hat folgende Form:

$$f(y, P) = \sum_{j=1}^m f(y, \theta_j) q_j \quad (2.2)$$

Dabei gibt m die Anzahl der Komponenten an, $P = \begin{pmatrix} \theta_1 & \dots & \theta_m \\ q_1 & \dots & q_m \end{pmatrix}$ ist die mischende Verteilung, θ_j steht für das Relative Risiko der j -ten Komponente und q_j ist das Gewicht der j -ten Komponente. $f(y, \theta)$ wird als *Kern* bezeichnet. Für die Gewichte q_j gilt:

$$q_j \geq 0, \quad j = 1, \dots, m \quad \text{und} \quad \sum_{j=1}^m q_j = 1.$$

Eine Einführung in den Themenkreis diskrete Mischmodelle ist bei Land [24] angegeben. Weitergehende Informationen sind in dem Buch von McLachlan [28] enthalten. Im Gegen-

satz zur endlichen Mischverteilung gibt es noch die stetige Mischverteilung, die folgende Form besitzt:

$$f(y, P) = \int_{-\infty}^{\infty} f(y, \theta)p(\theta)d\theta.$$

Die Dichte $p(\theta)$ ist in der Regel unbekannt. Sie kann aber in bestimmten Situationen durch die Gammafunktion (siehe Böhning [8], Seite 12) ersetzt werden. Da die Gammaverteilung schwer zu rechtfertigen ist, beschränkt sich diese Arbeit auf die diskrete Mischverteilung. In einer multizentrischen Studie ist im Allgemeinen nicht bekannt, welches Zentrum zu welcher Teilpopulation gehört. Dieser Fall wird als unbeobachtete Heterogenität bezeichnet.

Im folgenden werden die drei Modelle vorgestellt. Angefangen wird mit dem herkömmlichen Normalverteilungsansatz. Danach folgt das Kernstück dieser Arbeit, das Profil-Likelihood-Modell. Als letztes wird das häufig angewendete Multi-Level-Modell dargestellt.

2.2 Normalverteilungsansatz

Das ND-Modell geht zurück auf Laird [23] und besitzt als erste Modellannahme, dass das logarithmische Relative Risiko, definiert als

$$y_i = \log\left(\frac{x_i^T}{n_i^T}\right) - \log\left(\frac{x_i^C}{n_i^C}\right), \quad (2.3)$$

normalverteilt ist. Inwieweit diese Annahme bei den hier vorgestellten multizentrischen Studie zutrifft, wird später diskutiert. Der ND-Ansatz besitzt folgende Form:

$$f(y|P, \sigma^2) = \sum_{j=1}^m q_j \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y - \log(\theta_j))^2}{2\sigma^2}}. \quad (2.4)$$

(2.4) weist auf die zweite Modellannahme. In der Dichte ist die unbekannte Varianz des log. RR enthalten, die entsprechend geschätzt werden muss. Unter der Annahme, dass die Beobachtungen x_i^T und x_i^C binomialverteilt mit $x_i^T \sim \text{Bin}(p_i^T, n_i^T)$ und $x_i^C \sim \text{Bin}(p_i^C, n_i^C)$, ist die Varianz gegeben als

$$\begin{aligned} \text{Var}(y_i) &= \text{Var}\left[\log\left(\frac{x_i^T}{n_i^T}\right) - \log\left(\frac{x_i^C}{n_i^C}\right)\right] \\ &= \text{Var}\left[\log\left(\frac{x_i^T}{n_i^T}\right)\right] + \text{Var}\left[\log\left(\frac{x_i^C}{n_i^C}\right)\right]. \end{aligned} \quad (2.5)$$

In dieser Form kann die Formel (2.5) nicht angewendet werden, weil die Varianz von $\log\left(\frac{x_i^T}{n_i^T}\right)$ nicht bekannt ist. Mit Hilfe der δ -Methode ($\text{Var}(g(X)) = g'(E(X))^2 \text{Var}(X)$) kann die Varianz geschätzt werden. Dies führt zu folgender Varianzformel (Böhning [7, Seite 51]):

$$\hat{\sigma}_i^2 = \widehat{\text{Var}}(y_i) = \frac{1}{x_i^T} - \frac{1}{n_i^T} + \frac{1}{x_i^C} - \frac{1}{n_i^C}. \quad (2.6)$$

In den Studien 1.8, 1.9, 1.10 und 1.11, in denen Personenzeiten betrachtet werden, sind die Beobachtungen nicht mehr binomialverteilt. Vielmehr wird den Beobachtungen eine Poissonverteilung unterstellt. Für die Varianzschätzung gilt nun die Formel

$$\widehat{\text{Var}}(y_i) = \frac{1}{x_i^T} + \frac{1}{x_i^C}. \quad (2.7)$$

Diskussion

Zur Verdeutlichung der ersten Modellannahme dient die Abbildung 2.1. In der Grafik ist die empirische Verteilungsfunktion und die theoretische Normalverteilung mit den Parametern μ und σ abgebildet. μ und σ wurden aus den Daten geschätzt. Die Daten stammen aus der Nikotinstudie (Tabelle 1.12) mit 59 Zentren. Aus der Abbildung 2.1 kann nicht abgelesen werden, ob sich die empirische Verteilung signifikant von der Normalverteilung unterscheidet. Dafür gibt es statistische Testverfahren, welche die Anpassungsgüte überprüfen. Ein Anpassungstest ist der Kolmogorov-Smirnov-Test (Schlittgen [32] S. 394-401). Bei diesem heuristischen Test wird der Abstand zwischen der empirischen und der theoretischen Verteilungsfunktion $|F_n(x) - F_0(x)|$ überprüft. Der größte Abstand ist definiert als

$$D = \sup_{-\infty < x < \infty} \left| \hat{F}_n(x) - F_0(x) \right|.$$

Der Kolmogorov-Smirnov-Test (als Test auf die Verteilungsform) kann nur bei Studien mit Effekthomogenität angewendet werden. Unter Heterogenität könnte der Test fälschlicherweise abgelehnt werden, da nicht geprüft wird, ob die logarithmischen Relativen Risiken in den Unterpopulationen normalverteilt sind. In der Tabelle (2.2) sind die multizentrischen Studien überprüft worden, bei denen der Homogenitätstest nicht abgelehnt

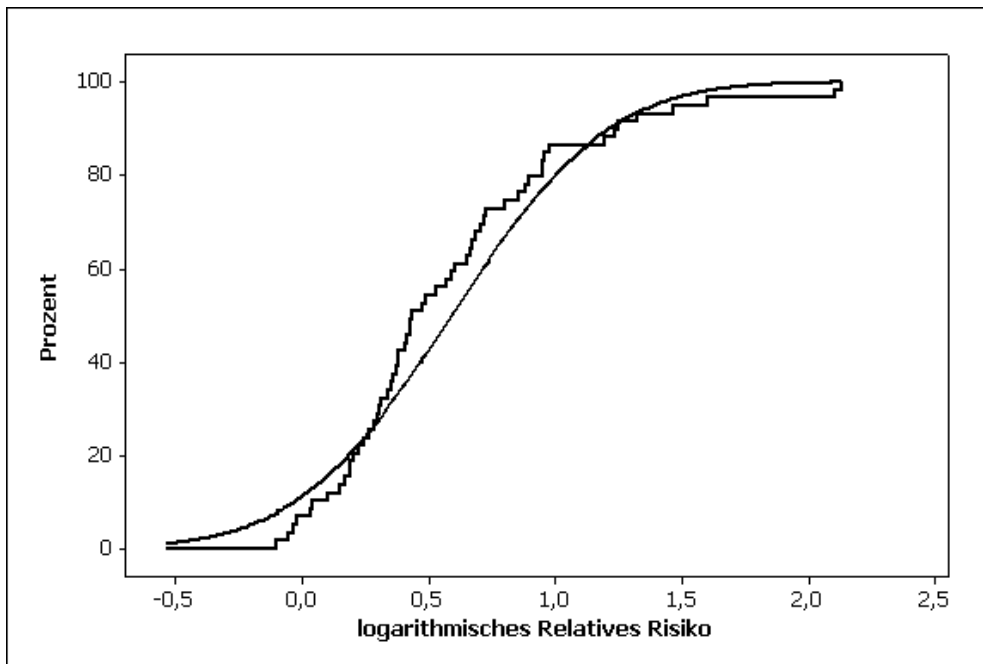


Abbildung 2.1: Empirische Verteilung des log Relativen Risikos im Vergleich mit der Normalverteilung (Daten sind aus Tabelle 1.12)

wurde. Bei der Studie 1.12 wird der Signifikanztest abgelehnt, da die Prüfgröße größer als der kritische Wert von 0.895 (aus Schlittgen [32] Seite 396, Kritische Werte der modifizierten Kolmogorov-Statistik bei Normalverteilung mit unbekanntem Parametern μ und σ) ist. Dennoch kann bei den anderen geprüften multizentrischen Studien der Anpassungstest nicht abgelehnt werden. Somit ist nach dem Kolmogorov-Smirnov-Test die erste wichtige Annahme bei 6 Studien erfüllt.

Des Weiteren gilt auch hier das Gesetz der großen Zahlen. Dieses besagt, je größer der Stichprobenumfang in jedem Zentrum ist, desto eher gilt die Normalverteilungsannahme des log. Relativen Risikos. Inwieweit sich dieses Gesetz auf die Schätzung der Heterogenität auswirkt, wird in der Simulationsstudie Kapitel 5.2 behandelt. Der vielleicht bedeutendste Nachteil in dem Modell ist, dass die Schätzung der Varianz zu schlecht sein könnte, um eine gute Approximation der wahren Varianz zu sein. Diese Ungenauigkeit in der Varianzschätzung wird besonders bei kleinen Stichprobenumfängen in das Modell miteinfließen.

Tabelle 2.2: Anpassungstest an die Normalverteilung bei einem 95% Konfidenzintervall (kritischer Wert bei 0.895)

Daten	D	Prüfgröße	Testergebnis
Tabelle 1.1	0.10157	0.479312	nicht ablehnen
Tabelle 1.4	0.15402	0.710501	nicht ablehnen
Tabelle 1.5	0.168302	0.464038	nicht ablehnen
Tabelle 1.6	0.0985919	0.550654	nicht ablehnen
Tabelle 1.11	0.169934	0.497003	nicht ablehnen
Tabelle 1.12	0.133894	1.02905	ablehnen
Tabelle 1.13	0.089264	0.697528	nicht ablehnen

2.3 Profil-Likelihoodansatz

Der PL-Ansatz ist für ein einfaches Kohortendesign in Böhning [7] erstmalig entwickelt worden. Eine Proportion bzw. Rate lässt sich sehr einfach durch eine Binomial- oder Poissonverteilung modellieren. Somit kann die Heterogenität in der Proportion bzw. Rate als Mischung von Binomial- oder Poissonverteilungen beschrieben werden. Dagegen lässt sich das Verhältnis zweier Raten durch keine elementare Verteilung mehr beschreiben und verbaut damit der Modellierung der Heterogenität den einfachen Zugang. Der Grund liegt in der Likelihoodfunktion. Unter der Annahme, dass die x_i^T [x_i^C] Poissonverteilt sind, mit $x_i^T \sim Poi(n_i^T p_i^T)$ [$x_i^C \sim Poi(n_i^C p_i^C)$] ergibt sich für das i -te Zentrum folgender Poissonlikelihood

$$L_{Poi}(p_i^T, p_i^C)_i = e^{-n_i^T p_i^T} \frac{(n_i^T p_i^T)^{x_i^T}}{x_i^T!} \times e^{-n_i^C p_i^C} \frac{(n_i^C p_i^C)^{x_i^C}}{x_i^C!} \quad (2.8)$$

Die parameterabhängige Poisson-Log-Likelihood ist

$$l_{Poi}(p_i^T, p_i^C)_i = x_i^T \log(p_i^T) - n_i^T p_i^T + x_i^C \log(p_i^C) - n_i^C p_i^C \quad (2.9)$$

Durch die Ersetzung von p_i^T durch $\theta_i p_i^C$ ($\theta_i = \frac{p_i^T}{p_i^C}$) wird der interessierende Parameter, das Relative Risiko eingeführt. Damit ergibt sich

$$\begin{aligned} l_{Poi}(\theta_i, p_i^C)_i &= x_i^T \log(\theta_i p_i^C) - n_i^T \theta_i p_i^C + x_i^C \log(p_i^C) - n_i^C p_i^C \\ &= x_i^T \log(\theta_i) + x_i \log(p_i^C) - (n_i^T \theta_i + n_i^C) p_i^C \\ \text{mit } x_i &= x_i^T + x_i^C \end{aligned} \quad (2.10)$$

Diese Log-Likelihoodfunktion (2.10) hängt von zwei Parametern, einmal vom Relativen Risiko und zum anderen vom Baselineparameter p_i^C oder auch Nuisance-Parameter genannt, ab. Der Nuisance-Parameter ist hier nicht von vordergründigem Interesse. Er ist aber notwendig für die vollständige Beschreibung der Likelihoodfunktion. Mit dem Profil-Maximum-Likelihood-Schätzer (PMLE) kann der Nuisance-Parameter aus dem Modell eliminiert werden. Dazu wird (2.10) nach p_i^C maximiert, indem die Nullgesetzte partielle Ableitung nach p_i^C aufgelöst wird. In dem Fall ist eine geschlossene Lösungsform möglich, mit

$$p_i^C = \frac{x_i}{n_i^C + n_i^T \theta_i}. \quad (2.11)$$

Der Profil-Maximum-Likelihood-Schätzer (2.11) wird nun in die Log-Likelihoodfunktion (2.10) eingesetzt. Das Ergebnis ist eine einfache und nur noch von θ_i abhängige Profil-Log-Likelihoodfunktion, mit der folgenden Form:

$$l_{Poi}(\theta_i)_i = x_i^T \log(\theta_i) - x_i \log(n_i^T \theta_i + n_i^C) \quad (2.12)$$

Alternativ kann (2.11) auch in (2.8) eingesetzt werden, mit

$$L_{Poi}^{Profil}(\theta_i)_i = e^{-x_i} \frac{\left(\frac{n_i^T \theta_i x_i}{n_i^C + n_i^T \theta_i}\right)^{x_i^T} \left(\frac{n_i^C x_i}{n_i^C + n_i^T \theta_i}\right)^{x_i^C}}{x_i^T! x_i^C!} \quad (2.13)$$

Wird auf (2.13) der Logarithmus angewendet erhält man wieder (2.12), wobei die parameterunabhängigen Faktoren ignoriert werden. Dieser Ansatz lässt sich auch auf den Binomiallikelihood anwenden. Dennoch ist der Aufwand weitaus höher, zumal der Profil-Maximum-Likelihood-Schätzer nicht mehr eindeutig ist. Da die Poissonverteilung eine hervorragende Approximation der Binomialverteilung ist, wird diese Modellfehlspezifikation, die nur bei Studien mit binomialverteilten Beobachtungen auftritt, in Kauf genommen.

2.4 Multi-Level-Modell

In diesem Kapitel wird das 2-Level-Modell eingeführt. Die Klasse der Multi-Level-Modelle (ML-Modelle) stammt aus der Familie der verallgemeinerten linearen Modelle (GLM). Die GLM werden in dem Buch von McCullagh und Nelder [26] eingeführt, und sind eine Weiterentwicklung der linearen Modelle. Biostatistische Daten haben oft eine hierarchische Struktur. Diese Struktur kommt in multizentrischen Studien vor, wo Patienten in Zentren eingeteilt sind und die Zentren wiederum als eine Studie aufgefasst. Auch

in Tierpopulationen tritt in natürliche Weise eine hierarchische Struktur auf. So ist ein Tier bestimmt durch dessen Eltern, die Eltern sind Teil einer Gruppe oder Herde und diese können in bestimmten Gebiete oder Zonen vorkommen. Diese verschachtelte Struktur in den Beobachtungen wird im ML-Modell in Einheiten (Levels) klassifiziert. Auf die multizentrischen Studien angewendet, ergeben sich zwei Levels. Das erste Level kann mit folgenden Modell formuliert werden

$$\log \left(\frac{p_x}{1 - p_x} \right) = \text{logit}(p_x) = \alpha + \beta x \quad (2.14)$$

p ist die Erkrankungswahrscheinlichkeit und der Indikator x gibt an, ob p zum Behandlungs- ($x = 1$) oder zum Kontrollarm ($x = 0$) gehört. 2.14 ist essentiell ein einfaches Level-Modell und wird üblicherweise als logistischer Regressionsansatz in der Logitform bezeichnet. α ist in dieser Form der Baseline- und β der Effektparameter. Der Effektparameter ist hier das log. Odds Ratio (Kreuzproduktverhältnis) und nicht das log. Relative Risiko. Über den Intercept- (α) und den Effektparameter kann aber das Relative Risiko berechnet werden. Das Logitmodell wird gewöhnlich bei binomialverteilten Beobachtungen verwendet. Die Dichte eines Zentrums wird formuliert als

$$\begin{aligned} f_{ML_{Bin}}(p_i^T, p_i^C)_i &= \binom{n_i^T}{x_i^T} (p_i^T)^{x_i^T} (1 - p_i^T)^{n_i^T - x_i^T} \\ &\quad \times \binom{n_i^C}{x_i^C} (p_i^C)^{x_i^C} (1 - p_i^C)^{n_i^C - x_i^C}. \end{aligned} \quad (2.15)$$

Die Baselineparameter werden im nächsten Schritt durch die Modellvorgaben ersetzt, mit $p_x = \frac{e^{\alpha + x\beta}}{1 + e^{\alpha + x\beta}}$. Die resultierende Dichte ist

$$\begin{aligned} f_{ML_{Bin}}(\alpha, \beta)_i &= \binom{n_i^T}{x_i^T} \left(\frac{e^{\alpha + \beta}}{1 + e^{\alpha + \beta}} \right)^{x_i^T} \left(\frac{1}{1 + e^{\alpha + \beta}} \right)^{n_i^T - x_i^T} \\ &\quad \times \binom{n_i^C}{x_i^C} \left(\frac{e^{\alpha}}{1 + e^{\alpha}} \right)^{x_i^C} \left(\frac{1}{1 + e^{\alpha}} \right)^{n_i^C - x_i^C}. \end{aligned} \quad (2.16)$$

Im Falle von Poissonverteilten Beobachtungen wird dagegen auf das Logmodell zurückgegriffen. Dieses Multi-Level-Modell hat dementsprechend die Form:

$$\log(p_x) = \alpha + \beta x,$$

mit der Dichte

$$f_{ML_{Poi}}(p_i^T, p_i^C)_i = e^{-p_i^T n_i^T} \frac{(p_i^T n_i^T)^{x_i^T}}{x_i^T!} \times e^{-p_i^C n_i^C} \frac{(p_i^C n_i^C)^{x_i^C}}{x_i^C!}. \quad (2.17)$$

Wieder werden die Baselineparameter durch die Modellvorgaben ($p_x = e^{\alpha+\beta x}$) ersetzt, mit dem Ergebnis

$$f_{ML_{Poi}}(\alpha, \beta)_i = e^{-e^{\alpha+\beta} n_i^T} \frac{(e^{\alpha+\beta} n_i^T)^{x_i^T}}{x_i^T!} \times e^{-e^{\alpha} n_i^C} \frac{(e^{\alpha} n_i^C)^{x_i^C}}{x_i^C!}.$$

Hier ist der Effektparameter β gleichzeitig der Schätzer für das logarithmische Relative Risiko.

Als eine weitere Verallgemeinerung wird angenommen, dass die α - und β - Werte Zufallsvariablen sind und eine bestimmte Verteilung besitzen. Diese Abstraktion wird als zweites Level bezeichnet. Der zufällige Effekt zwischen den Zentren wird hier mit einer Mischverteilung modelliert. Das daraus resultierende Modell hat nun die Form:

$$f_{ML_{Bin}}(P)_i = \sum_{j=1}^m f_{ML_{Bin}}(\alpha_j, \beta_j)_i q_j$$

für den Binomialfall und

$$f_{ML_{Poi}}(P)_i = \sum_{j=1}^m f_{ML_{Poi}}(\alpha_j, \beta_j)_i q_j$$

für den Poissonfall, mit

$$P = \begin{pmatrix} \alpha_1 & \dots & \alpha_m \\ \beta_1 & \dots & \beta_m \\ q_1 & \dots & q_m \end{pmatrix}$$

Damit sind nun die alle drei Modelle vorgestellt. Im nächsten Kapitel werden für die vorgestellten Modelle die Schätzer der mischenden Verteilung abgeleitet.

