Capture-Recapture Ansätze in Prävention und Epidemiologie

Dankmar Böhning

AG Epidemiologie und Biometrie

Institut für Internationale Gesundheitswissenschaften der FU/HU Berlin

Übersicht

- I: Früherkennungsuntersuchungen (Screening Verfahren)
- II: Schätzung der Ausmaße eines Gesundheitsproblems
- III: Vollzähligkeitsproblem bei Registerdaten

Wieviele **Fälle** (Erkrankte) **n** gibt es in einer spezifischen Population?

- Fälle werden durch eine oder mehrere (Informations-) **Quellen** erfasst
  - z. B. niedergelassene Ärzte
  - z. B. Krankenhäuser
    - ein Fall kann durch eine **einzige** Quelle erfasst werden (capture) oder durch **mehrere** Quellen wiedererfasst werden (recapture)
- identifiziert werden n_{obs} Fälle
- mit $n_{obs} < n$ (z.T. auch $n_{obs} << n$)

Problem: Schätzung für n ?

- p_0 Wahrscheinlichkeit einen Fall **nicht** zu erfassen
- **dann:**

$$n = \text{unbeobachtete} + \text{beobachtete Fälle}$$

$$= n\,p_0 + (1 - p_0)\,n$$

$$= n\,p_0 + n_{obs}$$

$$n_{HTP} = n_{obs} / (1 - p_0) \qquad \text{(Horwitz-Thompson)}$$

Horwitz-Thompson ?

- p_0 oft **nicht** bekannt
- Lösung dieses Problems hängt vom **Anwendungsfall** ab

I: Früherkennungsuntersuch-ungen (Screening Verfahren) [mit vollständig bekannten Krankheitsstatus der Testpositiven]

Digitale Rektale Untersuchung und Prostatakrebs (18527 Männer)

(Daten: Smith et al. 1997, Journ. of Urology)

	PK	Kein PK	
DRU +	316	1114	1430
DRU -	229 ?		17097
	545		18527

Annahme: Daten vorhanden zur Angabe der Sensitivität der DRU

Sensitivität = P(DRU+ | PK) = 1- p_0 = 0.58
(nach Brennecke/Schelp: Sozialmedizin 1993)

$$\mathbf{n_{HTE} = n_{obs} / (1- p_0) = 316 / 0.58 = 545}$$

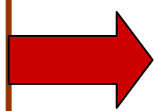
95% KI für n: (493, 597)

Häufig: Sensitivität **nicht** bekannt

Aber mehrere diagnostische Verfahren simultan verfügbar

Digitale Rektale Untersuchung und PSA Serum für Prostatakrebs (18527 Männer)

PK	PSA +	PSA -	
DRU +	179	137	316
DRU -	264	?	
	443		?



kein PK	PSA +	PSA -	
DRU +	138	976	1114
DRU -	717	?	
	855		?

Capture-Recapture Ansatz mit *zwei* Quellen (Screening Verfahren)

Datensituation:

Test 2 \ Test 1	+	-	
+	n_{11}	n_{10}	$n_{1.}$
-	n_{01}	n_{00}	.
	$n_{.1}$		n

Analogie zum CR-Ansatz:

Test positiv *entspricht* Quelle erfasst (Test 1 ist Capture-Stichprobe, Test 2 Recapture-Stichprobe);

n_{00} Anzahl der Personen, die in **keiner** Stichprobe auftauchen

Capture-Recapture Ansatz mit *zwei* Quellen (Screening Verfahren)

- Erfassungswahrsch.

p_{11}	p_{10}	$p_{1.}$
p_{01}	p_{00}	$p_{0.}$
$p_{.1}$	$p_{.0}$	1

- zugehörigen Daten

n_{11}	n_{10}	$n_{1.}$
n_{01}	n_{00}	.
$n_{.1}$		n

unter *Unabhängigkeit:* $p_{11} = p_{1.}\, p_{.1}$

$$n_{11} = n_{1.}/n \times n_{.1}/n$$

diese **Schätzgleichung** führt auf

$$\mathbf{n_{LP} = n_{1.}\ n_{.1} / n_{11}}$$

den Lincoln-Petersen Schätzer der Anzahl der Fälle n

Digitale Rektale Untersuchung und PSA Serum für Prostatakrebs (18527 Männer)

PK	PSA +	PSA -	
DRU +	179	137	316
DRU -	264	202	466
	443	339	782

$$\mathbf{n_{LP} = n_{1\cdot}\ n_{\cdot 1} / n_{11} = 316 \times 443 / 179 = 782}$$

95% KI für n: (724, 840)

Ausblick

- Ansatz erweiterbar auf Situationen mit differenziertem Erkrankungsstadium und multiple diagnostische Verfahren
- **Nachteil** des elementaren LP-Ansatzes: Nichteinhaltung der Randbedingung:

  Summe der geschätzte Fälle und Nicht-Fälle ist oft geringer als die Anzahl der Teilnehmer am Früherkennungsprogramm

DRU und PSA für Prostatakrebs (18527 Männer)

PK	PSA +	PSA -	
DRU +	179	137	316
DRU -	264	202	466
	443	339	782

kein PK	PSA +	PSA -	
DRU +	138	976	1114
DRU -	717	5071	5788
	855	6047	6902

zusammen: **7684**

Ausblick

- Möglicher Grund: positive Korrelation der Screening Tests
- Modifizierter Ansatz erlaubt Korrelation der Screening Tests bei Einhaltung der Randbedingung

- Böhning *et al.* (*Biostatistics*, 2003)

II: Schätzung der Ausmaße eines Gesundheitsproblems

Typische Fragestellungen ...

- „Estimating the number of opiate users in Amsterdam by capture-recapture“, Buster *et al.* (2001), *European Journal of Epidemiology*
- „A Capture-Recapture Analysis of Intussusception after Rotavirus Vaccination“, Verstraeten *et al.* (2001), *American Journal of Epidemiology*
- ...

Epidemiologische Studie einer Hepatitis A - Infektion

- Ort: Umgebung eines College im Nordosten Taiwans
- Zeit: April – Juli 1995
- Umfang: 271 beobachtete Fälle
- Drei Quellen
  - **A**: Serumtest durchgeführt vom nationalen Institut für Präventivmedizin
  - **B**: Fälle von lokalen, niedergelassenen Medizinern gemeldet
  - **C**: Untersuchung (Befragung) von Epidemiologen

Datensituation

(1 = Quelle erfasst)

mehrdimensionale Kreuztabelle

A	B	C	Freq
1	1	1	n_{111}
1	1	0	n_{110}
1	0	1	n_{101}
1	0	0	n_{100}
0	1	1	n_{011}
0	1	0	n_{010}
0	0	1	n_{001}
0	0	0	

Für die Hepatitis A Epidemie

A	B	C	Freq
1	1	1	28
1	1	0	21
1	0	1	17
1	0	0	69
0	1	1	18
0	1	0	55
0	0	1	63
0	0	0	?

Zwei-Quellen-Ansätze

A	B	C	Freq
1	1	1	28
1	1	0	21
1	0	1	17
1	0	0	69
0	1	1	18
0	1	0	55
0	0	1	63
0	0	0	?

A-B

49	86
73	

336

A-C

45	90
81	

378

B-C

46	76
80	

334

Realitätsgerechte Schätzung hier: **1300**

Log-lineare Modellierung (mit drei Quellen)

$$\log E(n_{ijk}) = \lambda + \lambda_A I_{\{i=0\}} + \lambda_B I_{\{j=0\}} + \lambda_C I_{\{k=0\}} + \lambda_{AB} I_{\{i=0\}} I_{\{j=0\}} + \lambda_{AC} I_{\{i=0\}} I_{\{k=0\}} + \lambda_{BC} I_{\{j=0\}} I_{\{k=0\}}$$

Auswahl eines Modells: z. B.

$$\log E(n_{ijk}) = \lambda + \lambda_A I_{\{i=0\}} + \lambda_B I_{\{j=0\}} + \lambda_C I_{\{k=0\}} + \lambda_{AB} I_{\{i=0\}} I_{\{j=0\}}$$

Prädiktion:

$$\log \hat{n}_{000} = \hat{\lambda} + \hat{\lambda}_A + \hat{\lambda}_B + \hat{\lambda}_C + \hat{\lambda}_{AB}$$

Vorteil: Modellierung der Abhängigkeitsstruktur

- durch Interaktionsterme im log-linearen Modell

Nachteil: Modellevaluation

- Bayesianische Informationskriterium:

$$BIC = 2 \text{ log-likelihood} - p \log (n_{obs})$$

p = Anzahl der Parameter im Modell

Einige mögliche Modelle

Name	Nummer	Terme im Modell
Konstante	1	λ
Vollst. Unabh.	2	$+\lambda_A +\lambda_B +\lambda_C$
A-C, B-C Unabh.	3	$+\lambda_{AB}$
A-B, B-C Unabh.	4	$+\lambda_{AC}$
A-B, A-C Unabh.	5	$+\lambda_{BC}$
B-C Unabh.	6	$+\lambda_{AC}+\lambda_{BC}$
A-C Unabh.	7	$+\lambda_{AB}+\lambda_{BC}$
A-B Unabh.	8	$+\lambda_{AC}+\lambda_{BC}$
A,B,C abh.	9	$+\lambda_{AC}+\lambda_{AB}+\lambda_{BC}$

Ergebnisse für Hepatitis A

Modell-Nr.	BIC	$\hat{n}_{000}$	$\hat{n}$
1	-123.33	39	310
2	-84.16	117	388
3	-86.55	145	416
4	-89.65	122	393
5	-86.73	142	413
6	-90.43	193	464
7	-84.21	256	527
8	-90.91	181	452
9	-76.61	1042	1313

Ausblick

- log-lineare Modellierung brauchbarer Ansatz für *mittlere* Quellenanzahl (3-4)
- bei sehr vielen Quellen wird Modellauswahl problematisch

- Böhning und Schön (2002), *Preprint des RKI*

III: Vollzähligkeitsproblem bei Registerdaten

Anwendung: Krebsregisterdaten des Saarlandes

Quelleninformationen in einem Krankheits(Krebs)register

- Ein Fall wird auf Grund wenigstens einer von mehreren Quellen wie Pathologie, Krankenhäusern, niedergelassenen Ärzten, Totenschein, ... in das Register aufgenommen
- weitere Kovariate möglicherweise verfügbar wie Alter bei Diagnosestellung, Geschlecht, ...

Datensituation

ID	Quelle A	Quelle B	Quelle C	...	Anzahl: Quellen
001	1	0	0	...	1
002	0	1	1	...	2
003	0	0	0		0
004	1	0	1	...	2
005	1	1	1	...	3
...	...	...	...	...	...

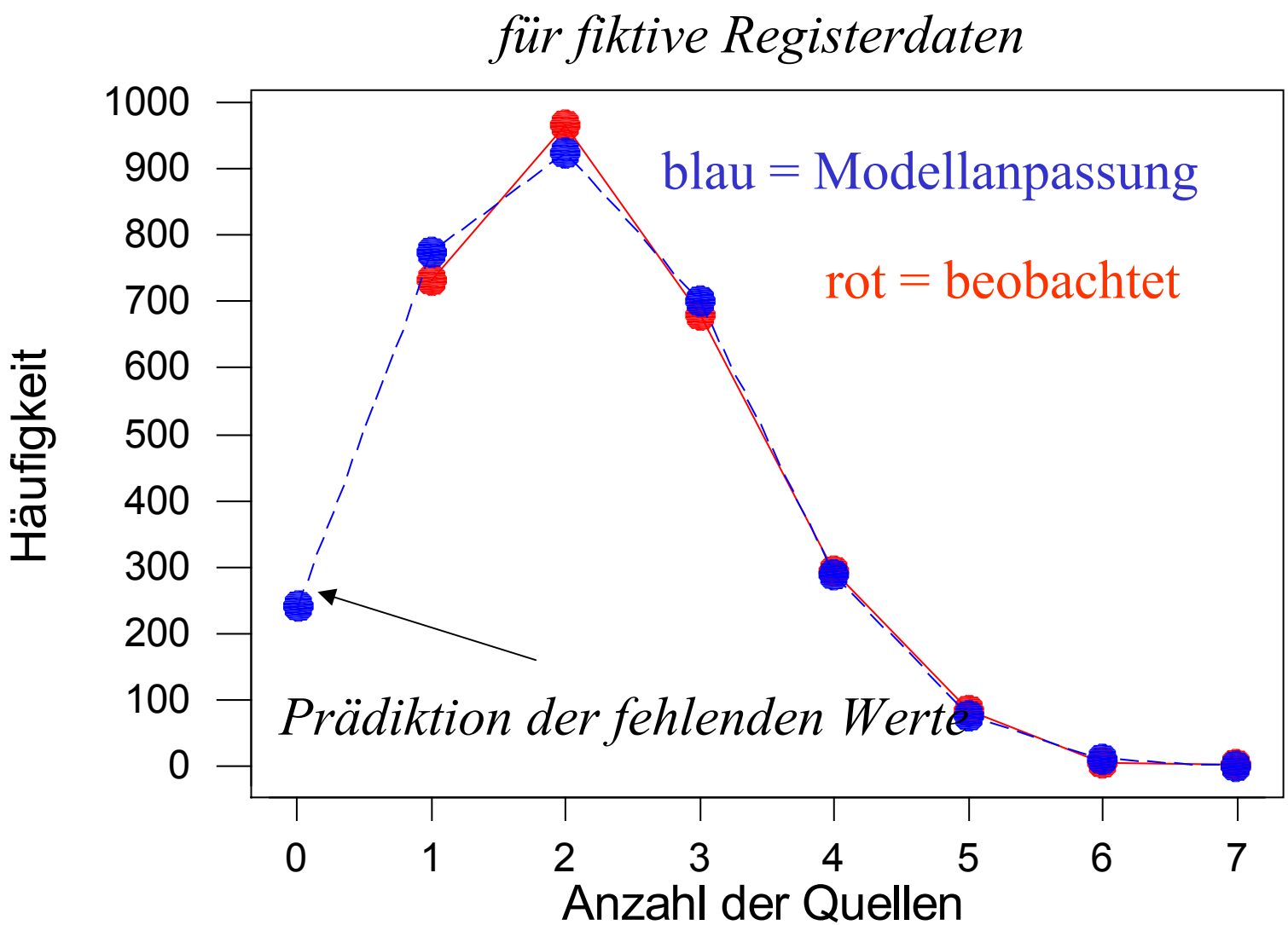
Die Zählverteilung

ID	Quelle A	Quelle B	Quelle C	...	Anzahl: Quellen
001	1	0	0	...	1
002	0	1	1	...	2
004	1	0	1	...	2
005	1	1	1	...	3
...	...	...	...	...	...

Warum Zählverteilung ?

- Log-lineare Modellierung wird komplex bei *vielen* Quellen; mit 4 Quellen hat man > 100, mit 5 Quellen mehr als 7000 mögliche Modelle
- Zählverteilung erfordert grundsätzlich nur eine Entscheidung über das Modell
- Zählverteilung erscheint robuster gegenüber Quellenauswahl (Totenschein)

Verteilung der beobachteten und angepassten Zählverteilungsdaten



Einfache Verteilungsmodelle für Zähldaten

Binomial

$$f(y,\theta) = \binom{m}{y} \theta^y (1-\theta)^{m-y}, \quad y=0,1,\ldots,m$$

(m = Anzahl der Quellen, θ = Erfassungswahrsch.)

Prädiktionswahrsch. für eine Null:

$$p_0 = f(y=0,\theta) = \binom{m}{0} \theta^0 (1-\theta)^{m-0} = (1-\theta)^m$$

Anzahl der Fälle n

Angenommen, θ wäre *bekannt*, dann

geschätzte Anzahl von Fällen

$$\hat{n} = n_{obs} / (1 - p_0)$$

mit $p_0 = (1-\theta)^m$.

Schätzung der Erfassungswahrscheinlichkeit θ

wäre n *bekannt*, dann kann θ durch den Mittelwert geschätzt werden:

$$\hat{\theta} = (n_0\ 0 + n_1\ 1 + n_2\ 2 + \ldots + n_m\ m)/(n\ m)$$

Schätzalgorithmus

Schritt 0. Festlegung des Anfangswertes
für $\theta = \theta^{(1)}$ (z. B. $\theta = 1/2$), $j=1$.

Schritt 1. Berechne $n^{(j)} = n_{obs} / (1 - p_0^{(j)})$,
wobei $p_0^{(j)} = (1-\theta^{(j)})^m$

Schritt 2. Berechne
$\theta^{(j+1)} = (n_1\ 1 + \ldots + n_m\ m)/(m\ n^{(j)})$

setze $j=j+1$, und gehe nach Schritt 1.

Version des EM Algorithmus (DLR 1977)

- Zur Konstruktion des Maximum Likelihood Schätzers von θ
  - Schritt 1: E-Schritt
  - Schritt 2: M-Schritt
- Imputation der fehlenden Daten (als Nebenprodukt)
- (starke) Konvergenz ist gesichert (für diesen Fall) (Dietz *et al.* 2000, CSDA)

Einfache Verteilungsmodelle für Zähldaten

Poisson

$f(y,\theta) = e^{-\theta}\,\theta^{y} \,/\, y\,!\,,\ y=0,1,\ldots$

(geeignet für großes m = Anzahl der Quellen)

Prädiktionswahrsch. für eine Null:

$p_0 = f(y=0,\theta) = e^{-\theta}\,\theta^{y} \,/\, y\,! \;= e^{-\theta}$

Schätzung der Parameter

Ähnlich zur Binomialverteilung:

einziger Unterschied ist die Prädiktion der fehlenden Werte:

Poisson: $p_0 = f(y=0,\theta) = e^{-\theta}$

Binomial: $p_0 = f(y=0,\theta) = (1-\theta)^m$

Größere Flexibilität und Robustheit durch Mischungen

- Einfache Zählverteilungen wie die Binomial oder Poisson treffen wenig *realitätsgerechte* Annahmen der Homogenität der Erfassungswahrscheinlichkeiten
- unterschiedliche Erfassungswahrscheinlichkeiten der Quellen in unterschiedlichen Subpopulationen erlaubt realitätsgerechtere Modellierungen

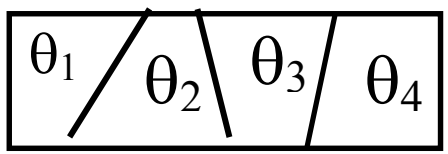
Mischverteilungsmodellierung in zwei Sätzen

Homogenität

einparametrige Dichte $f(y,\theta)$

(z.B. $f(y,\theta)$ = Binomial oder Poisson)

Heterogenität



Dichte in Subpopulation j: $f(y,\theta_j)$

Mischungen in zwei Sätzen

latente Variable Z beschreibt Populationszugehörigkeit zu k Subpopulationen

gemeinsame Dichte f(x,z) schreiben als:
$f(x,z) = f(x|z)f(z) = f(x,\theta_z)q_z$

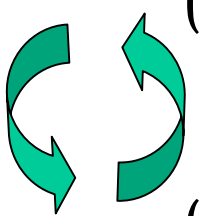
Marginal- oder Mischverteilungsdichte:
$f(x,Q) = f(x,\theta_1)q_1 + \ldots + f(x,\theta_k)q_k$

mit $Q = \begin{pmatrix} \theta_1 \ldots \theta_k \\ q_1 \ldots q_k \end{pmatrix}$ als *mischender Verteilung*

Schätzalgorithmus (Mischungen)

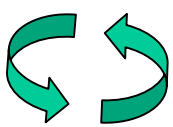
Geschachtelter EM Algorithmus

Schritt 0. Festlegung des Anfangswertes
für Q =$Q^{(1)}$ (z. B. Q empirische Dichte)



(E-)Schritt 1. Berechne $n^{(j)} = n_{obs} / (1 - p_0^{(j)})$,
wobei $p_0^{(j)} = f(0, Q^{(j)})$

(M-)Schritt 2. Berechne (NP)MLE $Q^{(j+1)}$ von Q für
imputierte Daten: $n_0^{(j)}, n_1, \ldots, n_m$



Schritt 2.1: E-Schritt für Mischungen.
Schritt 2.2: M-Schritt für Mischungen.

setze j=j+1 und gehe nach Schritt 1.

Spezielle Mischungen

von Binomialdichten

$$f(y, \theta_j, q_j) = \sum_{j=1}^{k} q_j \binom{m}{y} \theta_j^y (1-\theta_j)^{m-y}, \quad y=0,1,\ldots,m$$

(m = Anzahl der Quellen,
θ_j = Erfassungswahrscheinlichkeit in Subpopulation j,
q_j = Gewicht der Subpopulation j)

Prädiktionswahrscheinlichkeit für eine Null:

$$p_0 = f(y=0, \theta_j, q_j) = \sum_{j=1}^{k} q_j (1-\theta_j)^m$$

Spezielle Mischungen

von Poissondichten

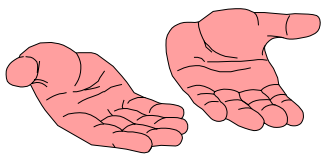
$$f(y, \theta_j, q_j) = \Sigma_{j=1}^{k} \; q_j \exp(-\theta_j) \; \theta_j^y \; / y \; ! \; , \; y=0,1, \ldots$$

(θ_j = Erfassungsparameter in Subpopulation j,
q_j = Gewicht der Subpopulation j)

Prädiktionswahrscheinlichkeit für eine Null:

$$p_0 = f(y=0, \theta_j, q_j) = \Sigma_{j=1}^{k} \; q_j \exp(-\theta_j)$$

Beispiel für geschachtelten EM Algorithmus



Start C.A.MUST

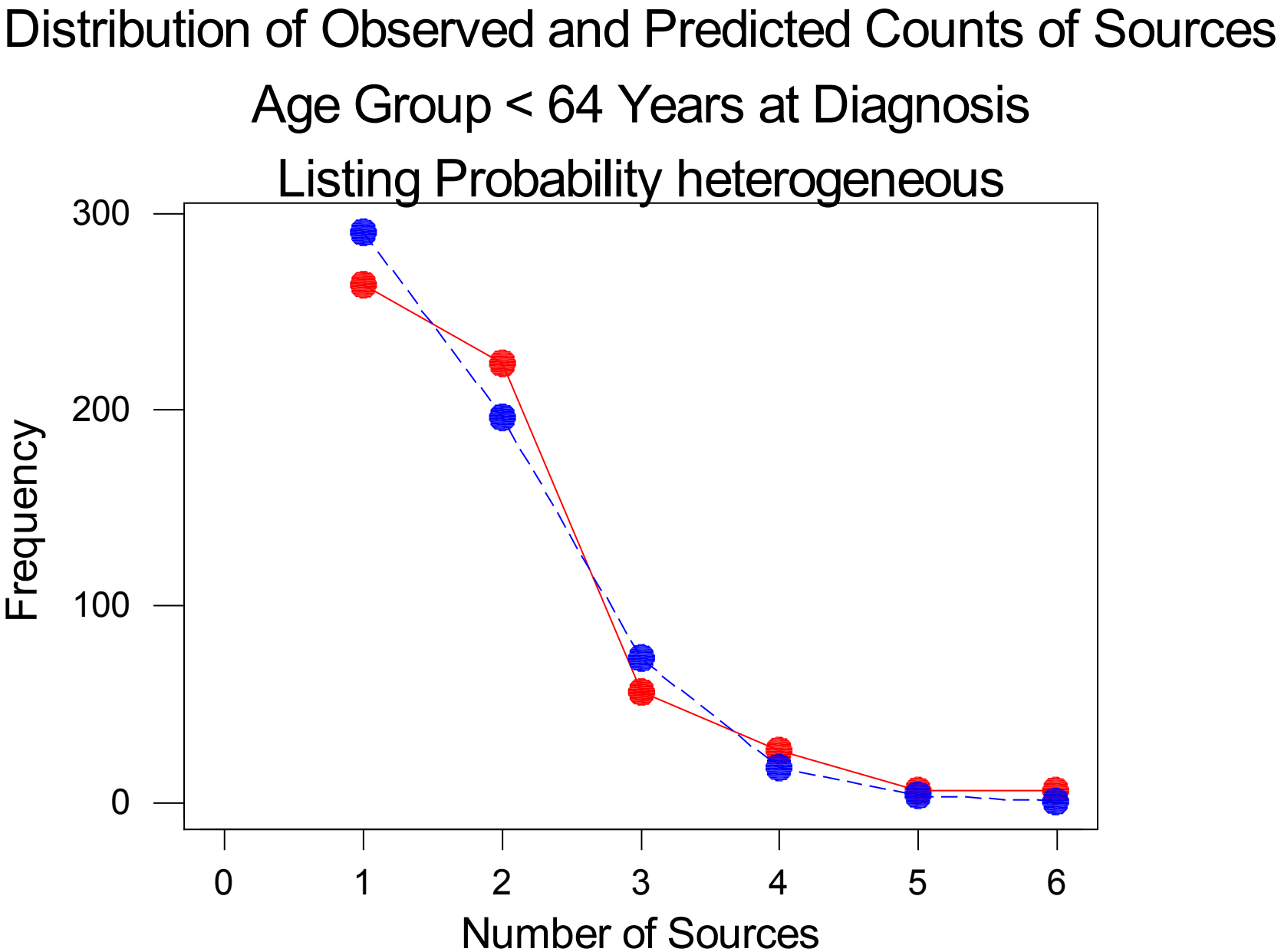
Hasselblad-Daten

0	162
1	267
2	271
3	185
4	111
5	61
6	27
7	8
8	3
9	1

Ergebnisse beim Krebsregister Saarland

- Gemeinsames Projekt mit Robert Koch Institut Berlin, Dachdokumentation Krebs (Dr. Dieter Schön)
- Sechs Hauptquellen aus 40 Krankenhaus-kategorien und 31 Abteilungskategorien
  - Zählvariable selten > 10
- Zeitrahmen: 1994 - 1998
- Hier 3 Lokalisationen: Lunge, Brust (w.) und die Prostata

Prostata - Krebs

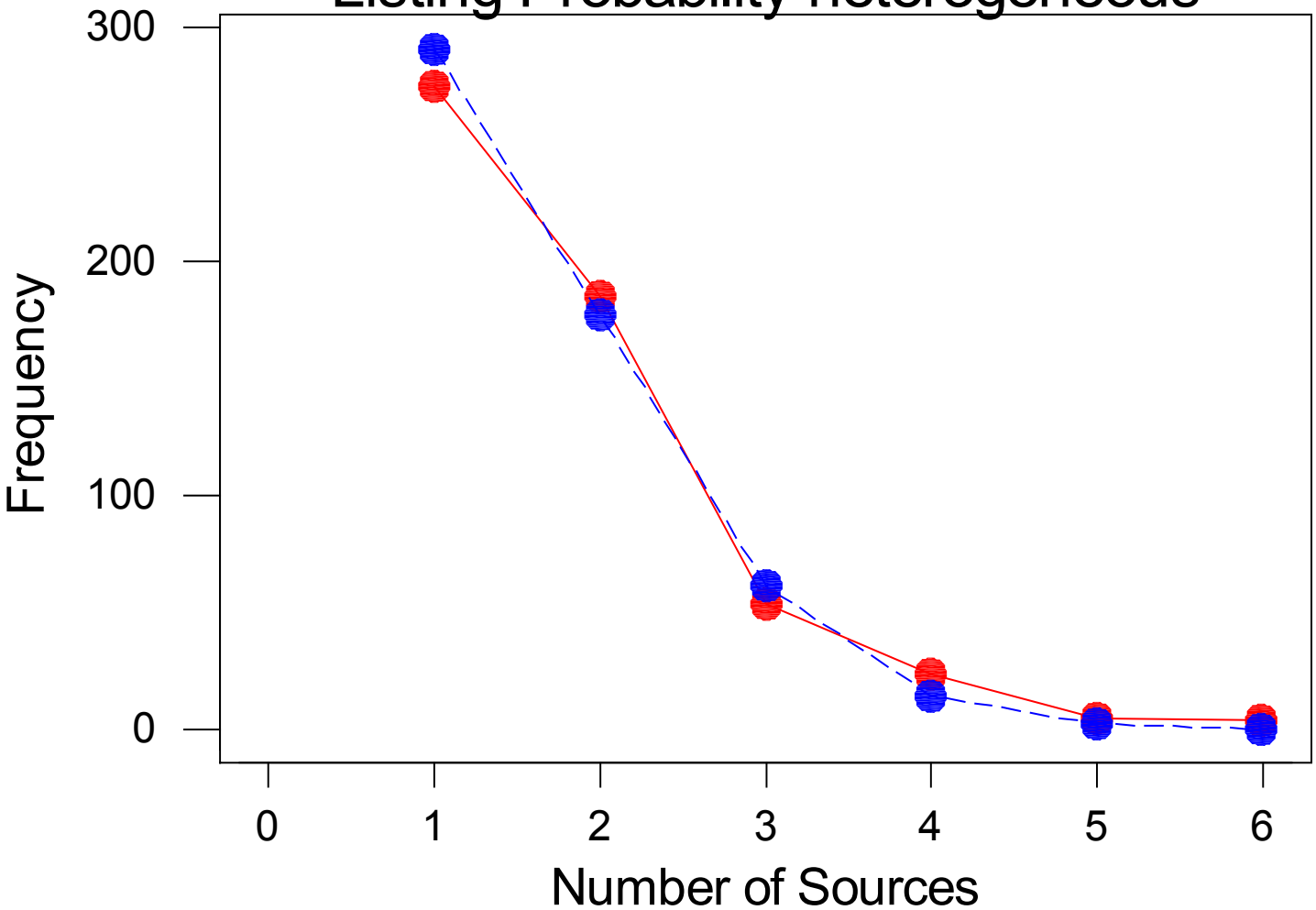


Prostate Cancer

Distribution of Observed and Predicted Counts of Sources

Age Group > 64 and < 70 Years at Diagnosis

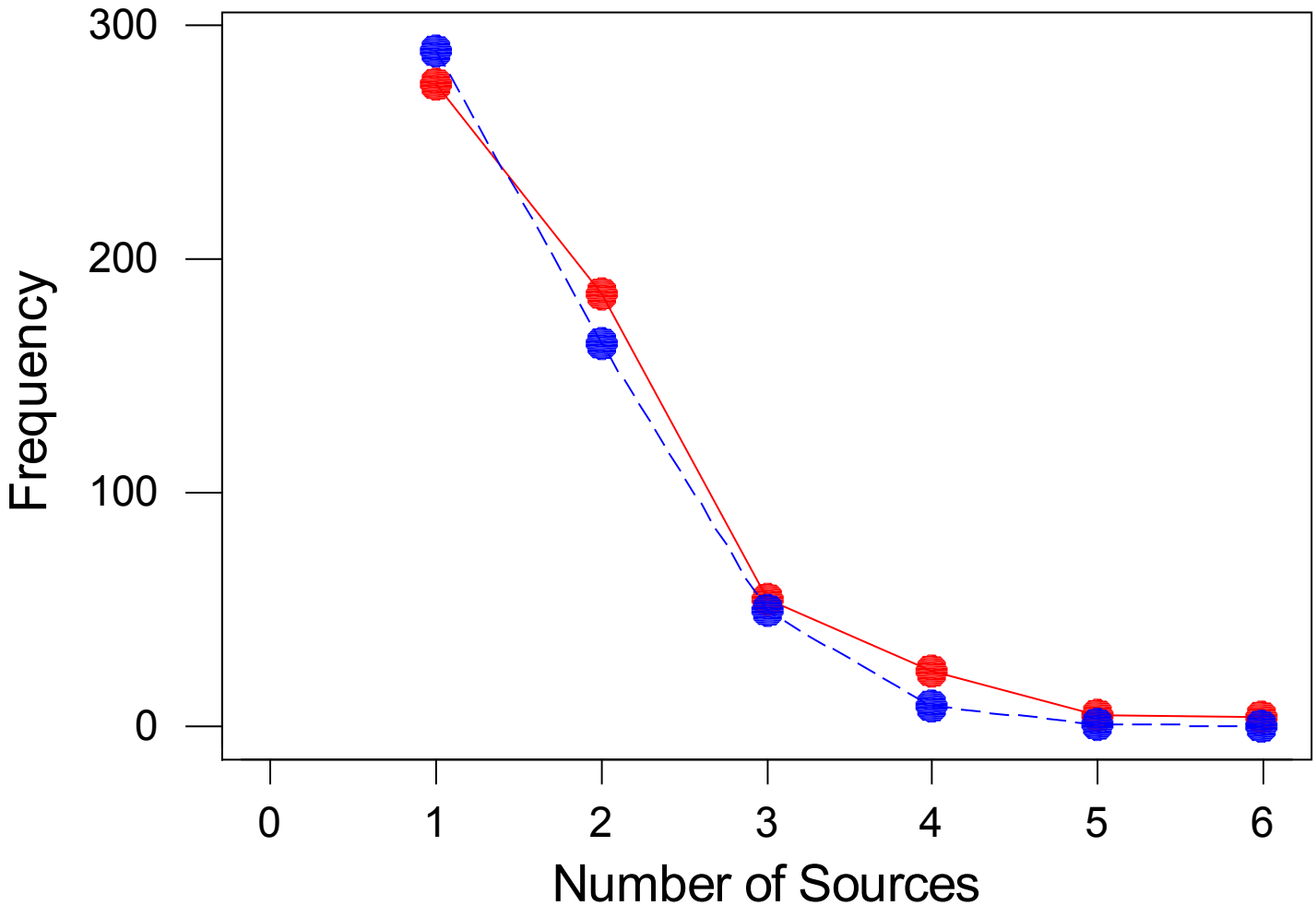
Listing Probability heterogeneous



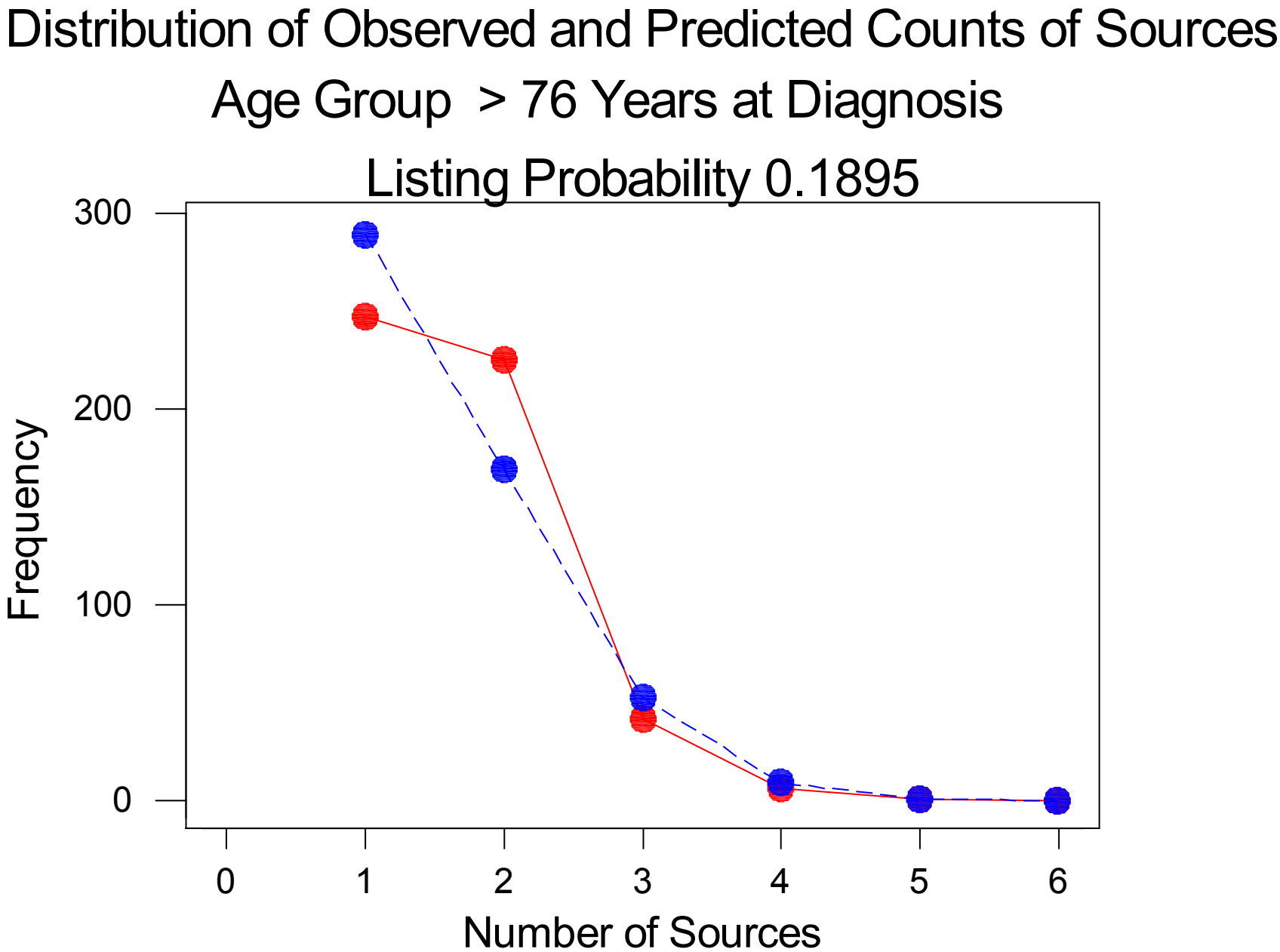
Prostate Cancer

Distribution of Observed and Predicted Counts of Sources

Age Group > 70 and < 76 Years at Diagnosis
Listing Probability 0.1848



Prostate Cancer



Prostate Cancer

Zusammenfassung für Prostatakrebs

Alter	n_{obs}	n_o	LP 1.	LP 2.	Gewicht 2. Komp.	Vollzählig-keit (%)
- 64	582	181	0.2111	0.5534	0.0184	76.2779
- 70	547	203	0.1910	0.4525	0.0368	72.9333
- 76	512	213	0.1848	-	1.0000	70.6207
> 76	521	206	0.1895	-	1.0000	71.6644
Total	**2162**	**803**	**0.2010**	**-**	**1.0000**	**72.9174**

Zusammenfassung Lungenkrebs

Alter	n_{obs}	n_o	Erfassungs wahrsch.	Vollzählig-keit (%)
- 59	866	47	0.3684	94.8521
- 67	926	64	0.3234	93.5354
- 73	792	67	0.3057	92.2002
> 73	785	143	0.2680	84.5905
Total	**3369**	**321**	**0.3125**	**91.3008**

Zusammenfassung Brustkrebs

Alter	n_{obs}	n_o	Erfassungs-wahrschein.	Vollzählig-keit (%)
- 52	877	91	0.2959	90.5992
- 63	942	101	0.2917	90.3164
- 73	845	121	0.2594	87.4741
> 73	817	143	0.2215	85.1042
Total	**3481**	**456**	**0.2687**	88.4176

Zusammenfassung

- für *wenige Quellen* (< 4-5) erscheint die log-lineare Modellierung brauchbar
- falls die *Anzahl der Quellen groß* wird: Zählverteilungsmodellierung attraktiver (durchaus historische Parallelergebnisse in den Sozialwissenschaften)
- Mischverteilungsmodelle: *Flexibilität* beim Abfangen von möglicher Heterogenität und Quellenabhängigkeit
- Böhning, Schön *et al.*(2003) *Ann. Inst. Stat. Math.*

Ausblick und offene Fragen

- *unterschiedliche Verteilungsfamilien* für Zählvariablen (z. B. erlauben Mischungen nur das Auffangen von Überdispersion)
- *zuverlässige* Bereitstellung der Standardfehler
- Evaluation der vorgeschlagenen Schätzmethoden auf der Basis *realer Registerdaten*

Danke für die Aufmerksamkeit!