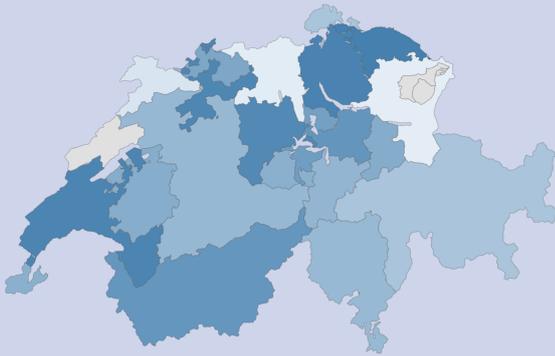


# Überprüfung der Methodik im Schweizer Atlas der Gesundheitsversorgung



## Schlussbericht

zuhanden des Schweizerischen Gesundheitsobservatoriums (Obsan)

30. März 2023

## Impressum

### **Bibliographische Angaben**

Schoch, T. und Thees, O. (2023). Überprüfung der Methodik im Schweizer Atlas der Gesundheitsversorgung, Olten. Bericht zuhanden des Schweizerischen Gesundheitsobservatoriums (Obsan).

### **Projektteam FHNW**

Tobias Schoch

Oscar Thees

### **Auftraggeber**

Schweizerisches Gesundheitsobservatorium (Obsan)

Espace de l'Europe 10

CH-2010 Neuchâtel

Der Bericht gibt die Auffassung der Autoren/Innen wieder, die nicht notwendigerweise mit derjenigen des Auftraggebers oder der Begleitorgane übereinstimmen muss.

Fachhochschule Nordwestschweiz

Hochschule für Wirtschaft

Institut ICC

Riggenbachstrasse 16

CH-4600 Olten

Schweiz

[www.fhnw.ch/de/die-fhnw/hochschulen/hsw/icc](http://www.fhnw.ch/de/die-fhnw/hochschulen/hsw/icc)

Prof. Dr. Tobias Schoch

Tel.: +41 (0) 62 957 21 02

E-Mail: [tobias.schoch@fhnw.ch](mailto:tobias.schoch@fhnw.ch)

# Inhaltsübersicht

<b>Abkürzungsverzeichnis</b>	<b>4</b>
<b>1 Einleitung</b>	<b>5</b>
1.1 Kriterien für die Methodenwahl . . . . .	6
1.2 R-Package sava . . . . .	6
1.3 Aufbau des Berichts . . . . .	6
<b>2 Grundlagen</b>	<b>7</b>
2.1 Population, Indikatoren und Daten . . . . .	7
2.2 Verabredung der Notation . . . . .	8
2.3 Schätzung der Raten . . . . .	9
2.4 Standardisierung der Raten . . . . .	11
2.5 Deskriptive Datenanalyse . . . . .	18
<b>3 Modellierung</b>	<b>22</b>
3.1 Systematic component of variation . . . . .	22
3.2 Poisson–Gamma Empirical Bayes . . . . .	28
3.3 Normal–Normal Empirical Bayes . . . . .	32
<b>4 Analyse</b>	<b>37</b>
4.1 SCV . . . . .	37
4.2 Normal–Normal und Poisson–Gamma Empirical Bayes . . . . .	40
4.3 Quantifizierung . . . . .	44
<b>5 Schlussfolgerungen und Empfehlungen</b>	<b>46</b>
<b>Literaturverzeichnis</b>	<b>48</b>
<b>Anhänge</b>	<b>51</b>
<b>A Test-Indikatoren</b>	<b>51</b>
<b>B Methoden</b>	<b>52</b>
<b>C Diagnostikplot nach Gail (1978)</b>	<b>56</b>
<b>D Verteilung</b>	<b>66</b>
<b>E R-Package: sava</b>	<b>68</b>

## Abkürzungsverzeichnis

<b>BFS</b>	Bundesamt für Statistik
<b>CIHB</b>	Conditionally independent hierarchical bayes
<b>EB</b>	Empirical Bayes
<b>HSA</b>	Hospital Service Area (Spitalregion)
<b>ISS</b>	Interne, indirekte Standardisierung
<b>ML</b>	Maximum Likelihood
<b>MedStat</b>	Medizinische Statistik der Krankenhäuser
<b>MSE</b>	Mean Square Error (mittlerer quadrat. Fehler)
<b>Obsan</b>	Schweizerisches Gesundheitsobservatorium
<b>QQ-Plot</b>	Quantil-Quantil-Plot
<b>QR</b>	Quintile Ratio
<b>SAE</b>	Small Area Estimation
<b>SAVA</b>	Small Area Variation Analysis
<b>SCV</b>	Systematic Component of Variation
<b>SMR</b>	Standardized Mortality Ratio
<b>STATPOP</b>	Statistik der Bevölkerung und der Haushalte
<b>VI</b>	Vertrauensintervall

# 1 Einleitung

Die Inanspruchnahme von Gesundheitsleistungen weist grosse regionale Unterschiede auf. Der Schweizer Atlas der Gesundheitsversorgung<sup>1</sup> (herausgegeben vom Schweizerischen Gesundheitsobservatoriums Obsan und dem Bundesamt für Gesundheit) bildet seit dem Relaunch im April 2023 einen grossen Teil der medizinischen Versorgung und deren Inanspruchnahme in über 100 Indikatoren ab. Die Indikatoren fokussieren auf die somatische Versorgung und umfassen sowohl stationäre als auch ambulante Behandlungen. Die Definition der Indikatoren erfolgte gestützt auf die Fachliteratur sowie unter Einbezug der medizinischen Fachgesellschaften und weiteren Expertinnen und Experten.<sup>2</sup>

## Standardisierung, Aggregierung und Darstellung

Die Indikatoren bilden die zentrale Grundlage, um geografische Muster bei der Inanspruchnahme von Leistungen der Gesundheitsversorgung zu erkennen und das Ausmass der regionalen Unterschiede zu quantifizieren. Die Daten zu den Indikatoren müssen jedoch vorgängig durch geeignete statistische Methoden standardisiert, aggregiert und dargestellt werden. Dies kann bspw. durch die Berechnung von Behandlungsraten erfolgen, die dann kartografisch dargestellt werden können.

## Zielsetzung

Im Rahmen des Relaunch des Versorgungsatlas sollte die statistische Methodik überprüft und gegebenenfalls angepasst werden. Der vorliegende Bericht hat sich dieser Zielsetzung angenommen und diskutiert statistische Methoden, um Raten und Kennzahlen zur Quantifizierung der regionalen Variation im Kontext des Versorgungsatlas zu berechnen. Dabei orientieren wird uns an den Anforderungen, die Lawson et al. (2000) an die Erstellung von *Disease Maps* (einem verwandten Forschungsgebiet) formuliert haben. Die wichtigsten Anforderungen aus Lawson et al. (2000, S. 2217) können wie folgt auf den Punkt gebracht werden: Disease Maps bzw. kartographische Darstellungen zum Versorgungsatlas:

- i) “describe the spatial variation in disease incidence [...]”;
- ii) “identify areas of unusually high risk so that action may be taken”;
- iii) “provide a „clean“ map of disease risk in a region to allow better resource allocation and risk assessment”

---

<sup>1</sup> Siehe <https://www.versorgungsatlas.ch>

<sup>2</sup> Siehe Bericht: Kaiser B., Essig S. und Zwahlen M. (2022). Indikatorenset für den Relaunch des Schweizerischen Versorgungsatlas. (Bericht im Auftrag des Schweizerischen Gesundheitsobservatoriums Obsan und des Bundesamts für Gesundheit BAG). Basel/ Bern/ Luzern. Der Bericht kann auf der Webseite <https://www.versorgungsatlas.ch> heruntergeladen werden.

## 1.1 Kriterien für die Methodenwahl

Zur Bewertung unterschiedlicher Methoden und Verfahren werden folgende Fragen/ Kriterien herangezogen.

- *Theoretische Fundierung.* Sind die Methoden auf einer theoretischen Fundierung abgestützt? Ist der theoretische Ansatz kompatibel mit empirischen Befunden?
- *Quantifizierung der regionalen Variabilität.* Erfolgt die Modellierung der regionalen Variabilität im Kontext der Theorie? Sind die Masse zur Quantifizierung valide und reliabel?
- *Bemessung von "too much variation".* Wie bemisst sich eine allfällige Überdispersion?
- *Effizienz der Schätzer/ Raten.* Sind die geschätzten Raten stabil und effizient geschätzt (d. h. möglichst kleine Zufallsvariation)?
- *Vertrauensintervalle für die geschätzten Raten.* Ist die vorgeschlagene Konstruktion von Vertrauensintervallen geeignet, die Variabilität korrekt wiederzugeben?

## 1.2 R-Package sava

Die Berechnungen im vorliegenden Bericht erfolgten mit der R-Statistiksoftware (Language and Environment for Statistical Computing, siehe [R Development Core Team, 2023](#)). Die im Bericht verwendeten Funktionen sind im R-Package sava implementiert und öffentlich zugänglich. Das Code Repository befindet sich auf GitHub und kann unter <https://github.com/tobiasschoch/sava> eingesehen werden. Die wichtigsten Funktionen des Packages sind in Anhang E dokumentiert.

## 1.3 Aufbau des Berichts

In Kapitel 2 werden die Grundlagen (Population, Indikatoren, HSA-Topologie und Notation) erarbeitet und die Daten mit den Methoden der deskriptiven Statistik untersucht. In Kapitel 3 stellen wir unterschiedliche Modelle und Methoden zur Quantifizierung der räumlichen Variation vor. Diese Methoden werden in Kapitel 4 mit empirisch Daten untersucht. In Kapitel 5 werden die wichtigsten Erkenntnisse zusammengefasst und Empfehlungen ausgesprochen.

## 2 Grundlagen

Der Bericht ist absichtlich formal gehalten (mit gewissen Einschränkungen für die Leserlichkeit). Dies sollte möglichen Unklarheiten bei den Definitionen und Annahmen vorgereifen.

### 2.1 Population, Indikatoren und Daten

Mit dem Relaunch im April 2023 umfasst der Versorgungsatlas über 100 **Indikatoren** zur medizinischen Versorgung in der Schweiz. Die Indikatoren fokussieren auf die somatische Versorgung und umfassen sowohl stationäre als auch ambulante Behandlungen. Sie sind nach Kanton und oftmals auch nach **Hospital Service Areas** (HSA, Spitalregionen) oder ausgewiesen.

Die Definition der Indikatoren erfolgte durch das Obsan (unter Einbezug der medizinischen Fachgesellschaften und weiteren Expertinnen und Experten). Eine detailliert Beschreibung der Indikatoren findet sich auf der Webseite <https://www.versorgungsatlas.ch>.

**Anmerkung.** Einzelne Indikatoren sind nur für Teilpopulation definiert; z. B. Mammografien (nur Frauen). In den meisten Fällen ist aus dem Kontext ersichtlich, auf welche Teilpopulation sich ein Indikator bezieht, so dass wir auf eine ausdrückliche Nennung verzichten. In anderen Fällen geben wir die Teilpopulation explizit an.

Die Operationalisierung der Indikatoren basiert in erster Linie auf zwei Datenquellen. Für den stationären Bereich ist dies die Medizinische Statistik der Krankenhäuser (MedStat) des Bundesamts für Statistik (BFS). Für den ambulanten Bereich wird vor allem der Tarifpool (TP) der SASIS AG genutzt; für einzelne Indikatoren werden zusätzliche Datenquellen hinzugezogen.

Für die empirische Überprüfung der Methoden werden **26 Test-Indikatoren** verwendet, siehe Anhang A. Zwei beispielhafte Indikatoren sind:

- MENK: Arthroskopische Meniskektomie am Knie
- DECO: Dekompression am Spinalkanal (ohne gleichzeitige Diskushernie)

Die absoluten Häufigkeiten (Fallzahlen) zu den Testindikatoren stammen aus der MedStat (Referenzjahre 2013–2015). Für die Berechnung der standardisierten Behandlungsraten werden Daten aus Statistik der Bevölkerung und der Haushalte (STATPOP) des BFS und der European Standard Population verwendet.

## 2.2 Verabredung der Notation

Alle HSA-spezifischen Grössen, wie z. B. die Populationsgrösse der  $i$ -ten HSA,  $n_i$ , oder die Behandlungsrate (Kurzform: Rate; engl. Crude Rate),  $r_i$ , werden mit Subskript oder Index  $i$  ( $i = 1, \dots, n$ ) gekennzeichnet. Die Anzahl der HSA wird mit  $n$  bezeichnet und hängt von der gewählten **HSA-Topologie** ab, die je nach Indikator unterschiedlich sein kann (z. B. unterschiedliche Topologien für Herzchirurgie und Rheumatologie). Die 26 Test-Indikatoren auf Basis der MedStat-Daten sind einer der folgenden HSA-Topologien zuwiesen (siehe Anhang A):

- HSA\_2013 (61 Regionen)
- HSA\_2013\_spine (20 Regionen)
- HSA\_2013\_herz (19 Regionen)

Die Namensgebung der HSA orientiert sich an den Namen der Kantone, in welchem sich die HSA befindet, z. B. BE05 oder ZH47.

**Anmerkung.** Die Wahl der HSA-Topologie ist nicht Gegenstand dieses Berichts. Die Topologien werden durch das Obsan festgelegt und nach der Methode von [Haynes et al. \(2020\)](#) erzeugt.

Grössen oder Kennzahlen ohne Indexierung beziehen sich auf die Schweiz. Die Indizes  $j$  und  $k$  werden verwendet, um Personen in Teilpopulation kenntlich zu machen (z. B. Altersgruppen).

Zur Vereinfachung der Notation verzichten wir darauf, die Grössen und Kennzahlen (z. B. Raten oder Häufigkeiten) mit einem *Indikator*-spezifischen Index zu versehen (z. B.  $r_{MENK}$  für die Behandlungsrate des Indikators MENK). Wir stellen sicher, dass es aus dem Kontext ersichtlich ist, auf welchen Indikator sich eine Grösse oder Kennzahl bezieht. Überdies verabreden wir die folgende Notation:

$r^{crude}$	Crude Rate in der Schweiz pro 1'000 Einwohner/Innen,
$r_i^{crude}$	Crude Rate in der $i$ -ten HSA ( $i = 1, \dots, n$ ) pro 1'000 Einwohner/Innen,
$r_{jk}^{crude}$	Crude Rate der $j$ -ten Altersgruppe ( $j = 1, \dots, J$ ) für das Geschlecht $k$ , mit $k \in \{M, F\}$ , pro 1'000 Einwohner/Innen,
$o_{ijk}$	beobachtete Anzahl der Fälle (observed) in der $i$ -ten HSA, der $j$ -ten Altersgruppe und Geschlecht $k$ ,
$o_{jk}$	beobachtete Anzahl der Fälle (observed) in der $j$ -ten Altersgruppe und Geschlecht $k$ ,
$o_i$	beobachtete Anzahl der Fälle (observed) in der $i$ -ten HSA,
$o$	beobachtete Anzahl aller Fälle, $o = \sum_{i=1}^n o_i$ ,
$e_i$	erwartete Anzahl der Fälle in der $i$ -ten HSA (siehe unten),
$e$	erwartete Anzahl aller Fälle, $e = \sum_{i=1}^n e_i$ ,
$n_{ijk}$	Anzahl Einwohner/Innen in der $i$ -ten HSA, der $j$ -ten Altersgruppe mit Geschlecht $k$ .
$n_i$	Anzahl Einwohner/Innen der $i$ -ten HSA,

**Anmerkung.**

- Die Definition der Altersgruppen und weitere Spezifikationen werden in den folgenden Kapiteln diskutiert. Wir haben die Notation bereits an dieser Stelle vollständig eingeführt, so dass die Lesenden alle Definitionen an einem Ort einsehen können.
- Wir verwenden den englischen Begriff der Crude Rate, weil es keinen entsprechend prägnanten Begriff in der deutschen Sprache gibt.

## 2.3 Schätzung der Raten

In Kern des Interesses steht die Anzahl der beobachteten Fälle  $o_i$  (für einen bestimmten Indikator) in Region  $i$  bzw. die Crude Rates

$$r_i^{crude} = \frac{o_i}{n_i} \quad (i = 1, \dots, n)$$

wobei  $o_i$  und  $n_i$  *bekannte* Kennzahlen der Grundgesamtheit sind.<sup>1</sup> Die Crude Rates  $r_i^{crude}$  müssen darum nicht geschätzt werden<sup>2</sup> und besitzen (im Kontext der Design-basierten Inferenz/ Randomisierungsinferenz) keine Stichprobenverteilung. Die Crude Rates sind jedoch *empirische* Kenngrößen. Falls es in einer Region keine Fälle gab (d. h.  $o_i = 0$ ), dann folgt  $r_i^{crude} = 0$ . Eine Crude Rate von null ist Ausdruck davon, dass es in einer bestimmten Region für eine bestimmte Zeitperiode keine Fälle gab, obschon die *wahre* (oder effektive), aber unbekannte, Rate grösser null ist.<sup>3</sup> Die Rate  $r_i^{crude} = 0$  ist ein Artefakt. Sie ist ungeeignet, um sich ein Bild zur Größenordnung der effektiven oder wahren Rate zu machen. Die Crude Rates sind nicht nur im Fall von  $o_i = 0$  problematisch, sondern weisen (im Kontext des Versorgungsatlas) generell – vor allem in kleinen Regionen – eine beträchtliche Variation auf, die sichtbar wird, wenn die Crude Rates mehrerer Jahre verglichen werden.

Der hohen zeitlichen Variation der Crude Rates und dem Artefakt für  $o_i = 0$  könnten wir dadurch begegnen, dass die Crude Rates nicht für einzelne Jahre sondern grössere Zeitfenster (z. B. 4 Jahre) berechnet werden. Damit ginge jedoch eine Reduktion der zeitlichen Auflösung einher, so dass kurzfristige Veränderungen nicht erkennbar sind. Eine Alternative dazu bildet der *Einsatz statistischer Modelle*, den wir in den nächsten Abschnitten genauer ausarbeiten wollen.<sup>4</sup>

<sup>1</sup> Für den Moment betrachten wir Zähldaten ohne Standardisierung nach Risikofaktoren wie Alter etc.

<sup>2</sup> Schätzer werden durch einen "Hut", z. B.  $\hat{r}_i$ , kenntlich gemacht.

<sup>3</sup> Es ist theoretisch denkbar, dass die wahre Rate null ist. Solche Fälle sind jedoch von akademischem Interesse und werden hier ausgeschlossen.

<sup>4</sup> Die Berechnung der Crude Rates für mehrere Jahre beruht (implizit) auch auf einem statistischen Modell. Es postuliert, dass die Anzahl Fälle in einem mehrjährigen Zeitfenster identisch verteilt sind.

## Statistische Modelle

In allen Modellen gehen wir von der folgenden Annahme aus (Cain und Diehr, 1992).

**Annahme 1.** Es bezeichne  $O_{il}$  die Anzahl der Behandlungen von Person  $l$  in HSA  $i$  in der Referenzperiode ( $i = 1, \dots, n, l = 1, \dots, n_i$ ). Wir nehmen an, dass die  $O_{il}$  unabhängig und identisch verteilte Zufallsvariable sind.

Unter dieser Annahme<sup>5</sup> ist es zugelassen, dass Mehrfachbehandlungen in der Referenzperiode auftreten können; folgerichtig nehmen die  $O_{il}$  Werte aus den natürlichen Zahlen inkl. null an. Die wichtigsten Verteilungen sind im nachfolgenden Exkurs charakterisiert.

**Bernoulliverteilung.** Sind Mehrfachbehandlungen (in der gleichen Beobachtungsperiode) ausgeschlossen, dann kann  $O_{il}$  nur Werte in  $\{0, 1\}$  annehmen und für die Verteilung gilt  $O_{il} \sim \text{Ber}(p_i)$  für  $l = 1, \dots, n_i$ . Die Summe von  $n_i$  unabhängig Bernoulli-verteilter Zufallsvariablen besitzt eine Binomialverteilung,  $O_i = \sum_{l=1}^{n_i} O_{il} \sim \text{Bin}(n_i, p_i)$ . Erwartungswert und Varianz für  $O_i$  sind gegeben durch  $E(O_i) = p_i$  und  $\text{Var}(O_i) = n_i p_i (1 - p_i)$ .

**Poissonverteilung.** Sind Mehrfachbehandlungen möglich, dann nimmt  $O_{il}$  Werte aus den natürlichen Zahlen (inkl. null) an. Die Verteilung von  $O_{il}$  kann durch die Poissonverteilung  $\text{Pois}(p_i)$  beschrieben werden (für  $l = 1, \dots, n_i$ ). Die Poissonverteilung ist reproduktiv, so dass gilt  $O_i = \sum_{j=1}^{n_i} O_{ij} \sim \text{Pois}(n_i p_i)$ . Erwartungswert und Varianz für  $O_i$  sind gegeben durch  $E(O_i) = \text{Var}(O_i) = n_i p_i$ .

Wenn die Wahrscheinlichkeit  $p_i$  sehr klein ist, dann wird die Binomialverteilung häufig durch eine Poissonverteilung approximiert. Dies erkennt man auch aus der Varianz der Binomialverteilung:  $\text{Var}(O_i) \approx n_i p_i$ , weil  $(1 - p_i) \approx 1$  ist. Obwohl die Poissonverteilung die Verteilung einer Zählvariablen beschreibt, ist sie zur Modellierung von Mehrfachbehandlungen ungeeignet, weil sie die effektive Varianz unterschätzt (Cain und Diehr, 1992). Die Autoren führen den sogenannten Multiple Admission Factor (MFA) ein, um die Varianz infolge von Mehrfachbehandlungen schätzen zu können. Unsere empirischen Untersuchungen zeigen, dass der MFA einen relativ kleinen Einfluss hat.

Wir wollen an dieser Stelle noch einmal auf die Crude Rates zurückkommen. Jetzt wird unterstellt, dass die Anzahl der Fälle in den Regionen  $i = 1, \dots, n$  als unabhängige Zufallsvariablen mit Poissonverteilung modelliert werden, formal

$$O_i \mid \theta_i \stackrel{\text{ind}}{\sim} \text{Pois}(\theta_i n_i) \quad (2.1)$$

<sup>5</sup> Die Wahrscheinlichkeit, dass eine Person  $l$  einer Behandlung unterzogen wird, hängt nicht davon ab, ob eine andere Person in HSA  $i$  dieser Behandlung unterzogen wurde. Diese Annahme ist für die meisten im Versorgungsatlas abgebildeten Behandlungen sinnvoll. Diese Annahme wäre verletzt, wenn es sich z. B. um Behandlungen infolge von ansteckenden Krankheiten handeln würde.

Die Annahme besagt auch, dass alle Personen in HSA  $i$  die gleiche Wahrscheinlichkeit  $p_i$  besitzen, der Behandlung unterzogen zu werden.

wobei  $\theta_i$  die unbekannte, wahre Rate in Region  $i$  bezeichnet. Unter dem Modell können die Crude Rates nach der Maximum-Likelihood-Methode (ML) geschätzt werden. Der Schätzer und seine geschätzte Varianz sind gegeben durch

$$\widehat{r}_i^{crude} = \frac{O_i}{n_i} \quad \text{und} \quad \widehat{\text{Var}}(\widehat{r}_i^{crude}) = \frac{O_i}{n_i}. \quad (2.2)$$

Für  $O_i = 0$  erzeugt der Schätzer  $\widehat{r}_i^{crude}$  das Artefakt einer geschätzten Rate von null. Die Varianz ist umgekehrt proportional zur Anzahl Personen  $n_i$ . Für kleine Regionen ist die Varianz des Schätzers tendenziell grösser als in grossen Regionen.

Die Crude Rates sind einfach zu berechnen. Sie berücksichtigen jedoch keine regionsspezifischen Risikofaktoren, so dass Vergleiche zwischen Regionen irreführend sein können, falls sich die Regionen in Bezug auf Risikofaktoren (z. B. Alter) systematisch unterscheiden. Durch eine vorgängige Standardisierung können Risikofaktoren (z. B. alters- und geschlechtsspezifische Effekte) eliminiert werden. Dies soll im nachfolgenden Kapitel genauer besprochen werden.

## 2.4 Standardisierung der Raten

Die Analyse oder graphische Darstellung nicht standardisierter regionaler Raten kann irreführend sein (Cressie, 1992; Waller und Carlin, 2010). Dies ist insbesondere dann der Fall, wenn sich die Regionen hinsichtlich ihrer Grösse oder des datengenerierenden Prozesses unterscheiden; siehe bspw. Fleiss et al. (2003, Kap. 19) oder Cressie (1992). Im Kontext von Gesundheitsdaten dient die Standardisierung der Korrektur der Raten für bekannte Risikofaktoren (siehe bspw. Waller und Gotway, 2004, Kap. 2.3.4). Breslow und Day (1987, S. 61) haben das Ziel der Standardisierung treffend umschrieben:

*A major goal of standardization, besides combining a set of age-/stratum-specific rates into a synoptic figure, is to provide a quantitative measure of the difference in rates between the study cohort and a standard population or other comparison group that is free from the effects of age or other confounding variables.*

Der Versorgungsatlas führt eine Standardisierung nach Alter und Geschlecht durch. Dies ist einerseits der Verfügbarkeit von Daten zu Risikofaktoren geschuldet und andererseits dadurch begründet, dass der Atlas eine Vielzahl von Indikatoren abbildet, was eine "Kompromiss"-Auswahl an Risikofaktoren nötig machte. Für die vertiefte Analyse eines Einzelindikators, kann es unter Umständen angezeigt sein, weitere Risikofaktoren einzubeziehen.

**Tabelle 2.1** Tafel zur direkten und indirekten Standardisierung (nach Alter; hypothetische Population)

	Studienpopulation	Referenzpopulation (standard)
1. Altersgruppen	$1, \dots, J$	$1, \dots, J$
2. Anzahl Individuen (nach Altersgruppe)	$n_1^S, \dots, n_J^S$	$n_1^R, \dots, n_J^R$
3. Altersspezifische Rate	$r_1^S, \dots, r_J^S$	$r_1^R, \dots, r_J^R$
4. Total Fälle	$\sum r_j^S n_j^S = \text{Total}^S$	$\sum r_j^R n_j^R = \text{Total}^R$
5. Rate (Crude Rate)	$r^S = \frac{\text{Total}^S}{\sum n_j^S}$	$r^R = \frac{\text{Total}^R}{\sum n_j^R}$
A. Indirekte Standardisierung (ISR):	$r^{ind} = SMR \cdot r^R$	mit $SMR = \frac{\text{Total}^S}{\sum r_j^R n_j^S}$
B. Direkte Standardisierung (DSR):	$r^{dir} = \frac{\sum r_j^S n_j^R}{\sum n_j^R}$	
a. Schätzung ISR	$\widehat{r}^{ind} = \widehat{SMR} \cdot r^R$	mit $\widehat{SMR} = \frac{\text{Total}^S}{\sum \widehat{r}_j^R n_j^S}$
b. Schätzung DSR	$\widehat{r}^{dir} = \frac{\sum \widehat{r}_j^S n_j^R}{\sum n_j^R}$	

Anmerkung: Das Symbol  $\sum$  bezeichnet die Summe über den Laufindex  $j = 1, \dots, J$ . Geschätzte Größen sind mit einem "Hut" versehen.

## Methoden der Standardisierung

Wir wollen die Notation in diesem Abschnitt einfach halten und verzichten darum auf eine *kombinierte* Standardisierung nach Alter und Geschlecht. Es soll nur nach Alter standardisiert werden.

In der Literatur wird zwischen *indirekter* und *direkter* Standardisierung unterschieden (Waller und Gotway, 2004, Kap. 2); siehe Tabelle 2.1. In beiden Fällen wird eine Studienpopulation und eine Referenzpopulation (Standardpopulation) definiert. Die *indirekte* Standardisierung (siehe Zeile A in Tabelle 2.1) setzt Kenntnis der altersspezifischen Raten in der Referenzpopulation,  $r_j^R$ , voraus (die Altersgruppen sind mit  $j = 1, \dots, J$  indiziert; das Superskript R steht für Referenzpopulation; die Studienpopulation wird mit Superskript S gekennzeichnet). Diese Raten sind unbekannt und müssen für die Referenzpopulation geschätzt werden (die Schätzung ist mit  $\widehat{r}_j^R$  bezeichnet, siehe Zeile a). Diese Art der Standardisierung wird auch als *internal indirect* Standardisierung (IIS) bezeichnet, weil die Studienpopulation eine Teilmenge der Referenzpopulation ist. Für die *direkte* Standardisierung (siehe Zeile B) werden die altersspezifischen Rate in der *Studienpopulation*,  $r_j^S$ , benötigt. Auch sie müssen geschätzt werden (siehe Zeile b).

## Vergleichbarkeit von Raten

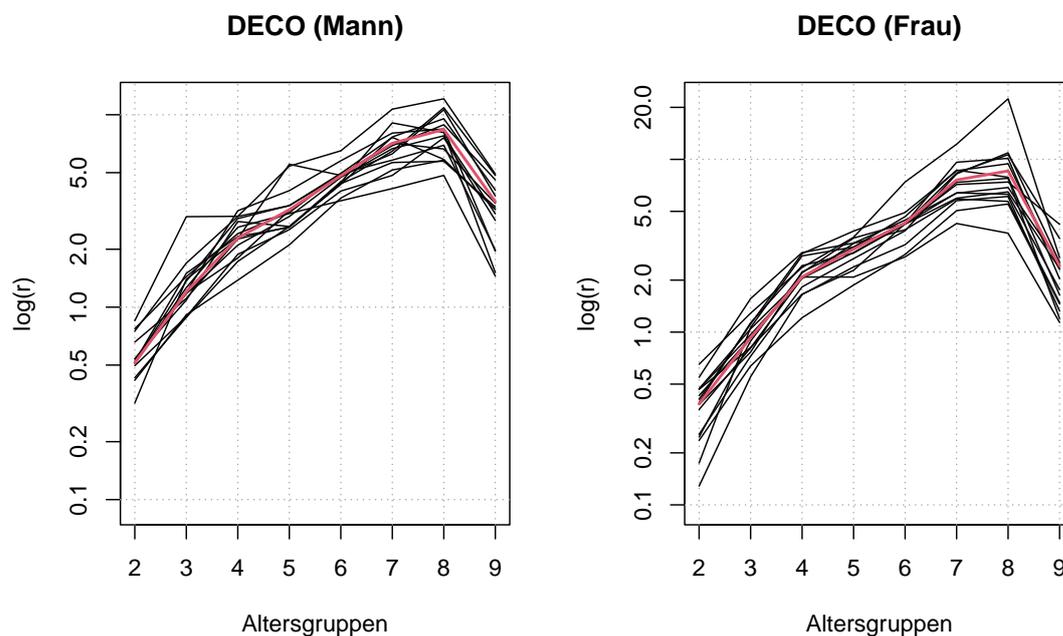
Die nach der direkten bzw. indirekten Standardisierung berechneten Raten sind, je nach Anwendungsfall und Fragestellung, besser oder weniger gut geeignet. Wir wollen auf die wichtigsten Unterschiede der beiden Standardisierungsmethoden kurz eingehen.

Die direkte Standardisierung ist vorteilhaft – so eine weitverbreitete Meinung, vgl. bspw. [Miittinen \(1972\)](#) –, um Vergleiche zwischen Raten anzustellen. Einschränkend muss jedoch festgehalten werden, dass Vergleiche nur dann sinnvoll sind, wenn die zu vergleichenden Raten für die gleiche Referenzpopulation (z. B. European Standard Population) berechnet wurden und die Studienpopulation(en) eine grosse Ähnlichkeit mit der Referenzpopulation aufweist bzw. aufweisen. Direktstandardisierte Raten sind also nicht per se vergleichbar.<sup>6</sup> Bei der indirekten Standardisierung ist Problematik der Vergleichbarkeit noch viel ausgeprägter. Durch die unterschiedlichen Strukturen der Population können Vergleiche verzerrt sein ( $\Rightarrow$  Bias). Gemäss [Breslow und Day \(1987, S. 73\)](#) war die Bias-Problematik der indirekten Standardisierung ausschlaggebend, warum viele Autorinnen und Autoren, z. B. [Miittinen \(1972\)](#), indirekte-standardisierte Raten für nicht vergleichbar hielten bzw. halten. Um den Sachverhalt besser zu verstehen, ist es hilfreich Bedingungen herauszuarbeiten, unter welchen die direkte und indirekte Standardisierung zu den gleichen Resultaten führen.

[Pickle und White \(1995, S. 617–618\)](#) haben drei Bedingungen formuliert, unter welchen die Raten nach der direkten und indirekten Standardisierung identisch sind (oder zumindest kaum unterscheidbar sind). Falls eine der Bedingungen erfüllt ist, dann erübrigt sich die Bias-Problematik bei der indirekten Standardisierung, weil  $\widehat{r}_i^{ind} = \widehat{r}_i^{dir}$  gilt. In diesem Fall ist die indirekt-standardisierte Rate der direkt-standardisierten Rate bezüglich Bias nicht unterlegen.

Die ersten zwei Bedingungen von [Pickle und White \(1995\)](#) sind eher von akademischem Interesse, weil sie idealisierte Population betreffen. Die dritte Bedingung fordert ein, dass die Raten (nach Alter und Geschlecht) in der Studienpopulation *proportional* zu denjenigen der Referenzpopulation sind. Formal muss also  $r_{ij}^S = \alpha_i r_{ij}^R$  gelten, wobei  $\alpha_i$  ein regionenspezifischer Proportionalitätsfaktor ist; der Index  $j$  indexiert die Altersgruppen (und weitere Risikofaktoren). Ebendiese Proportionalitätsbedingung kann mit empirischen Daten getestet werden. [Gail \(1978\)](#) hat hierzu einen Diagnostikplot vorgeschlagen, in welchem  $\log(r_{ij})$  als Liniengrafik (für jede einzelne Region  $i = 1, \dots, n$ ) gegen die Altersgruppe  $j$  aufgetragen wird. Falls alle Linien durch eine vertikale Parallelverschiebung ineinander überführt werden können, dann ist die Bedingung erfüllt. Der Diagnostikplot für den Indikator DECO (siehe Abb. 2.4) zeigt 20 Linien (für jedes Geschlecht) weil die HSA-Topologie HSA\_spine\_2013 20 Regionen umfasst (zusätzlich ist der Verlauf der Rate für die Schweiz rot eingezeichnet). Wir erkennen aus der Abbildung, dass die regionenspezifischen

<sup>6</sup> [Breslow und Day \(1987, S. 55–56\)](#) diskutieren beispielhaft welchen Effekt die Wahl der Referenzpopulation für die Raten zur Magenkrebsinzidenz in Island und Japan hat. Die Autoren folgern, dass “[...] the relative positions of the two countries as expressed in the age-standardized rate will depend to a large extent on the choice of the standard. If the standard population is heavily weighted towards the elderly, Iceland will have a relatively higher age standardized rate, while the reverse will be true if the standard population is younger.”



**Abbildung 2.1** Diagnostikplot nach Gail (1978) für den Indikator Dekompression am Spinalkanal (ohne gleichzeitige Diskushernie (DECO), nach Geschlecht. Jede Linie repräsentiert das Altersgruppen-Profil der logarithmierten Raten für eine HSA; die rote Linie zeigt den Verlauf für die Schweiz. (HSA-Topologie: HSA\_spine\_2013: 20 Regionen, Altersgruppen in 10-Jahresschritten; Daten: MedStat 2013–2015, aggregiert).

Linien (mehrheitlich) parallel zu einander verlaufen. Für den Indikator DECO ist die Proportionalitätsbedingung (mehrheitlich) erfüllt.

**Anmerkung.** Bei der überwiegenden Mehrheit der Indikatoren ist die Proportionalitätsbedingung erfüllt. Alle Diagnostikplots nach Gail (1978) sind in Anhang C aufgeführt.

## Effizienz und Verzerrung

In der Praxis sind die (wahren) Raten  $r_i^{dir}$  und  $r_i^{ind}$  unbekannt und müssen durch  $\hat{r}_i^{dir}$  und  $\hat{r}_i^{ind}$  geschätzt werden. Um die Schätzer zu charakterisieren, schlagen Breslow und Day (1987, S. 73) vor, die Verzerrung (Bias) und die Varianz des Schätzers (Effizienz) zu berücksichtigen. In der älteren Literatur – vgl. z. B. Miettinen (1972) – wurden Effizienzüberlegungen kaum beachtet. Wir definieren den mittleren quadratischen Fehler (mean square error, MSE) eines Schätzers  $\hat{r}_i$ , der in zwei Komponenten

$$\text{MSE}(\hat{r}_i) = \text{Variance}(\hat{r}_i) + \text{Bias}^2(\hat{r}_i)$$

zerlegt werden kann. Wir haben gesehen, dass die Proportionalitätsbedingung von Pickle und White (1995) mehrheitlich erfüllt ist. Deshalb ist die Bias-Problematik für die nach der indirekten Standardisierung berechneten Raten gebannt. Wie sieht es mit der Varianz aus? Hier ist die

nach der *direkten* Standardisierung geschätzte Rate problematisch, weil sie aufgrund der geringen Fallzahlen (höchst) instabil ist; siehe [Breslow und Day \(1987, S. 66\)](#) oder [Waller und Gotway \(2004, S. 14\)](#).<sup>7</sup> Daraus resultiert eine grosse Varianz und damit eine grosse Schätzungenauigkeit. Schwerer wiegt, dass es keine Möglichkeit gibt, diese Problematik zu beheben. Die Schätzer nach der indirekten Standardisierung haben hingegen eine wesentlich kleinere Varianz und sind (sofern die Proportionalitätsbedingung erfüllt ist) der direkten Standardisierung klar vorzuziehen. Wir illustrieren die Effizienzüberlegungen beispielhaft im untenstehenden Exkurs.

**Exkurs.** Die Darstellung in Tabelle 2.1 wurde bewusst vereinfacht. Wir wollen nun die Schätzung nach der direkten und indirekten Methode auf die Daten der MedStat (standardisiert nach Alter und Geschlecht) anwenden und verlassen darum die vereinfachte Darstellung. Wir erhalten die folgenden Schätzer der Raten

$$\hat{r}_i^{ind} = \frac{o_i}{\sum_{j=1}^J \sum_{k \in \{M,F\}} \hat{r}_{jk} n_{ijk}} \cdot \frac{o}{n} \quad \text{und} \quad \hat{r}_i^{dir} = \frac{\sum_{j=1}^J \sum_{k \in \{M,F\}} \hat{r}_{ijk} n_{ijk}^{ESP}}{\sum_{j=1}^J \sum_{k \in \{M,F\}} n_{ijk}^{ESP}},$$

wobei  $o$  und  $n$ , respektive, die Anzahl Fälle und Anzahl Personen in der MedStat bezeichnen; ESP steht für European Standard Population. Für die folgende Besprechung konzentrieren wir uns auf die geschätzten Raten (hervorgehoben in roter Schriftfarbe). Der indirekt-standardisierte Schätzer beruht auf dem Schätzer  $\hat{r}_{jk} = o_{jk}/n_{jk}$ , wobei die Indizes  $j$  und  $k$ , respektive, für die Altersgruppen und das Geschlecht stehen. Der direkt-standardisierte Schätzer basiert auf dem Schätzer  $\hat{r}_{ijk} = o_{ijk}/n_{ijk}$  (Index  $i$  steht für die Region). Wir wollen nun die weitere Diskussion konkret gestalten und wählen darum einen Indikator mit der HSA\_2013-Topologie. Diese Topologie zählt 61 Regionen. In der MedStat werden 21 Altersklassen und 2 Geschlechter unterschieden. Für die direkte Standardisierung müssen also folglich  $61 \times 21 \times 2 = 2'562$  einzelne Raten  $\hat{r}_{ijk}$  geschätzt werden. Für die indirekte Standardisierung sind es hingegen nur 42 verschiedene Raten (21 Altersklassen  $\times$  2 Geschlechter). Die Anzahl der Fälle aus der MedStat für ein bestimmtes Jahr ist eine feste Grösse. Es macht nun einen grossen Unterschied, ob 42 oder 2'562 Raten berechnet werden. Es ist offensichtlich, dass die nach der direkten Methode standardisierte Rate nur mit deutlich geringerer Präzision geschätzt werden kann.

## Implementation

**Anmerkung.** Alle Raten beziehen sich auf die Basis von 1'000 Einwohner/Innen der ständigen Wohnbevölkerung der Schweiz.

Zur Vereinfachung der Notation verzichten wir in diesem Abschnitt darauf, die geschätzten Raten durch einen Hut zu kennzeichnen, z. B.  $\hat{r}_{ij}^{crude}$ , und schreiben stattdessen  $r_{ij}$ .

<sup>7</sup> "For rare diseases, the incidence proportions observed,  $r_j = y_j/n_j$  [in unserer Notation  $o_j$  anstelle von  $y_j$ ], may be statistically unstable, particularly if  $n_j$  is small, in that the addition or deletion of a single case within a particular age group could drastically change the value of  $r_j$ ." ([Waller und Gotway, 2004, S. 14](#)).

Die indirekt-standardisierten Raten können mithilfe des **Standardized Mortality Ratios (SMR)** ausgedrückt werden. Das SMR in HSA  $i$  ist gegeben durch

$$\text{SMR}_i = \frac{o_i}{e_i},$$

wobei  $o_i$  und  $e_i$ , respektive, die **beobachteten** und **erwarteten** Anzahl Fälle in HSA  $i$  bezeichnen,

$$e_i = \sum_{j=1, \dots, J} \sum_{k \in \{M, F\}} n_{ijk} r_{jk}, \quad (2.3)$$

wobei

- $r_{jk}$  die mit den Daten aus der MedStat geschätzte Crude Rate für die  $j$ -te Altersgruppe ( $j = 1, \dots, J$ ) und das Geschlecht  $k \in \{M, F\}$  bezeichnet (Altersgruppen: Alter < 1 Jahr, 1 – 4 Jahre, 5 – 9 Jahre, . . . , 85 – 89 Jahre, 90 – 94 Jahre und  $\geq 95$  Jahre);
- $n_{ijk}$  die Anzahl der Personen in der MedStat in der  $i$ -ten HSA, der  $j$ -ten Altersgruppe mit Geschlecht  $k$  bezeichnet.

**Annahme 2.** Die erwartete Anzahl der Fälle,  $e_i$ , ist nicht-stochastisch, d. h.  $e_i$  ist eine feste Grösse und keine Zufallsvariable (für alle  $i = 1, \dots, n$ ).

Diese Annahme wird bei der statistischen Modellierung eine zentrale Rolle spielen. Wir haben sie bereits hier eingeführt, um sie im Kontext der Standardisierung besprechen zu können. Aus (2.3) ist ersichtlich, dass  $e_i$  eine Funktion der alters- und geschlechtsspezifischen Raten ist. Die  $e_i$  sind formal betrachtet auch Schätzgrössen. Der Einfluss der Schätzung schwindet jedoch mit der Anzahl Regionen. Bei einer grossen Anzahl an Regionen kann der Einfluss vernachlässigt werden; entsprechend werden die  $e_i$  als feste Grössen aufgefasst; siehe [Waller und Carlin \(2010, S. 219\)](#) oder auch [Rao \(2003, S. 206\)](#).

Die nach der **indirekten** Standardisierung berechnete Rate in der  $i$ -ten HSA ist nun definiert durch

$$r_i^{\text{ind}} = \text{SMR}_i \cdot r_i \quad (i = 1, \dots, n). \quad (2.4)$$

### Direkte Standardisierung

Die nach der **direkten** Standardisierung berechnete Rate in der  $i$ -ten HSA ist definiert durch

$$r_i^{\text{dir}} = \frac{\sum_{j=1}^J \sum_{k \in \{M, F\}} r_{ijk} n_{ijk}^{\text{ESP}}}{\sum_{j=1}^J \sum_{k \in \{M, F\}} n_{ijk}^{\text{ESP}}},$$

wobei  $r_{ijk}$  die mit den Daten aus der MedStat geschätzte Crude Rate für die HSA  $i$ , die  $j$ -te Altersgruppe ( $j = 1, \dots, J$ ) und das Geschlecht  $k \in \{M, F\}$  bezeichnet;  $n_{ijk}^{\text{ESP}}$  ist die Anzahl der Personen in der European Standard Population in der  $i$ -ten HSA, der  $j$ -ten Altersgruppe mit Geschlecht  $k$ .

In Anhang [B.1](#) wird die Berechnung von Vertrauensintervallen für direkt-standardisierte Raten besprochen.

## 2.5 Deskriptive Datenanalyse

In diesem Kapitel geben wir einen Überblick über die wichtigsten Charakteristika der Daten, die bereits in Kapitel 2.1 eingeführt wurden. Im Mittelpunkt des Interesses steht die regionale Verteilung der Test-Indikatoren nach HSA, wobei die absoluten Häufigkeiten (Fallzahlen) und Crude Rates einzeln ausgewiesen werden.

Es liegen 26 Testindikatoren basierend auf den Daten der Medizinischen Statistik der Krankenhäuser vor. Sie beziehen sich auf die HSA-Topologien HSA\_2013, HSA\_2013\_herz und HSA\_2013\_spine. Die Verteilung der Referenzbevölkerung ( $n_i$ ) in den HSA ist in Tab. 2.2 charakterisiert. Die kleinste HSA (Topologie HSA\_2013) umfasst 26'588 Personen der Referenzbevölkerung. Die grösste HSA umfasst 1'224'734 Personen (HSA\_2013\_herz). Die Grösse der Referenzbevölkerung darf nicht darüber hinwegtäuschen, dass die beobachteten Fallzahlen,  $o_i$ , bei einigen Indikatoren sehr gering sind.

**Tabelle 2.2** Grösse der HSA

HSA Topologie	Min.	1. Quartil	Median	3. Quartil	Max.
HSA_2013	26'588	52'588	74'551	129'077	474'477
HSA_2013_herz	43'608	166'440	299'569	474'616	1'224'734
HSA_2013_spine	25'789	139'662	214'070	449'124	1'177'070

Daten: MedStat 2014

In den nächsten Abschnitten stehen die Fallzahlen und die Raten im Vordergrund.

In Anhang D untersuchen wir mittels Quantil-Quantil-Plot (QQ-Plot), wie gut die empirische Verteilung von  $(o_i - e_i)/e_i$  durch die Normalverteilung beschrieben werden kann. Die erwartete Anzahl der Fällen,  $e_i$ , wird nach der indirekten Standardisierung berechnet.

### HSA-Topologie: HSA\_2013

Abb. 2.2 zeigt die absolute Häufigkeit der Fälle ( $o_i$ ) je Indikator und HSA  $i = 1, \dots, n$  (Panel a) und die Crude Rates pro 1'000 Personen (Panel b). Die Häufigkeitsverteilungen zeigen eine grosse Streubreite; sie reicht von einer Handvoll von Fällen (z. B. 9 Fälle in einer HSA bei EGDB) bis zu knapp 2'000 Fällen pro HSA (z. B. 1'883 Fälle bei MENK); siehe auch Tab. 2.3. Die Verteilungen der Crude Rates ist fast symmetrisch (mit Ausnahme von MENK).

### HSA-Topologie: HSA\_herz\_2013

Für die HSA-Topologie HSA\_herz\_2013 (MedStat 2014) zeigt die Verteilung der Crudes Rates nach HSA keine besonderen Auffälligkeiten; siehe Abb. 2.3. Die Verteilungen sind leicht rechtsschief. Die

absolute Häufigkeit der Fälle bewegt sich bei den vier Indikatoren BYPA, HKAT, PTCA und STEN zwischen 23 und 6'656; siehe Tab. 2.4.

### HSA-Topologie: HSA\_spine\_2013

Für die HSA-Topologie HSA\_spine\_2013 (Medstat 2014) sind in der Verteilung der Crude Rates nach HSA für die Indikatoren DECO, FUSI, und VERT einige Ausreisser erkennbar. Die Ausreisser-HSA sind mit Namen versehen; siehe Abb. 2.4. Bei den Indikatoren KYPH und TDRE gibt es HSA, deren absolute Häufigkeit bei den Fallzahlen sehr gering oder sogar null ist; siehe Tab. 2.5.

**Tabelle 2.3** Min. und max. Häufigkeit der Fälle ( $o_i$ ) je HSA und Indikator und nationale Crude Rates

Indikator	APPE	CHOL	EGDB	EGDO	EHLB	HALL	HTEP	INGH	KTEP	MENK
Min.	10	28	9	41	11	23	64	29	66	49
Max.	593	1084	1039	747	551	455	1646	1269	1421	1883
Rate	1.21	2.05	1.92	1.89	1.14	0.95	2.77	2.37	2.41	3.11

(Forts.)

Indikator	RAFX	ROTM	SEKT	TUHS	TURP	VSTR
Min.	23	19	74	9	13	11
Max.	588	807	1952	455	775	783
Rate	1.18	1.48	4.10	0.68	1.44	1.07

Topologie: HSA\_2013; Daten: MedStat 2014

**Tabelle 2.4** Min. und max. Häufigkeit der Fälle ( $o_i$ ) je HSA und Indikator und nationale Crude Rates

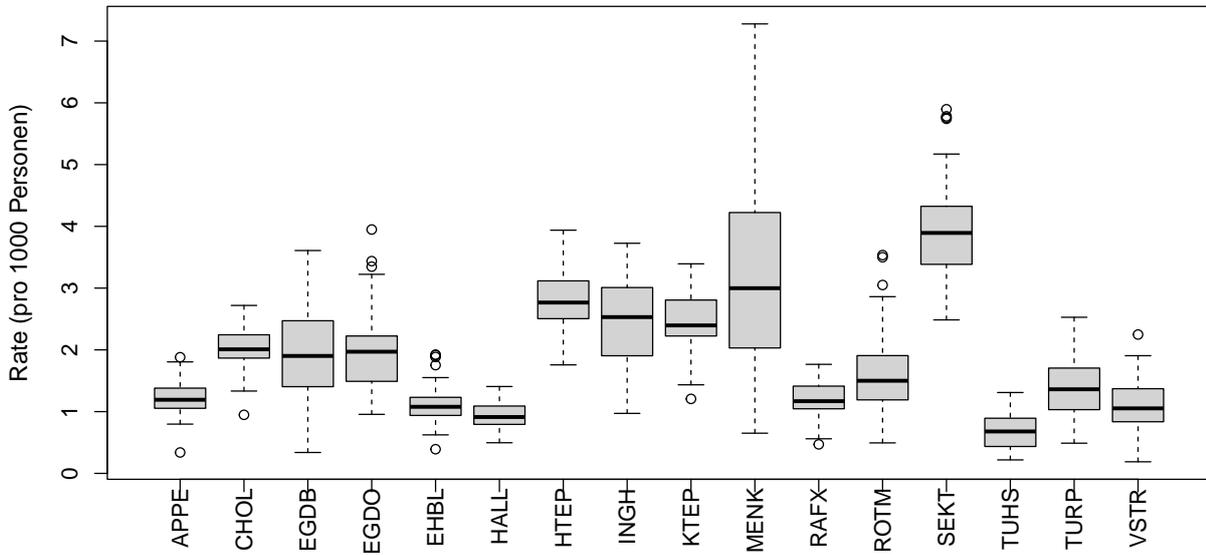
Indikator	BYPA	HKAT	PTCA	STEN
Min.	23	154	112	114
Max.	809	6656	3643	3665
Rate	0.56	3.87	2.49	2.50

Topologie: HSA\_2013\_herz; Daten: MedStat 2014

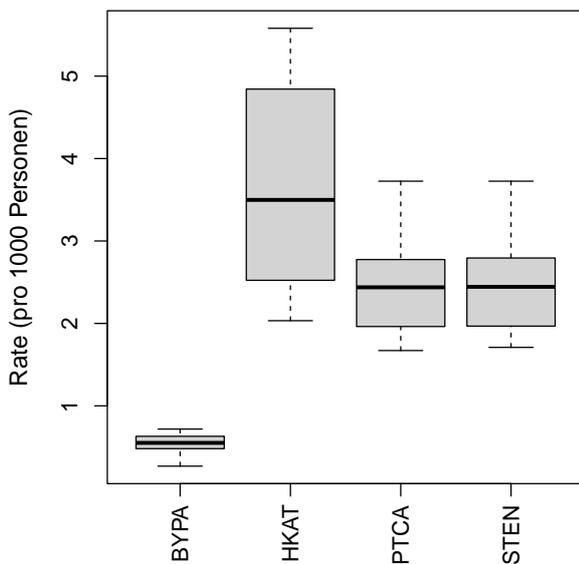
**Tabelle 2.5** Min. und max. Häufigkeit der Fälle ( $o_i$ ) je HSA und Indikator und nationale Crude Rates

Indikator	DECO	FUSI	KYPH	TDRE	VERT
Min.	120	44	1	0	4
Max.	4876	1717	210	212	692
Rate	3.38	1.27	0.16	0.09	0.28

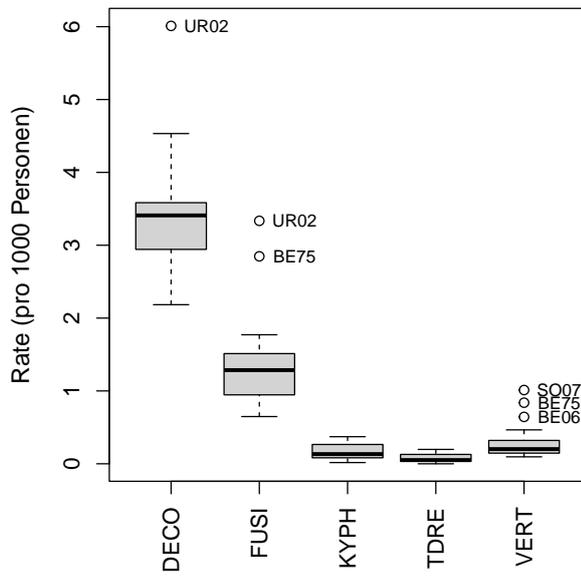
Topologie: HSA\_2013\_spine; Daten: MedStat 2014



**Abbildung 2.2** Regionale Verteilung der Crude Rates je Indikator (Daten: MedStat 2014; Topologie: HSA\_2013, 61 HSA)



**Abbildung 2.3** Regionale Verteilung der Crude Rates je Indikator (Daten: MedStat 2014; Topologie: HSA\_2013\_herz, 19 HSA)



**Abbildung 2.4** Regionale Verteilung der Crude Rates je Indikator (Daten: MedStat 2014; Topologie: HSA\_2013\_spine, 20 HSA)

### 3 Modellierung

In Kapitel 2.3 haben wir gesehen, dass der ML-Schätzer der Crude Rates und dessen Varianzschätzer unter dem Poissonmodell (in HSA  $i$ ) durch

$$\widehat{r}_i^{crude} = \frac{o_i}{n_i} \quad \text{und} \quad \widehat{\text{Var}}(\widehat{r}_i^{crude}) = \frac{o_i}{n_i} \quad (2.2)$$

gegeben sind. Für den Fall  $o_i = 0$  erzeugt der Schätzer  $\widehat{r}_i^{crude}$  das Artefakt einer geschätzten Rate von null. In Regionen mit einer kleinen Anzahl an Fällen  $n_i$  ist die geschätzte Varianz des Schätzers tendenziell gross (et vice versa).

Wir wollen nun weitere Methoden und Modelle diskutieren, die nicht zu den genannten Artefakten und Problemen führen. Hierzu greifen wir auf Erkenntnisse aus den Disziplinen Small Area Estimation (SAE, siehe bspw. Rao und Molina, 2015) und Spatial Statistics (siehe bspw. Diggle, 2010) zurück. Insbesondere die Methoden aus dem Gebiet Disease Mapping (siehe bspw. Waller und Carlin, 2010) – einem Überlappungsbereich von SAE und Spatial Statistics – sind gut geeignet für den Einsatz in der Small Area Variation Analysis (SAVA).

**Anmerkung.** Die beiden Masse zur Quantifizierung der regionalen Variation:

- Extremal Quotient (Verhältnis der grössten Crude Rate zur kleinsten Crude Rate),
- Variationskoeffizient (gewichtet nach Populationsgrösse)

werden in der neueren Literatur selten verwendet. Dies liegt auch daran, weil sie einigermaßen schwerwiegende theoretische Nachteile aufweisen; siehe bspw. Diehr et al. (1990) und Ibáñez et al. (2009). Aus diesen Gründen gehen wir auf diese Masse nicht näher ein.

Wir haben darauf verzichtet, Methoden zu berücksichtigen, welche Korrelationsstrukturen und Spatial Patterns (z. B. Besag et al., 1991) explizit modellieren.

Der Rest dieses Kapitels ist in drei Teile gegliedert:

- Systematic Component of Variation (siehe Kapitel 3.1)
- Empirical-Bayes Schätzer für das Poisson–Gamma-Modell (siehe Kapitel 3.2)
- Empirical-Bayes Schätzer für das Normal–Normal-Modell (siehe Kapitel 3.3)

#### 3.1 Systematic component of variation

Die *systematic component of variation* (SCV) von McPherson et al. (1982) spielt eine zentrale Rolle in der SAVA-Literatur zur Quantifizierung der regionalen Variabilität. Allerdings hatten es die

Autoren unterlassen, eine hinreichend genaue Definition des Masses anzugeben und Bezüge zur statistischen Fachliteratur herzustellen. Die Bedingungen, unter welchen das Mass wohldefiniert ist, blieben lange Zeit weitgehend unbekannt. Es ist wohl auch diesem Umstand geschuldet, dass die SCV – trotz grosser Beliebtheit in Praxis – seit Jahren in der Kritik steht; siehe bspw. [Diehr et al. \(1990\)](#); [Shwartz et al. \(1994\)](#).

Die theoretischen Grundlagen, auf welchen die SCV fusst, wurden von [Shwartz et al. \(1994\)](#) aufgearbeitet. Letztere stellen auch einen Bezug zur statistischen Fachliteratur her. Wir erachten es als zentral, die wichtigsten Pfeiler, auf welchen die SCV beruht, zu besprechen. Nur so können die Vor- und Nachteile der SCV diskutiert werden. Wir greifen an dieser Stelle auf die bereits eingeführte Notation zurück. Es bezeichnen  $o_i$  und  $e_i$  die beobachtete und die erwartete Häufigkeiten. Wir schreiben  $O_i$  und meinen damit die entsprechende Zufallsvariablen. Die SCV kann aus dem folgenden *Conditionally Independent Hierarchical Bayes* (CIHB) Modell abgeleitet werden (angelehnt an [Shwartz et al., 1994](#), S. 200)

$$O_i \mid \Lambda_i = \lambda_i \stackrel{ind}{\sim} \text{Pois}(\lambda_i e_i), \quad i = 1, \dots, n, \quad (3.1)$$

$$\Lambda_i \stackrel{iid}{\sim} (1, \sigma^2), \quad i = 1, \dots, n. \quad (3.2)$$

Die Symbole  $\stackrel{ind}{\sim}$  und  $\stackrel{iid}{\sim}$  bedeuten, respektive, “ist unabhängig verteilt” und “ist unabhängig und identisch verteilt”;  $\text{Pois}(\lambda)$  ist die Verteilungsfunktion einer Poisson-verteilten Zufallsvariable mit Rate  $\lambda > 0$  und  $(\mu, \sigma^2)$  steht für eine beliebige Verteilung, die nur durch den Erwartungswert  $\mu \in \mathbb{R}$  und die Varianz  $\sigma^2 \geq 0$  bestimmt ist. Im Kontext der Bayes’schen Inferenzstatistik zum CIHB-Modell werden die Parameter  $\mu$  und  $\sigma^2$  als bekannte Grössen betrachtet. In der frequentistischen Inferenzstatistik sind die Parameter unbekannt und müssen geschätzt werden ( $\Rightarrow$  Empirical-Bayes Ansatz).

### Diskussion der Verteilungsannahmen

Im CIHB-Kontext werden die Formulierungen (3.1) und (3.2) als, respektive, das *sampling model* und die A-priori-Wahrscheinlichkeitsverteilung (kurz A-priori-Verteilung) der Zufallsvariablen  $\Lambda_i$  bezeichnet. Die A-priori-Verteilung modelliert die Poisson-Überdispersion (*overdispersion* oder *extra-Poisson variation*). Man spricht von Überdispersion, wenn die empirische Varianz in den Daten grösser ist (z. B. aufgrund von regionalen Effekten), als die vom Poissonmodell postulierte Varianz.

In der Spezifikation von [McPherson et al. \(1982\)](#) ist die Verteilung in (3.2) eine uneigentliche (improper) A-priori-Verteilung, weil sie nur durch die ersten beiden zentralen Momente definiert ist. Der Erwartungswert ist 1.0 und postuliert, dass die Behandlungsrate im Mittel über alle  $n$  Regionen den Wert 1 besitzt. Falls  $\sigma^2 > 0$  ist, dann besitzt die Verteilung eine Überdispersion (i. Vgl. z. Poissonmodell). Wenn hingegen  $\sigma^2$  bzw. dessen Schätzung den Wert null hat, dann gibt es keine systematische Variation (im Sinne der SCV). [McPherson et al. \(1982\)](#) begründen die Wahl des

Poissonmodells in (3.1) dadurch, dass “surgery [...] a relatively rare event” sei. Diese Annahme erscheint uns in den meisten Fällen tauglich. Eine analoge Begründung für die Poissonverteilung wird auch im Gebiet Disease Mapping für seltene, nicht-ansteckende Krankheiten häufig angeführt; siehe bspw. Mollié (1996, S. 360). Diehr und Grembowski (1990) und Diehr et al. (1992) geben allerdings auch Beispiele aus der SAVA an, bei welchen die Poissonverteilungsannahme die empirische Verteilung unzureichend beschreibt.

Im CIHB-Kontext beruht die Bayes'sche statistische Inferenz auf Grundlage der A-posteriori-Verteilung. Die frequentistische Inferenzstatistik zum Modell beruht auf Schätzungen nach dem Empirical Bayes (EB) Ansatz. Dabei ist der Parameter  $\sigma^2$  in (3.2) unbekannt und soll an der Randverteilung der  $O_i$  geschätzt werden. Wir gehen auf den EB Ansatz später noch genauer ein; siehe Kapitel 3.2.

### 3.1.1 Herleitung der SCV

Ausgehend vom Modell in (3.1 – 3.2) kann die SCV hergeleitet werden. Im ersten Schritt definieren McPherson et al. (1982, S. 1314) die Zufallsvariable

$$Y_i = \ln \left( \frac{O_i}{e_i} \right) \quad (3.3)$$

und führen folgende Approximation von  $Y_i$  ein

$$Y_i \approx \frac{O_i - e_i}{e_i}. \quad (3.4)$$

Diese Approximation wird im weiteren Verlauf der Diskussion eine wichtige Rolle spielen. Zunächst halten wir fest, dass McPherson et al. (1982, S. 1314) mit Hilfe der eingeführten Approximation – unter den Annahmen des Modells (3.1 – 3.2) – folgern, dass für alle  $i = 1, \dots, n$  gilt:  $E(Y_i) \approx 0$  und

$$E(Y_i^2) \approx \frac{1}{e_i^2} \text{Var}(O_i - e_i) = \frac{1}{e_i^2} (e_i^2 \sigma^2 + e_i) = \sigma^2 + \frac{1}{e_i}. \quad (3.5)$$

Auf der Grundlage von (3.5) leiten die Autoren den Momenten-basierten Schätzer von  $\sigma^2$  her, der als SCV  $\hat{\sigma}_{scv}^2$  bezeichnet wird und durch

$$\hat{\sigma}_{scv}^2 = \frac{1}{n} \sum_{i=1}^n \left( Y_i^2 - \frac{1}{e_i} \right) \quad (3.6)$$

definiert ist. Für diese Herleitung nutzen McPherson et al. (1982) die Erkenntnis, dass  $Y_i^2 = \sigma^2 + 1/e_i$  als eine “Schätzung” von  $E(Y_i^2) \approx \sigma^2 + 1/e_i$  in (3.5) aufgefasst werden kann. Den Ausdruck in (3.6) erhält man schliesslich als arithmetisches Mittel über alle  $Y_i^2 = \sigma^2 + 1/e_i$ ,  $i = 1, \dots, n$  (und Umstellung der Formel). Weil das arithmetische Mittel und nicht ein gewichtetes Mittel verwendet wird, haben alle Regionen den gleichen Einfluss auf die SCV, unabhängig von deren Grösse und

der Präzision der Schätzung.

Der Ausgangspunkt der SCV ist gemäss [McPherson et al. \(1982\)](#) das Poissonmodell in (3.1). Allerdings verlassen die Autoren diese theoretische Fundierung gleich wieder und formulieren (faktisch) ein Modell, das (implizit) unterstellt, dass die Verteilung von  $\ln(o_i/e_i)$  durch eine Normalverteilung angenähert werden kann. Dieses Vorgehen wird von den Autoren nicht begründet. Vielmehr erschliesst es sich einem aus dem Kommentar, den die Autoren gegen Schluss ihrer Besprechung nachreichen: "A plot of the  $\log_e(O_i/E_i)$  in each region [...] does not show any systematic departure from a symmetric Gaussian distribution." ([McPherson et al., 1982](#), S. 1314). Anstatt mit dem Ausdruck  $\ln(o_i/e_i)$  zu arbeiten, führen [McPherson et al. \(1982\)](#) die Approximation in (3.4) ein. Hierbei handelt es sich um die Taylor-Entwicklung erster Ordnung des natürlichen Logarithmus mit Entwicklungsstelle 1. Diese Approximation ist tauglich wenn sich die beobachteten Werte  $o_i$  und die Erwartungen  $e_i$  nur geringfügig unterscheiden. Dann besitzt das Verhältnis  $o_i/e_i$  einen Wert nahe 1.0. Für Verhältnisse, die deutlich von 1.0 abweichen, ist die Approximation ungeeignet. Dies tritt bspw. dann auf, wenn die Daten Überdispersion aufweisen.

**Anmerkung.** Unsere deskriptiven Untersuchungen in Kapitel 2.5 und Anhang D zeigen, dass die empirischen Verteilungen der meisten Indikatoren relative gut durch die Normalverteilung approximiert werden können. Wir werden in den Kapiteln 3.2 und 3.3 auf Methoden eingehen, welche entweder durchgehend mit dem Poissonmodell arbeiten oder direkt auf der Approximation  $(o_i - e_i)/e_i$  abstellen.

### 3.1.2 Eigenschaften der SCV

Der Schätzer der SCV nach [McPherson et al. \(1982\)](#) in (3.6)

- a) kann negative Werte annehmen;
- b) nimmt den Wert  $-\infty$  an, falls eine oder mehrere Regionen keine beobachteten Fallzahlen aufweisen (d. h. wenn  $o_i = 0$  ist).

In Fall a) ist es angezeigt,  $\hat{\sigma}_{scv}^2 = 0$  zu setzen; siehe bspw. [Diehr et al. \(1990, S. 754\)](#). Fall b) ist problematischer, weil es keine einfache Abhilfe gibt. [Diehr et al. \(1990, S. 754\)](#) formulieren den Schätzer der SCV leicht verändert und zwar als

$$\hat{\sigma}_{\text{Diehr}}^2 = \frac{1}{n} \sum_{i=1}^n \left\{ \left( \frac{O_i - e_i}{e_i} \right)^2 - \frac{1}{e_i} \right\} \quad (3.7)$$

d. h. anstelle von  $Y_i$  – siehe (3.3) – verwenden [Diehr et al. \(1990\)](#) die Approximation in (3.4); siehe auch [Diehr et al. \(1992, S. 485\)](#). Diehr und ihre Koautoren geben keine Begründung an, weshalb sie (3.7) der ursprünglichen Formulierung der SCV vorziehen. Es ist jedoch einfach zu sehen, dass

(3.7) stets wohl definiert ist und nicht den Wert  $-\infty$  annimmt, falls in einzelnen Regionen keine Fälle beobachtet werden.<sup>1</sup>

Der Schätzer  $\hat{\sigma}_{Diehr}^2$  ist Teil einer Klasse von nicht-iterativen Schätzern und ist mit dem Schätzer von Marshall (1991), der im Bereich Disease Mapping geläufig ist, verwandt; siehe Exkurs.

**Exkurs.** Böhning et al. (2002) haben eine Klasse nicht-iterativer, linear unverzerrter Schätzer von  $\sigma^2$  vorgestellt. Die SCV (in der Formulierung von Diehr et al., 1992) ist Teil dieser Klasse und wird bei (Böhning et al., 2002) als  $\hat{\tau}_1^2$  bezeichnet. Die Klasse ist definiert durch

$$\hat{\sigma}_\circ^2 = \frac{\sum_{i=1}^n w_i Z_i}{\sum_{i=1}^n w_i}, \quad (3.8)$$

wobei  $\circ$  als Platzhalter steht, mit

$$Z_i = \frac{(Y_i - e_i \mu)^2 - e_i \mu}{e_i^2}.$$

Die SCV von McPherson et al. (1982) in der Formulierung von Diehr et al. (1990), siehe  $\hat{\sigma}_{Diehr}^2$  in (3.7), erhält man für  $w_i = 1/n$  und  $\mu = 1$ . Die Verallgemeinerung für den Fall  $\mu \neq 1$  wurde von Böhning (2000) zit. nach Böhning et al. (2002) vorgeschlagen. Für  $w_i = e_i$  erhält man den Schätzer  $\hat{\sigma}_M^2$  von Marshall (1991). Die Wahl  $w_i = e_i^2$  führt zum Schätzer  $\hat{\sigma}_B^2$  von Bautista (1997) zit. nach Böhning et al. (2002).

Böhning et al. (2004) haben die Schätzer in der besagten Klasse in einer Simulationsstudie untersucht. Sie kamen zum Schluss, dass der Schätzer  $\hat{\sigma}_M^2$  von Marshall (1991) bezüglich des mittleren quadratischen Fehlers (mean square error) besser ist als die Schätzer  $\hat{\sigma}_{Diehr}^2$  (SCV) und  $\hat{\sigma}_B^2$ . Der Schätzer von Marshall (1991) wurde in der SAVA-Literatur (wie unsere Recherche zeigt) nur selten verwendet.

## Bewertung der SCV in der Literatur

Diehr et al. (1990, S. 754) haben in einer umfangreichen Simulationsstudie die Eigenschaften des SCV (und anderer Masse) untersucht. Es zeigt sich, dass die SCV sehr stark reagiert ("is sensitive") in Bezug auf veränderte Annahmen betreffend: a) Rate, b) Populationsgrösse, c) Variabilität der Regionengrösse und d) Mehrfachbehandlungen. In ihrer Würdigung der SCV formulieren (Diehr et al., 1990, S. 755) folgendes (sehr hartes) Urteil.

*The SCV was developed as a measure for comparing several surgery types, or the same surgery in two different regions. Its sensitivity to many factors other than true variability among small areas suggests that it does not fulfill this purpose, at least when the null hypothesis is true. (Diehr et al., 1990, S. 755)*

<sup>1</sup> Eine weitere Alternative, die auch nicht zum einem Schätzwert von  $-\infty$  führt, besteht darin,  $Y_i$  in (3.6) mit  $\tilde{Y}_i$  zu ersetzen, wobei  $\tilde{Y}_i = \ln\{(o_i + 0.5)/e_i\}$  die ln-Transformation mit einer Stetigkeitskorrektur darstellt.

Das Urteil ist in dieser Vehemenz überzogen. Weitere Einwände, die gegen die Verwendung der SCV in der Literatur vorgetragen wurden, sind nachfolgend zusammengetragen.

- Die Crude Rates spielen bei der SCV von [McPherson et al. \(1982\)](#) eine zentrale Rolle, obschon sie irreführend sein können bzw. sind, weil diese Raten je nach Grösse der Region und Häufigkeit der Fälle von unterschiedlicher Präzision sind ([Cressie, 1992](#); [Mollié, 1996](#); [Lawson et al., 2000](#)). Die Varianz der Crude Rates ist in kleinen Regionen gross — und entsprechend klein in vergleichsweise grossen Regionen. Es wäre jedoch wünschenswert, die regionale Analyse auf Kennzahlen abzustützen, die nicht durch die Grösse der Region und andere Drittfaktoren beeinflusst werden. Falls in zwei Regionen keine Fälle beobachtet werden, dann sind die jeweiligen Crude Rates in beiden Regionen null, unabhängig davon, ob die Regionen effektiv das gleiche Risikoprofil besitzen. Die Crude Rate kann nicht zwischen den beiden Fällen differenzieren.

Bei der Schätzung der SCV werden die Kennzahlen aller Regionen gemittelt, unabhängig davon, wie präzise die regionalen Kennzahlen sind; siehe [Shwartz et al. \(1994\)](#) und unsere Diskussion im Zusammenhang mit Formel (3.6). Andere Ansätze zur Quantifizierung der regionalen Variabilität (siehe Kapitel 3.2 und 3.3) verwenden ein gewichtetes Mittel, wobei die Gewichte die Präzision der Schätzung reflektieren.

- Ein Punkt, der in der Literatur (scheinbar) vergessen wurde, ist eine theoretische Inkonsistenz des SCV-Ansatzes. Und zwar folgender: Für die Schätzung der Überdispersion, beruht der Ansatz auf dem in Modell (3.1 – 3.2) bzw. einer Approximation. Es erscheint uns nun zwingend, dass auch die Berechnung der Vertrauensintervalle für die Raten in den Regionen auf Basis des Modells formuliert werden. Warum sollte ein 2-stufiges Modell für die Schätzung der SCV verwendet werden, das dann nicht für die Konstruktion von Vertrauensintervallen genutzt wird? Technisch ist es jedoch eine Herausforderung korrekte Vertrauensintervalle anzugeben. Dies liegt insbesondere daran, dass die A-priori-Verteilung in (3.2) nur durch die beiden ersten zentralen Momente charakterisiert ist.

### 3.1.3 Quantifizierung der regionalen Variabilität

Der Schätzer  $\hat{\sigma}_{scv}^2$  schätzt die regionale Überdispersion für das Modell (3.1 – 3.2). Um zu beurteilen, wie stark die Überdispersion ist, haben [McPherson et al. \(1996\)](#) die in Tab. 3.1 aufgeführten Intervalle und deren Bewertung vorgelegt. Wir konnten die Herleitung dieser Intervalle nicht nachvollziehen, weil der Research Report von [McPherson et al. \(1996\)](#) uns nicht zugänglich war.<sup>2</sup>

---

<sup>2</sup> Gemäss [www.worldcat.org](http://www.worldcat.org) [abgefragt am 14.1.2023] sind nur sechs gedruckte Exemplare verfügbar, davon fünf in Bibliotheken in Grossbritannien und ein Exemplar in der National Library of Medicine (USA). Keine Bibliothek verleiht das Dokument in digitaler Form.

**Tabelle 3.1** Intervalle für die Bewertung der SCV nach [McPherson et al. \(1996\)](#)

Bewertung:	tief	moderat	hoch	sehr hoch
Wertebereich:	< 3.0	3.0–5.4	5.4–10.0	>10.0

**Anmerkung.** Angesichts der in [Diehr et al. \(1990, S. 754\)](#) vorgelegten Resultate, dass die SCV sehr unterschiedliche Werte in Bezug auf veränderte Annahmen betreffend Populationsgrösse und Variabilität der Regionengrösse annimmt, erachten wir es als problematisch, allgemeingültige Intervalle oder Grenzwerte festlegen zu wollen.

### 3.1.4 Zwischenfazit

Das Zwischenfazit zur SCV mit Bezug auf die in Kapitel 1 formulierten Kriterien fällt – in aller Kürze – wie folgt aus (Symbolerklärung: ✓ mehrheitlich erfüllt; ~ gemischte Befunde; ✗ mehrheitlich nicht erfüllt).

- ~ Theoretische Fundierung
- ✓ Quantifizierung der regionalen Variabilität
- ~ Bemessung von “too much variation”
- ✗ Effizienz der Schätzer/ Stabilität der geschätzten Raten
- ✗ Vertrauensintervalle für die geschätzten Raten

## 3.2 Poisson–Gamma Empirical Bayes

Ausgangspunkt für die Methoden in diesem Kapitel ist ein *Conditionally Independent Hierarchical Bayes* (CIHB) Modell, analog zur Formulierung in Kapitel 3.1. Allerdings ist die A-priori-Verteilung nicht nur durch die beiden ersten Momente spezifiziert, sondern durch eine parametrische Verteilung. Die Parameter der A-priori-Verteilung sind unbekannt und sollen nach der *Empirical Bayes* (EB) Methode an der Randverteilung geschätzt werden.

### Modell und Verteilungsannahme

[Clayton und Kaldor \(1987\)](#) haben einen EB-Ansatz für die Modellierung alters-standardisierter relativer Risiken (Raten) für das Disease Mapping vorgeschlagen. Dieser Ansatz lässt sich ohne

grössere Anpassungen für eine Anwendung im Versorgungsatlas übernehmen. Das Modell von [Clayton und Kaldor \(1987\)](#) formuliert ein 2-stufiges CIHB-Modell,

$$O_i \mid \Theta_i = \theta_i \stackrel{ind}{\sim} \text{Pois}(\theta_i e_i), \quad i = 1, \dots, n, \quad (3.9)$$

$$\Theta_i \stackrel{iid}{\sim} \text{Gamma}(\alpha, \nu), \quad i = 1, \dots, n, \quad (3.10)$$

wobei  $\text{Gamma}(\alpha, \nu)$  die Gammaverteilung mit Skalenparameter  $\alpha > 0$  und Formparameter  $\nu > 0$  bezeichnet. Im Kontext der Bayes'schen Inferenzstatistik werden die Parameter  $\alpha$  und  $\mu$  als bekannte Grössen betrachtet. In der frequentistischen Inferenzstatistik sind die Parameter unbekannt und müssen geschätzt werden ( $\Rightarrow$  Empirical-Bayes Ansatz, siehe unten).

Analog zum 2-stufigen Modell bei der SCV (siehe Kapitel 3.1), wird auch hier die Poissonverteilung auf der ersten Stufe postuliert. Im Gegensatz zur Herleitung der SCV, beruht der Ansatz von [Clayton und Kaldor \(1987\)](#) durchgehend auf der Poissonverteilung (und wird nicht durch die Normalverteilung approximiert wie bei der SCV). Auf der zweiten Stufe wird eine Gammaverteilung spezifiziert, weil sie zur Poissonverteilung konjugiert ist. Diese Wahl wurde von [Clayton und Kaldor \(1987\)](#) getroffen, weil sie zu Schätzern in geschlossener Form führt.

### Spezialfall: Interne, indirekte Standardisierung

Für den Fall, dass die Standardisierung der Raten nach der *internal, indirect standardization* (IIS) Methode erfolgt, lässt sich einfach zeigen, dass  $\sum_{i=1}^n o_i = \sum_{i=1}^n e_i$  gilt (siehe bspw. [Böhning et al., 2004](#)). Daraus folgt

$$\frac{\sum_{i=1}^n o_i}{\sum_{i=1}^n e_i} = 1,$$

was die Wahl einer A-priori Verteilung mit Erwartungswert von 1 nahelegt ([Martuzzi und Hills, 1995](#)). Für die Gammaverteilung in (3.10) fordern wir darum ein, dass  $\alpha = \nu$  gelten soll. Diese Spezifikation stellt sicher, dass die Zufallsvariablen  $\Theta_i$  in (3.10) Erwartungswert 1 haben. Eine weitere Konsequenz der  $(\alpha = \nu)$ -Spezifikation ist, dass das Modell nur durch einen (Hyper-) Parameter (z. B. den Skalenparameter  $\alpha$ ) bestimmt ist, im Gegensatz zum ursprünglichen Modell von [Clayton und Kaldor \(1987\)](#). Das Ein-Parametermodell von [Martuzzi und Hills \(1995\)](#) stellt insofern einen Spezialfall des Modells von [Clayton und Kaldor \(1987\)](#) dar.

Die Eigenschaft, dass der Erwartungswert der  $\Theta_i$  im Ein-Parametermodell den Wert 1 besitzt, kann (wie oben angesprochen) als Folge der IIS-Methode hergeleitet werden. Sie besagt in Worten, dass das *nationale* Mittel der  $\theta_i e_i$  in (3.9) den Wert 1 hat. Die Werte der  $\theta_i e_i$  für die einzelnen Regionen können davon abweichen (und tun dies auch). Es ist an dieser Stelle wichtig zu betonen, dass die Standardisierung nach der IIS-Methode nicht zwingend das Ein-Parametermodell zufolge hat. Es ist ohne Weiteres denkbar (und je nach nach Anwendungsgebiet angezeigt), trotz Standardisierung nach der IIS-Methode das Zwei-Parametermodell von [Clayton und Kaldor \(1987\)](#) einzusetzen.

zen.<sup>3</sup> Wir bevorzugen das Ein-Parametermodell, weil es ein *Nullmodell* unter der folgenden Hypothese darstellt: Die regionale Überdispersion der Fallzahlen wird in Bezug zur *nationalen Referenz* bestimmt. Der Bezug zur nationale Referenz wird durch zwei Aspekte charakterisiert: 1) Die erwarteten Fallzahlen,  $e_i$ , sind nationale Kennzahlen, die nach der IIS-Standardisierung berechnet wurden. 2) Die regionale Überdispersion ist als multiplikativer Faktor in Bezug zur nationalen Ebene spezifiziert, und zwar so, dass der über die Nation gemittelte Faktor auf den Wert 1 normiert ist.

Die regionale Überdispersion wird im Ein-Parametermodell durch  $\text{Var}(\Theta_i) = 1/\alpha$  beschrieben. Falls die (geschätzte) Varianz null ist oder sich kaum von null unterscheidet, dann liegt keine (nennenswerte) Überdispersion vor. Je grösser die (geschätzte) Varianz, desto grösser ist (*ceteris paribus*) die (geschätzte) Überdispersion.<sup>4</sup>

### 3.2.1 Parameterschätzung

#### Zwei-Parametermodell

Die Schätzung der Parameter  $\alpha$  und  $\nu$  nach der Maximum-Likelihood-Methode (ML) ist numerisch instabil (Konvergenzprobleme). Darum haben Clayton und Kaldor (1987), Marshall (1991) und Lahiri und Maiti (2002) alternative Schätzmethoden vorgeschlagen.

#### Ein-Parametermodell

Die Parameterschätzung nach der ML-Methode ist für das Ein-Parametermodell numerisch wesentlich stabiler. Martuzzi und Hills (1995) verwenden eine ad-hoc-Implementation.<sup>5</sup> Mit den heutigen Computern kann die ML-Methode direkt berechnet werden. Eine notwendige Bedingung für Existenz der ML-Schätzung für  $\alpha$  ist

$$\sum_{i=1}^n (o_i - e_i)^2 > \sum_{i=1}^n e_i.$$

Falls diese Bedingung nicht erfüllt ist, dann gibt es keine Überdispersion und wir setzen  $\hat{\alpha} = 0$ .<sup>6</sup>

<sup>3</sup> Die Parameterschätzung für das Zwei-Parametermodell ist allerdings ineffizient, wenn das Ein-Parametermodell der datenerzeugende Prozess ist.

<sup>4</sup> Im Zwei-Parametermodell gilt  $\text{Var}(\Theta_i) = \nu/\alpha^2$ , wobei  $\alpha$  und  $\nu$  in der Anwendung durch Schätzungen ersetzt werden. Die Interpretation der regionalen Überdispersion ist analog zum Ein-Parametermodell.

<sup>5</sup> Sie berechnen die Likelihood-Funktion für feste Parameterwerte, die über ein Intervall gleichverteilt sind und erklären denjenigen Parameterwert zur ML-Schätzung, der den grössten Funktionswert besitzt.

<sup>6</sup> Wir machen den Beweis auf Anfrage gerne zugänglich.

### 3.2.2 Vorhersage/ Schätzung der Raten

Beim EB-Ansatz werden vorhergesagte Raten anstelle der Crude Rates verwendet, weil sie effizienter sind. Die Vorhersagen leiten sich aus dem Erwartungswert der A-posteriori Verteilung her

$$E(\Theta_i | O_i; \alpha, \nu) = \frac{O_i + \nu}{e_i + \alpha}, \quad i = 1, \dots, n. \quad (3.11)$$

Anstelle von  $\alpha$  und  $\nu$  werden die Schätzwerte  $\hat{\alpha}$  und  $\hat{\nu}$  in (3.11) eingesetzt. Der resultierende EB-Prädiktor für die Raten ist gegeben durch (Clayton und Kaldor, 1987)

$$\hat{\theta}_i = 1'000 \cdot \frac{O_i + \hat{\nu}}{e_i + \hat{\alpha}} = 1'000 w_i \cdot \frac{O_i}{e_i} + 1'000(1 - w_i) \cdot \frac{\hat{\nu}}{\hat{\alpha}}, \quad i = 1, \dots, n. \quad (3.12)$$

mit  $w_i = e_i / (e_i + \alpha)$ . Formel (3.12) vereinfacht sich für das Ein-Parametermodell, weil  $\hat{\alpha} = \hat{\nu}$  gilt. In beiden Modellen ist die vorhergesagte Rate  $\hat{\theta}_i$  ein gewichtetes Mittel (mit Gewichten  $w_i$ ) aus der Crude Rate  $O_i/e_i$  und  $\hat{\nu}/\hat{\alpha}$  bzw. 1. Die Gewichte  $w_i$  sind umgekehrt proportional zur Varianz von  $O_i/e_i$ . Für kleine Regionen mit vergleichsweise wenigen Fällen ist die Varianz gross und  $w_i$  entsprechend klein, so dass  $\hat{\theta}_i$  in stärkerem Mass von  $\hat{\nu}/\hat{\alpha}$  bzw. 1 als von  $O_i/e_i$  abhängt.

#### Vertrauensintervalle

Vertrauensintervalle für die vorhergesagte Raten  $\hat{\theta}_i$  in (3.12) können auf Grundlage des geschätzten Mittleren Quadratischen Fehlers (mean square error, MSE)  $MSE(\hat{\theta}_i)$  berechnet werden. Der MSE kann mit einem der folgenden Verfahren geschätzt werden:

- Jackknife-Schätzer von Jiang et al. (2002) oder Rao (2003, Kapitel 9.5.1)
- Bootstrap-Schätzer von Lahiri und Maiti (2002).

### 3.2.3 Quantifizierung der regionalen Variabilität

Die regionale Variabilität wird im Modell (3.9 – 3.10) durch die A-priori-Verteilung bestimmt. Die Varianz der  $\Theta_i$  ist durch  $\text{Var}(\Theta_i) = \nu/\alpha^2$  gegeben für das Zwei-Parametermodell (Clayton und Kaldor, 1987), resp.  $\text{Var}(\Theta_i) = 1/\alpha$  für das Ein-Parametermodell. Bei der geschätzten Varianz werden die Schätzwerte  $\hat{\alpha}$  und  $\hat{\nu}$  anstelle von  $\alpha$  und  $\nu$  eingesetzt.

Es bezeichne  $\{\hat{\theta}_i : i = 1, \dots, n\}$  die Menge der vorhergesagten Raten. Damit wird nun das *Quintile Ratio* berechnet (Coory und Gibberd, 1998)

$$\text{QR}(80/20) = \frac{Q_{80\%}(\hat{\theta}_i)}{Q_{20\%}(\hat{\theta}_i)}$$

wobei  $Q_{\alpha\%}(\hat{\theta}_i)$  das  $\alpha\%$ -Quantil der  $\hat{\theta}_i$  bezeichnet. Das QR ist also der Quotient zwischen einer hohen zu einer niedrigen Rate, wobei die hohe Rate dem 80%-Quantil und die niedrige Rate dem

20%-Quantil der vorhergesagten Raten entspricht. Sie drückt aus, um welchen Faktor die hohe Rate grösser ist als die niedrige Rate (z. B. Faktor 2.3). Die Wahl der Quantile (80% und 20%) stellt sicher, dass die Verhältniszahl nicht den Quotienten von Extremfällen oder Ausreissern abbildet (wie dies z. B. für das 95%- und das 5%-Quantil der Fall wäre).

**Anmerkung.** Das Quintile Ratio sollte nur für Topologien mit  $n > 5$  Regionen ausgegeben werden. Mit dem Schwellenwert  $n > 5$  ist sichergestellt, dass das erste Quintil grösser als der kleinste Wert ist und das QR nicht eine Verhältniszahl von Extrema ist.

Wir weisen noch einmal darauf hin, dass das Quintile Ratio auf Grundlage der vorhergesagten Raten  $\hat{\theta}_i$  berechnet wird, nicht den Crude Rates. Damit unterscheidet es sich massgeblich vom Extremal Quotient (EQ), das in der älteren SAVA-Literatur verwendet wurde. Das EQ wird bspw. in [Diehr et al. \(1990\)](#) besprochen.

### 3.2.4 Zwischenfazit

Das Zwischenfazit zum Poisson–Gamma-Modell in Bezug auf die in Kapitel 1 formulierten Kriterien fällt – in aller Kürze – wie folgt aus (Symbolerklärung: ✓ mehrheitlich erfüllt; ~ gemischte Befunde; ✗ mehrheitlich nicht erfüllt; ? keine Aussage möglich).

- ✓ Theoretische Fundierung
- ✓ Quantifizierung der regionalen Variabilität
- ? Bemessung von “too much variation”
- ✓ Effizienz der Schätzer/ Stabilität der geschätzten Raten
- ✓ Vertrauensintervalle für die geschätzten Raten

## 3.3 Normal–Normal Empirical Bayes

Das Normal–Normal CIHB-Modell nimmt für beide Stufen die Normalverteilung an. Es ist dasjenige CIHB-Modell, das am besten erforscht ist, was angesichts der herausragenden Stellung, welche die Normalverteilung einnimmt, nicht erstaunt. Wesentliche Beiträge zur EB-Schätzung für dieses Modell gehen auf [Efron und Morris \(1973\)](#) und [Efron und Morris \(1975\)](#) zurück.

Es bezeichne  $Y_i$  die *relative* Behandlungsrate ( $i = 1, \dots, n$ ). Sie ist definiert durch ([Efron und Morris, 1975](#); [Shwartz et al., 1994](#))

$$Y_i = \frac{O_i - e_i}{e_i}, \quad i = 1, \dots, n, \quad (3.13)$$

wobei  $O_i$  und  $e_i$  für, respektive, die Anzahl der beobachteten bzw. erwarteten Behandlungen in HSA  $i$  steht; siehe Kapitel 2.4. Die Grössen  $O_i$  sind Zufallsvariablen; wir schreiben  $o_i$  und meinen

damit eine Ausprägung von  $O_i$ . Unter Annahme 2 sind die  $e_i$ ,  $i = 1, \dots, n$ , keine Zufallsgrößen. Daraus folgt, dass die Varianz von  $Y_i$  durch

$$\widehat{\text{Var}}(Y_i) = D_i \approx \frac{O_i}{e_i^2}, \quad i = 1, \dots, n, \quad (3.14)$$

angenähert werden kann (Shwartz et al., 1994, S. 200). Die Varianzen in (3.14) werden in der Folge mit  $D_i$  bezeichnet.

### 3.3.1 Modell und Verteilungsannahme

Im weiteren Verlauf der Diskussion folgen wir Efron und Morris (1973, 1975) und formulieren das folgende CIHB-Modell

$$Y_i \mid \Theta_i = \theta_i \stackrel{ind.}{\sim} N(\theta_i, D_i), \quad i = 1, \dots, n, \quad (3.15)$$

$$\Theta_i \stackrel{iid}{\sim} N(0, A), \quad i = 1, \dots, n, \quad (3.16)$$

wobei  $N(\mu, \sigma^2)$  für die kumulative Verteilungsfunktion der Normalverteilung steht, mit Erwartungswert  $\mu$  und Varianz  $\sigma^2 \geq 0$ . Im Kontext der Bayes'schen Inferenzstatistik werden die Parameter  $\mu$  und  $\sigma^2$  als bekannte Größen betrachtet. In der frequentistischen Inferenzstatistik sind die Parameter unbekannt und müssen geschätzt werden ( $\Rightarrow$  Empirical-Bayes Ansatz).

#### Erste Stufe

Das Modell unterstellt auf der ersten Stufe, dass die  $Y_i$  eine (approximative) Normalverteilung besitzen; siehe (3.15). Diese Annahme kann, für grosse Fallzahlen, mit dem Zentralen Grenzwertsatz begründet werden; siehe bspw. Cressie (1992). Dies gilt unabhängig davon, ob die Behandlungsaktivität auf der Stufe des Individuums als unabhängige Bernoulli- oder Poisson-verteilte Zufallsvariable aufgefasst wird.

**Anmerkung.** Unsere deskriptiven Untersuchungen in Kapitel 2.5 und Anhang D weisen darauf hin, dass die empirischen Verteilungen der meisten Indikatoren durch die Normalverteilung approximiert werden können.

Das Modell in (3.15 – 3.16) kann auch als Approximation eines Poisson-Log-Normal CIHB-Modells aufgefasst werden (Ibáñez et al., 2009). Sei  $O_i \mid \Lambda_i = \lambda_i \sim \text{Pois}(e_i \lambda_i)$  die Verteilungen auf der ersten Stufe. Die Verteilung der zweiten Stufe ist in diesem Fall durch  $\ln(\Lambda_i) \sim N(0, A)$  gegeben ( $i = 1, \dots, n$ ).

**Exkurs.** Wie wir bereits ausgeführt haben, kann die Verteilung der relativen Behandlungsrate, wie sie in (3.13) definiert ist, für die Daten zu den Test-Indikatoren gut durch die Normalverteilung angenähert werden. Wenn dies nicht der Fall wäre, könnte man versuchen, die  $O_i$  derart

zu transformieren, dass deren Varianz nicht bzw. nur schwach vom Mittelwert abhängt (z. B. mit der Wurzeltransformation nach Freeman–Tukey); siehe bspw. [Cressie und Chan \(1989\)](#) und [Cressie \(1993, Kapitel 6.2\)](#).

## Zweite Stufe

Die Normalverteilung auf der 2. Stufe besagt, dass die relativen Raten  $\Theta_i$  eine Normalverteilung besitzen,  $i = 1, \dots, n$ ; siehe (3.16). Diese Annahme kann nicht direkt empirisch überprüft werden. Die A-priori-Verteilung in (3.16) hat Erwartungswert null. Dies ist eine Folge der linearen Beschränkung der  $Y_i$ -Werte; siehe (3.17) in der untenstehenden Anmerkung. Die Varianz der A-priori-Verteilung wird mit  $A$  bezeichnet und muss, im Kontext des EB-Ansatzes, geschätzt werden.

**Anmerkung.** Die relativen Behandlungsraten  $Y_i$  in (3.13) unterliegen einer linearen Beschränkung. Das heisst, es gibt Konstanten  $\{\alpha_i > 0 : i = 1, \dots, n\}$ , so dass gilt ([Efron und Morris, 1975](#))

$$\sum_{i=1}^n \alpha_i Y_i = 0. \quad (3.17)$$

Daraus folgt, dass der Erwartungswert  $E(Y_i) \approx 0$  ist für alle  $i = 1, \dots, n$ .

## 3.3.2 Parameterschätzung

Die Lokation in der A-priori-Verteilung in (3.16) ist auf null festgeschrieben und muss nicht geschätzt werden. Der Maximum-Likelihood-Schätzer  $\hat{A}$  von  $A$  ist als Nullstelle der Schätzgleichung ([Efron und Morris, 1975](#))

$$\sum_{i=1}^n \left( \frac{Y_i}{A + D_i} \right)^2 - \frac{1}{A + D_i} = 0$$

definiert, falls die Nullstelle positiv ist, sonst wird  $\hat{A} = 0$  gesetzt.

## 3.3.3 Vorhersage der Raten

Die Vorhersage der Raten unter dem Modell in (3.15–3.16) folgt aus dem Erwartungswert der A-posteriori-Verteilung und ist definiert durch ([Efron und Morris, 1975](#))

$$\hat{\theta}_i = (1 - \hat{B}_i)r_i \quad \text{mit} \quad \hat{B}_i = \frac{D_i}{\hat{A} + D_i}, \quad i = 1, \dots, n, \quad (3.18)$$

wobei  $r_i = 1'000 \cdot o_i/e_i$  die Crude Rate ist und  $\hat{A}$  den ML-Schätzer von  $A$  bezeichnet. Für den Schätzer  $\hat{B}_i$  in (3.18) ist bekannt, dass er die zu schätzende Grösse  $B_i = D_i/(A + D_i)$  überschätzt ([Morris, 1983](#)). Darum hat [Morris \(1983\)](#) eine korrigierte Variante vorgeschlagen, die durch den folgenden Ausdruck definiert ist

$$\hat{B}_i^* = \frac{(n-3)D_i}{(n-2)(\hat{A} + D_i)}. \quad (3.19)$$

Wir verwenden in der empirischen Analyse die korrigierte Variante von [Morris \(1983\)](#).

**Anmerkung.** Die Vorhersage der Raten in (3.18) bzw. (3.19) ist bei Vorliegen von Ausreißern ungeeignet. Als Alternative kann die *Limited Translation Rule* (LTR) von [Efron und Morris \(1972\)](#) verwendet werden. Diese Möglichkeit wird in [Shwartz et al. \(1994\)](#) angedeutet, aber nicht verwendet. In der deskriptiven Analyse in Kapitel 2.5 haben wir festgestellt, dass es nur wenige ausreisserbesetzte Verteilungen gibt, so dass man auf die LTR verzichten kann.

**Exkurs.** Eine weitere nennenswerte Alternative zu  $\hat{\theta}_i$  in (3.18) wurde von [Molina et al. \(2015\)](#) vorgeschlagen und beruht auf Vorarbeiten von [Datta et al. \(2011\)](#). Die alternative Methode schlägt einen Schätzer vor, der nur dann auf  $\hat{\theta}_i$  in (3.18) beruht, falls Überdispersion vorliegt. Unsere empirische Untersuchung zeigt, dass alle Test-Indikatoren eine deutliche Überdispersion aufweisen. Aus diesem Grund verzichten wir auf die Implementation der Methoden in [Molina et al. \(2015\)](#).

## Vertrauensintervalle

Vertrauensintervalle für die vorhergesagten Raten  $\hat{\theta}_i$  in (3.18) können auf Grundlage des geschätzten Mittleren Quadratischen Fehlers (mean square error, MSE) berechnet werden. Der MSE kann mit einem der folgenden Verfahren geschätzt werden:

- Taylor-Approximation von [Prasad und Rao \(1990\)](#)
- Jackknife-Verfahren von [Jiang et al. \(2002\)](#)
- Diverse Bootstrap-Verfahren; siehe bspw. [Rao und Molina \(2015\)](#)

Als Alternative können Vertrauensintervalle auf Grundlage der A-posteriori Varianz geschätzt werden; vgl. [Morris \(1983\)](#). Wir bevorzugen das Jackknife-Verfahren von [Jiang et al. \(2002\)](#).

### 3.3.4 Quantifizierung der regionalen Variabilität

Die regionale Variabilität wird im Modell (3.15 – 3.16) durch die A-priori-Verteilung bestimmt. Die Varianz der  $\Theta_i$  ist  $A$  und kann mit dem ML-Schätzer  $\hat{A}$  geschätzt. Falls  $\hat{A} > 0$  ist, dann gibt es Überdispersion.

Wir schlagen vor, auch für die empirische Analyse auf Basis des Modells in (3.15 – 3.16) das Quintile Ratio der vorhergesagten Raten zu berechnen; siehe Kapitel 3.2.3.

### 3.3.5 Zwischenfazit

Das Zwischenfazit zum Normal–Normal-Modell in Bezug auf die in Kapitel 1 formulierten Kriterien fällt – in aller Kürze – wie folgt aus (Symbolerklärung: ✓ mehrheitlich erfüllt; ~ gemischte Befunde; ✗ mehrheitlich nicht erfüllt; ? keine Aussage möglich).

- ✓ Theoretische Fundierung
- ✓ Quantifizierung der regionalen Variabilität
- ? Bemessung von “too much variation”
- ✓ Effizienz der Schätzer/ Stabilität der geschätzten Raten
- ✓ Vertrauensintervalle für die geschätzten Raten

In der Literatur (z. B. [Shwartz et al., 1994](#)) wird die EB-Schätzung für das Normal–Normal der SCV nach [McPherson et al. \(1982\)](#) vorgezogen. Zu dieser Einschätzung gelangen auch [Ibáñez et al. \(2009\)](#): “EB’s good behaviour is consistent with Shwartz’s results [12], and confirms that this statistic should become an essential part of SAVA studies.”

## 4 Analyse

In den vorangegangenen Kapiteln haben wir drei Methoden zur Quantifizierung der regionalen Variabilität vorgestellt und eine erste Bewertung vorgenommen, ob die Methoden für den geplanten Einsatz tauglich sind. Wir wollen nun den Einsatz der Methoden empirisch untersuchen (mit den in Kapitel 2 vorgestellten Testindikatoren), um so zu einer umfassenderen Einschätzung oder Bewertung zu gelangen. Die empirische Analyse verfolgt drei Ziele.

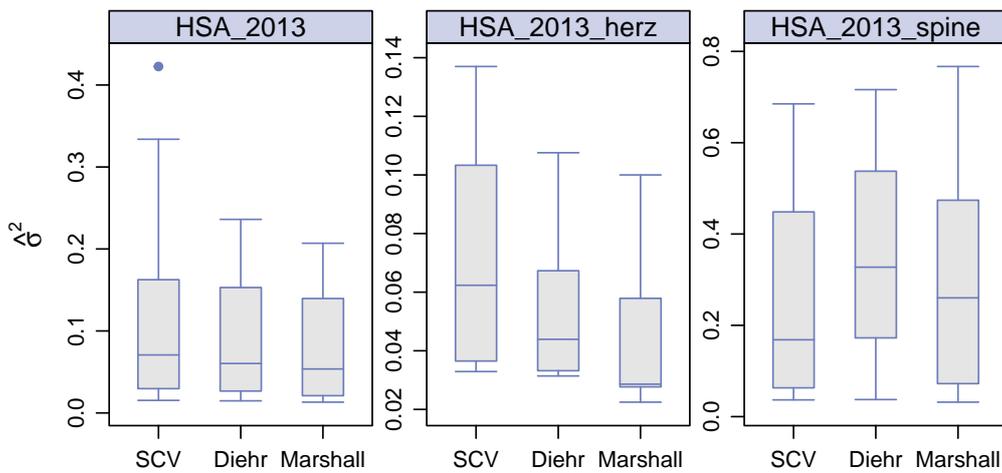
- Wir haben in Kapitel 3.1.2 mehrere Schätzer der SCV vorgestellt: den Schätzer von [McPherson et al. \(1982\)](#) in (3.6), den alternativen Schätzer von [Diehr et al. \(1990\)](#) in (3.7) und eine Schätzerklasse von [Böhning et al. \(2002\)](#), die unter anderen auch den Schätzer von [Marshall \(1991\)](#) beinhaltet; siehe (3.8). Wir wollen nun die konkurrierenden Schätzer mit den Daten zur MedStat untersuchen.
- In den Kapiteln 3.2 und 3.3 wurden die Grundlagen der EB-Methoden für das Poisson-Gamma- bzw. Normal-Normal-Modell besprochen. Nun wollen wir die beiden EB-Methoden mit den MedStat-Daten empirisch analysieren.
- Das Quintile Ratio QR(80/20) ([Coory und Gibberd, 1998](#)) wurde in den Kapiteln 3.2.3 und 3.3.4 als Mass vorgeschlagen, um die regionale Variabilität und eine allfällige Überdispersion "intuitiv" (Begriff aus [Coory und Gibberd, 1998](#)) zu verstehen. Wir haben das QR(80/20) für alle Testindikatoren der MedStat berechnet und wollen die Erkenntnisse dazu in diesem Kapitel besprechen.

### 4.1 SCV

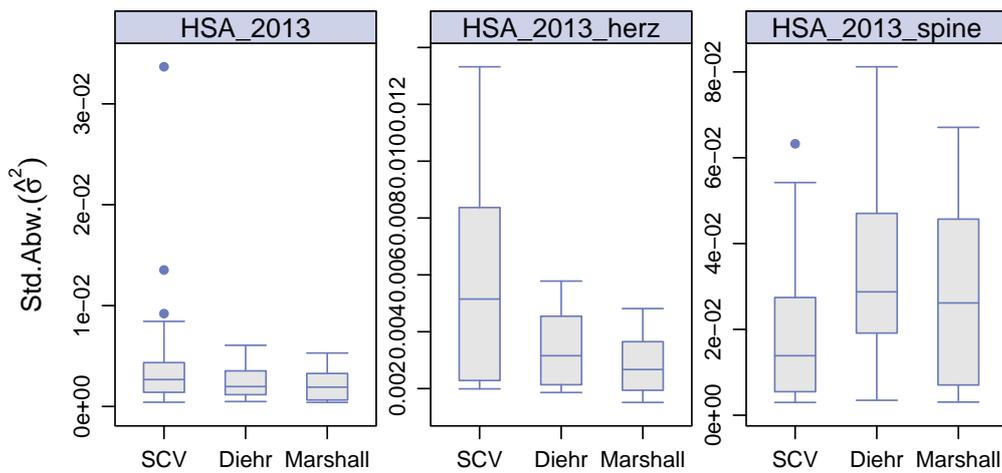
In Kapitel 3.1 haben wir drei Schätzmethoden vorgestellt um den Varianzparameter/ Überdispersionsparameter  $\sigma^2$  zu schätzen. Der bekannteste Schätzer ist die SCV von [McPherson et al. \(1982\)](#). Die anderen Schätzer (als "Diehr" und "Marshall" bezeichnet) wurden von [Diehr et al. \(1990\)](#) und [Marshall \(1991\)](#) vorgeschlagen. Wir verwenden für die Schätzer die generische Notation  $\hat{\sigma}^2$ ; die spezifischen Schätzer werden mittels Subskript gekennzeichnet, z. B.  $\hat{\sigma}_{Diehr}^2$ .

Für alle Indikatoren (gruppiert nach HSA-Topologie) wurden die drei Schätzer von  $\sigma^2$  mit den Daten zur MedStat 2013–2015 einzeln berechnet und in Abb. 4.1 als Boxplot dargestellt. Wir erkennen, dass die Schätzung des SCV-Schätzers von [McPherson et al. \(1982\)](#) tendenziell grösser ist als die Schätzungen der anderen Schätzer (Ausnahme: HSA\_2013\_herz, die nur 19 Regionen zählt). Aus der Abbildung geht auch hervor, dass der SCV-Schätzer eine grössere Streuung aufweist.

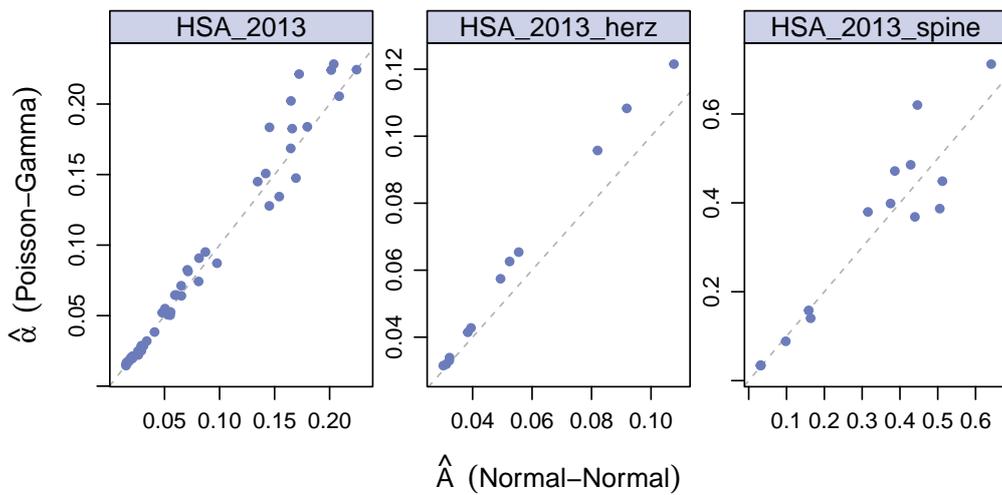
Ob die beobachteten Unterschiede auf eine systematische Überschätzung oder eine geringere Effizienz zurückzuführen ist, kann nicht beurteilt werden, weil der "wahre" Wert des Varianzparameters nicht bekannt ist. Zu weiteren Einsichten gelangen wir, wenn wir die Varianzschätzer betrach-



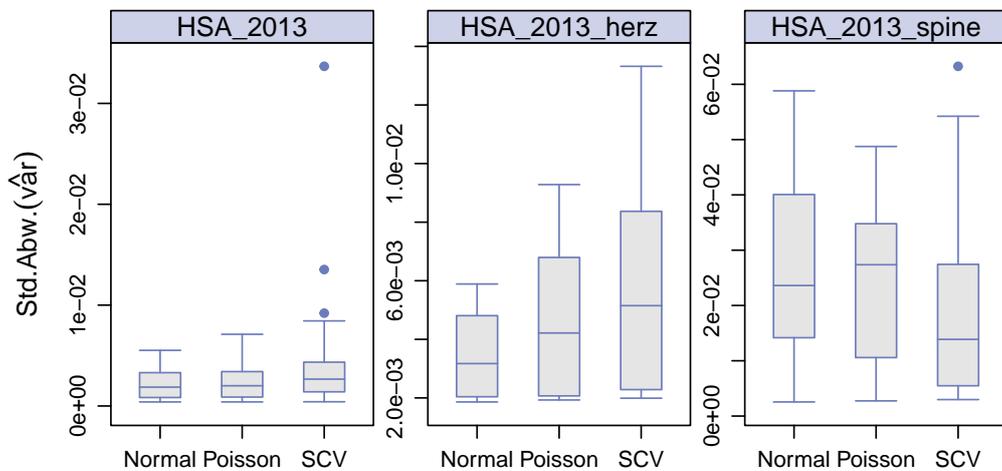
**Abbildung 4.1** Vergleich von drei Schätzern des Varianz- oder Überdispersionsparameters  $\sigma^2$ : SCV-Schätzer nach [McPherson et al. \(1982\)](#) und die alternativen SCV-Schätzer von [Diehr et al. \(1990\)](#) und [Marshall \(1991\)](#) (Daten: MedStat 2013–2015)



**Abbildung 4.2** Geschätzte Standardabweichung für die Schätzer von  $\sigma^2$ ; drei Schätzmethoden werden verglichen: SCV-Schätzer nach [McPherson et al. \(1982\)](#) und die alternativen SCV-Schätzer von [Diehr et al. \(1990\)](#) und [Marshall \(1991\)](#) (Varianzschätzer: Jackknife; Daten: MedStat 2013–2015)



**Abbildung 4.3** Gegenüberstellung der Schätzer des Varianzparameters in der A-priori-Verteilung ( $\alpha$  bzw.  $A$ ). Jeder Punkt repräsentiert einen Indikator. EB-Schätzer Normal-Normal ( $\hat{A}$ ) vs. Poisson-Gamma ( $\hat{\alpha}$ ), differenziert nach HSA-Topologie (gestrichelte Linie: 45-Grad-Linie, Daten: Med-Stat 2013–2015)



**Abbildung 4.4** Geschätzte Standardabweichung für die Schätzer des Varianzparameters in der A-priori-Verteilung ( $\sigma^2$ ,  $A$  oder  $\alpha$ ); drei Schätzmethoden werden verglichen: SCV-Schätzer nach [McPherson et al. \(1982\)](#) und die EB-Methoden Normal-Normal und Poisson-Gamma (Varianzschätzer: Jackknife; Daten: MedStat 2013–2015)

ten. Abb. 4.2 zeigt die geschätzte Standardabweichung (Std. Abw.) der Schätzer, Std.Abw. ( $\hat{\sigma}^2$ ), wobei  $\hat{\sigma}^2$  als Platzhalter für die Schätzer nach SCV, Diehr und Marshall steht. Die geschätzte Std. Abw. wurde nach dem Jackknife-Verfahren (vgl. bspw. Shao und Tu, 1996, Kapitel 2) berechnet. Aus Abb. 4.2 erkennen wir, dass der SCV-Schätzer von McPherson et al. (1982) (mit Ausnahme der HSA-Topologie HSA\_2013\_spine) die grösste Std. Abw. aufweist. Die Schätzer nach Diehr und Marshall sind etwa ähnlich effizient.

### Fazit

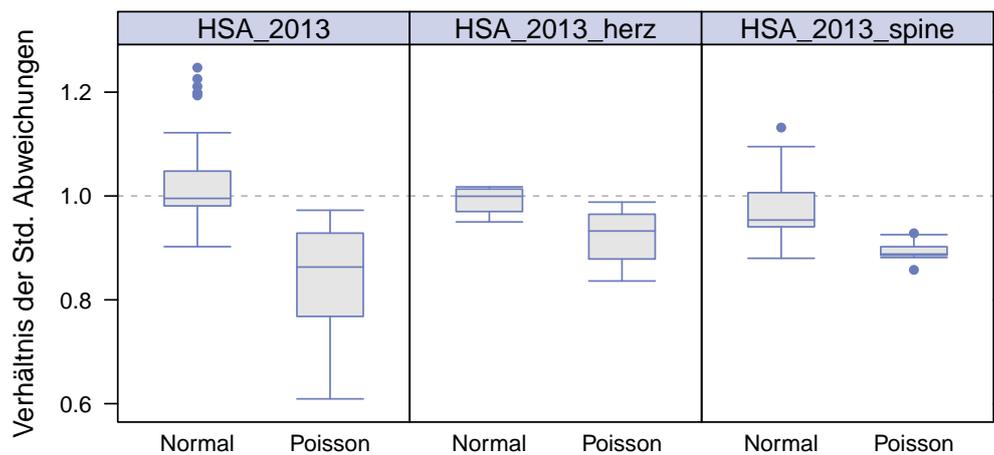
Der SCV-Schätzer nach McPherson et al. (1982) besitzt den Nachteil, dass er den Wert  $-\infty$  annimmt, falls eine oder mehrere Regionen keine Beobachtungen aufweisen. Darauf sind wir bereits in Kapitel 3.1 eingegangen. Der alternative Schätzer von Diehr et al. (1990) weist diese Problematik nicht auf und hat, wie wir oben diskutiert haben, in der Tendenz eine geringere Variabilität. Der Schätzer von Marshall (1991) war noch etwas effizienter. Wir schlagen dennoch vor, die SCV mit dem Schätzer von Diehr et al. (1990) zu berechnen, weil die Effizienzgewinne des Schätzers von Marshall (1991) überschaubar sind. Für den Schätzer von Diehr et al. (1990) spricht überdies, dass er dem Schätzer von McPherson et al. (1982) ähnlicher ist als der Marshall'sche Schätzer.

## 4.2 Normal–Normal und Poisson–Gamma Empirical Bayes

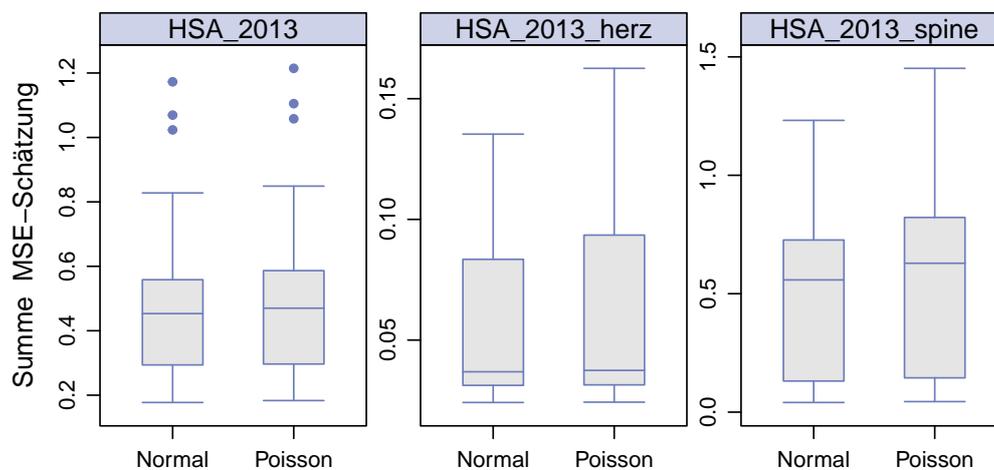
In diesem Kapitel wollen wir die EB-Schätzer der Modelle Normal–Normal und Poisson–Gamma vergleichen. Einen ersten Eindruck erhalten wir aus Abb. 4.3, in der die geschätzten Varianzparameter der A-priori-Verteilungen einander gegenüber gestellt werden. Jeder Punkt repräsentiert die Schätzung zu einer Region. Wir erkennen, dass die Schätzungen der Varianzparameter trotz unterschiedlicher Modelle einander relativ ähnlich sind. Grössere Unterschiede treten vor allem bei den Indikatoren der HSA-Topologie HSA\_2013\_herz auf.

Abb. 4.4 zeigt die geschätzte Standardabweichung (Std. Abw.) der Schätzer, Std.Abw. ( $\hat{\sigma}^2$ ), wobei  $\hat{\sigma}^2$  als Platzhalter für die EB-Schätzer, Normal (-Normal) und Poisson (-Gamma) und den SCV-Schätzer steht. Die geschätzte Std. Abw. wurde nach dem Jackknife-Verfahren (vgl. bspw. Shao und Tu, 1996, Kapitel 2) berechnet. Wir entnehmen der Abbildung, dass der EB-Schätzer für das Normal–Normal-Modell bei den HSA-Topologien HSA\_2013 und HSA\_2013\_herz die kleinste Varianz besitzt, gefolgt vom Schätzer zum Poisson–Gamma-Modell. Für die HSA-Topologie HSA\_2013\_spine ist die Streuung der geschätzten Std. Abw. bei allen drei Schätzern relativ ähnlich; der SCV-Schätzer weist im Median die kleinste Std. Abw. auf. Bei den anderen HSA-Topologien hingegen zeigt der SCV-Schätzer eine grössere Variabilität.

Wir wollen die Effizienz der beiden EB-Schätzer noch etwas genauer untersuchen und haben hierzu die unter dem Normal–Normal bzw. Poisson–Gamma-Modell vorhergesagten Raten  $\hat{\theta}_i$  und deren Std. Abw. berechnet. Die Std. Abw. ist definiert als die Wurzel des geschätzten mittleren qua-

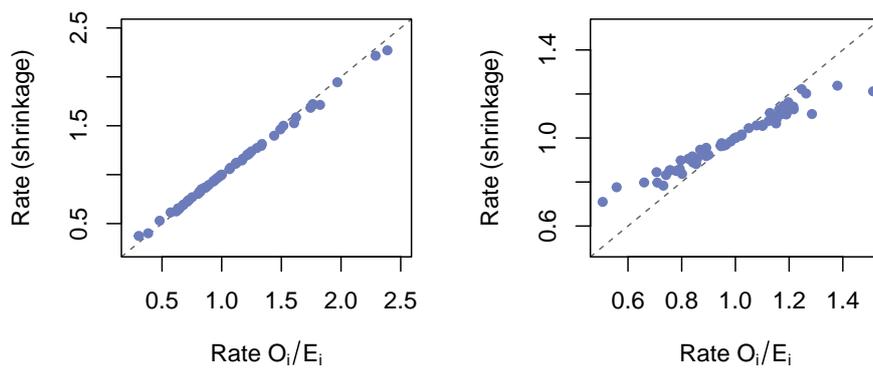


**Abbildung 4.5** Verhältnis Std.Abw.  $(o_i/e_i)$ /Std.Abw.  $(\hat{\theta}_i)$ , wobei Std.Abw.  $(\hat{\theta}_i)$  die Wurzel des geschätzten mittleren quadratischen Fehlers (MSE) der vorhergesagten Rate  $\hat{\theta}_i$  ist. Der MSE wurde für beide Modelle, Normal (-Normal) und Poisson (-Gamma), nach dem Jackknife-Verfahren berechnet. (Daten: MedStat 2013–2015)

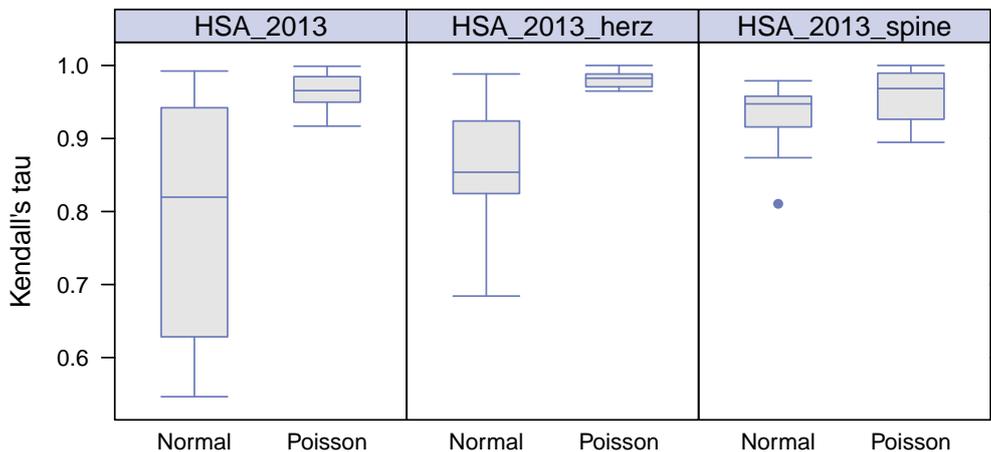


**Abbildung 4.6** Summe des geschätzten  $MSE\hat{\theta}_i$ , wobei über alle HSA  $i = 1, \dots, n$  summiert wird (Daten: MedStat 2013–2015)

dratischen Fehlers (MSE) der  $\hat{\theta}_i$  (siehe Kapitel 3.2.2 und 3.3.3). Der MSE wurde nach dem Jackknife-Verfahren geschätzt. Abb. 4.5 zeigt das Verhältnis der Std. Abw. der Crude Rates zur geschätzten Std. Abw. der  $\hat{\theta}_i$ . Wir sehen, dass die Schätzer der regionenspezifischen Raten für das Normal-Normal-Model klare Effizienzgewinne gegenüber den Schätzern der Crude Rates besitzen. Die Schätzer für das Poisson-Gamma sind weniger effizient. Die Unterschiede zwischen den beiden EB-Methoden sind bei der HSA-Topologie HSA\_2013 besonders ausgeprägt. Bei den restlichen HSA-Topologien sind Unterschiede wesentlich kleiner.



**Abbildung 4.7** Beispielhafte Illustration des Shrinkage-Verhaltens des EB-Schätzers (Poisson-Gamma) für die Test-Indikatoren ROTM (links) und HALL (rechts). Die Crude Rate,  $o_i/e_i$ , ist gegen die nach der EB-Methode vorhergesagte Rate,  $\hat{\theta}_i$ , aufgetragen (Daten: MedStat 2014)



**Abbildung 4.8** Rangkorrelation zwischen den Crude Rates  $o_i/e_i$  und den nach den EB-Methoden, Normal-Normal und Poisson-Gamma, berechneten Raten,  $\hat{\theta}_i$ , (Daten: MedStat 2013–2015)

Eine weitere Charakteristik, die wir untersuchen wollen, ist die Summe der geschätzten  $MSE(\hat{\theta}_i)$  für einen Indikator (die Summe wird über alle HSA gebildet). Die Resultate zu den beiden EB-

Methoden sind in Abb. 4.6 dargestellt. Auch hier sehen wir, dass die Schätzer für das Normal-Normal-Modell effizienter sind. Allerdings sind die Effizienzgewinne in Bezug auf die Schätzer des Poisson-Gamma-Modells nicht sehr ausgeprägt. Wir können festhalten, dass die EB-Methode zum Normal-Normal-Modell für *einzelne* Regionen wesentlich effizienter sind als diejenigen des Poisson-Gamma-Modells. In der *Summe* (d. h. für das Ensemble aller HSA) sind die Effizienzunterschiede zwischen den beiden Methoden deutlich geringer.

Die beiden EB-Schätzer zeigen das für sie charakteristische *shrinkage*-Verhalten. Wir haben dies für zwei Indikatoren (ROTM und HALL) und das Poisson-Gamma-Modell in Abb. 4.7 beispielhaft illustriert. Die Abbildung zeigt eine Gegenüberstellung der Crude Rates mit den nach dem Modell vorhergesagten Raten  $\hat{\theta}_i$  für alle HSA ( $i = 1, \dots, n$ ). Im linken Streudiagramm (Indikator ROTM) ist das shrinkage sehr schwach ausgeprägt. Die vorhergesagten Raten sind nahezu deckungsgleich mit den Crude Rates (mit Ausnahme der Werte an den Rändern, d. h. bei den Werten kleiner 0.5 und grösser 2.0). Anders sieht es beim Indikator HALL aus (rechtes Streudiagramm): sehr kleine und sehr grosse Crude Rates werden zum Wert 1.0 (d. h. dem Erwartungswert der A-priori-Verteilung) hin "geschrumpft".

Infolge des Shrinkage-Verhaltens, das die EB-Methoden einsetzen, kommt es zu Rangverschiebungen bei den vorhergesagten Raten im Vergleich zu den Crude Rates.<sup>1</sup> Zum Beispiel (fiktiv): Für die vorhergesagten Raten habe HSA 8 die grösste Rate und damit Rang 1. Bei den Crude Rates sei hingegen HSA 2 auf dem ersten Rang. Keines der beiden Rankings ist das "korrekte" Ranking. Letzteres beruht auf den wahren Raten; die sind jedoch unbekannt. Von beiden verfügbaren Rankings, welches ist das bessere? Das kann nicht abschliessend beantwortet werden. Das Ranking der vorhergesagten Raten wird bevorzugt.<sup>2</sup> Es ist insbesondere im Kontext der Amtlichen Statistik problematisch, wenn zwei von einander abweichende Rankings veröffentlicht werden. Das kann zu Widersprüchen führen, die kaum aufgelöst werden können. Es ist darum wünschenswert, EB-Methoden einzusetzen, deren vorhergesagten Raten sich in Bezug auf das Ranking nicht allzu stark vom Ranking der Crude Rates unterscheiden.

Um das Ausmass der Rangverschiebungen zu untersuchen, berechnen wir das Kendall'sche tau (Rangkorrelation) für die Ränge der Crudes Rates und den vorhergesagten Raten für jeden Indikator. Das Kendall'sche tau nimmt den Wert 1.0 wenn, die Ränge identisch sind und den Wert 0 falls keine Übereinstimmung zwischen den Rängen besteht. Die Resultate sind Abb. 4.8 dargestellt. Wir erkennen grosse Unterschiede zwischen den beiden EB-Methoden. Die nach dem Normal-Normal-Modell vorhergesagten Raten weichen in Bezug auf die Rangfolge deutlich von den Crude Rates ab. Für das Normal-Normal-Modell und die HSA-Topologie HSA\_2013 besitzt das Kendall'sche tau einen Median (über alle Indikatoren) von 0.82.

<sup>1</sup> Die hier diskutierten EB-Methoden beruhen zur Vorhersage der Raten auf dem Mittelwert der A-posteriori Verteilung. Für ebendiese Methoden ist bekannt, dass "[r]anking posterior means can perform poorly" (Shen und Louis, 1998); siehe auch Laird und Louis (1989).

<sup>2</sup> Sieht Rao (Kapitel 9.6.3 2003) für eine Diskussion der Rankingproblematik.

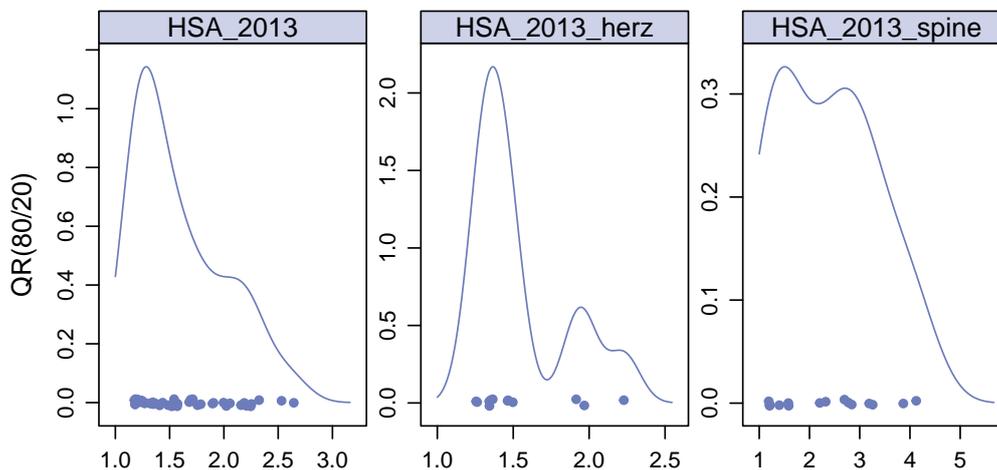
Beim Poisson–Gamma-Modell hat der Median für das Kendall'sche tau (gleiche HSA-Topologie) einen Wert von 0.96. Bei den anderen HSA-Topologien finden wir das gleiche Muster vor, allerdings sind die Werte für das Kendall'sche tau grösser und die Werte haben eine kleinere Streuung.

### Fazit

Die EB-Methode für das Normal–Normal-Modell ist in der Tendenz effizienter als die EB-Methode für das Poisson–Gamma-Modell. Diese Effizienzgewinne sind die Folge einer relativ starken Shrinkage der Raten. Mit der starken Schrumpfung gehen jedoch auch starke Rangverschiebungen im Vergleich zu den Crude Rates einher, was nicht wünschenswert ist. Aus diesem Grund (und weil die Effizienzunterschiede zwischen den Ansätzen nicht ausnehmend gross sind) bevorzugt wird die EB-Methode für das Poisson–Gamma-Modell.

## 4.3 Quantifizierung

Wir haben das Quintile Ratio QR(80/20) (Coory und Gibberd, 1998) für alle Testindikatoren der MedStat (Jahre 2013–2015) berechnet und in Tab. 4.1 aufgeführt. Das Quintile Ratio nimmt, je nach Indikator, Werte im Bereich 1.18 – 4.12 an. Die Verteilung der QR-Werte nach HSA-Topologie ist in Abb. 4.9 dargestellt. Die grösste Streuung der QR-Werte ist bei der Topologie HSA\_2013\_spine zu verzeichnen. Nennenswert ist auch die bimodale Verteilung bei HSA\_2013\_herz.



**Abbildung 4.9** Dichtefunktionen zur Verteilung des Quintile Ratio QR(80/20) nach HSA-Topologie (unterschiedliche Skalierung, Daten: MedStat 2013–2015)

**Tabelle 4.1** Quintile Ratio QR(80/20) (Poissonmodell, Daten: MedStat 2013–2015)

HSA	Indikator	2013	2014	2015
HSA_2013	APPE	1.32	1.23	1.19
HSA_2013	CHOL	1.18	1.21	1.18
HSA_2013	EGDB	2.00	2.02	2.06
HSA_2013	EGDO	1.57	1.54	1.52
HSA_2013	EHLB	1.41	1.35	1.35
HSA_2013	HALL	1.27	1.26	1.27
HSA_2013	HTEP	1.18	1.19	1.20
HSA_2013	INGH	1.57	1.68	1.71
HSA_2013	KTEP	1.23	1.21	1.18
HSA_2013	MENK	2.64	2.25	2.32
HSA_2013	RAFX	1.33	1.37	1.44
HSA_2013	ROTM	1.76	1.79	1.90
HSA_2013	SEKT	1.24	1.18	1.21
HSA_2013	TUHS	2.19	2.16	2.53
HSA_2013	TURP	1.49	1.69	1.56
HSA_2013	VSTR	2.21	1.90	2.25
HSA_2013_herz	BYPA	1.26	1.26	1.50
HSA_2013_herz	HKAT	1.97	1.91	2.23
HSA_2013_herz	PTCA	1.37	1.34	1.47
HSA_2013_herz	STEN	1.35	1.34	1.46
HSA_2013_spine	DECO	1.21	1.19	1.21
HSA_2013_spine	FUSI	1.59	1.58	1.40
HSA_2013_spine	KYPH	2.84	2.70	3.26
HSA_2013_spine	TDRE	4.12	3.87	2.77
HSA_2013_spine	VERT	3.19	2.32	2.21

## 5 Schlussfolgerungen und Empfehlungen

In der Literatur wurde eine Vielzahl von Methoden zur Quantifizierung der regionalen Variabilität von Behandlungsraten vorgeschlagen. Wir haben die folgenden drei Methoden herausgegriffen, weil sie für den Einsatz im Versorgungsatlas geeignet sind:

- Systematic Component of Variation ([McPherson et al., 1982](#))
- Empirical-Bayes Schätzer für das Poisson–Gamma-Modell ([Clayton und Kaldor, 1987](#); [Martuzzi und Hills, 1995](#))
- Empirical-Bayes Schätzer für das Normal–Normal-Modell ([Efron und Morris, 1975](#); [Shwartz et al., 1994](#))

Alle drei Methoden basieren auf den indirekt-standardisierten Raten und stellen nur minimale Anforderungen an die Verteilung der Daten. Wir haben insbesondere darauf verzichtet, Methoden zu berücksichtigen, welche Korrelationsstrukturen und Spatial Patterns (z. B. [Besag et al., 1991](#)) explizit modellieren.

In der empirischen Untersuchung zu 26 Testindikatoren auf Grundlage der Medizinischen Statistik der Krankenhäuser (Jahre 2013–2015) haben sich das Poisson–Gamma-Modell und die Schätzer von [Clayton und Kaldor \(1987\)](#) resp. [Martuzzi und Hills \(1995\)](#) als bester Kompromiss erwiesen. Die Schätzer für das Normal–Normal-Modell sind zwar tendenziell effizienter. Die für dieses Modell vorhergesagten Raten weisen jedoch (im Vergleich zu den Crude Rates und im Gegensatz zu den Schätzern unter dem Poisson–Gamma-Modell) massive Rangverschiebungen auf. Weil die Effizienzunterschiede zwischen den Ansätzen nicht ausnehmend gross sind, empfehlen wir das Poisson–Gamma- dem Normal–Normal-Modell vorzuziehen.

Zur Beurteilung des Ausmasses der regionalen Variation schlagen wir das Quintile Ratio ([Coory und Gibberd, 1998](#)) vor. Es handelt sich dabei um den Quotienten zwischen einer hohen zu einer niedrigen Rate, wobei die hohe Rate dem 80%-Quantil und die niedrige Rate dem 20%-Quantil der (nach der Empirical-Bayes-Methode vorhergesagten) Raten entspricht. Sie drückt aus, um welchen Faktor die hohe Rate grösser ist als die niedrige Rate (z.B. Faktor 2.3).

Die Systematic Component of Variation ([McPherson et al., 1982](#)) ist als Schätzer tendenziell weniger effizient als die Empirical-Bayes-Methoden für das Poisson–Gamma- und Normal–Normal-Modell. Weil die SCV jedoch in der Scientific Community der Small Area Variation Analysis eine etablierte Kennzahl ist, schlagen wir vor, die SCV im Versorgungsatlas als komplementäre Kennzahl aufzuführen, um die Anschlussfähigkeit an die Scientific Community sicherzustellen.

Das Obsan benutzt auf der Webseite mehrheitlich direkt-standardisierte Raten (in Bezug auf die European Standard Population). Um die Vergleichbarkeit zwischen den veröffentlichten Kennzahlen des Obsan sicherzustellen, ist es sinnvoll, im Versorgungsatlas auch die nach der direkten

Standardisierung berechneten Raten auszuweisen. Für die direkt-standardisierten Raten empfehlen wir die Konstruktion von Vertrauensintervallen nach der Methode von [Fay und Feuer \(1997a\)](#), weil es eine der konservativsten Methoden ist.

## Literaturverzeichnis

- Besag, J., York, J. und Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**, S. 1–20, DOI:10.1007/BF00116466.
- Böhning, D., Malzahn, U., Sarol, J., Rattanasiri, S. und Biggeri, A. (2002). Efficient Non-Iterative and Nonparametric Estimation of Heterogeneity Variance for the Standardized Mortality Ratio. *Annals of the Institute of Statistical Mathematics* **54**, S. 827–839, DOI:10.1023/A:1022419603608.
- Böhning, D., Sarol, J., Rattanasiri, S., Viwatwongkasem, C. und Biggeri, A. (2004). A comparison of non-iterative and iterative estimators of heterogeneity variance for the standardized mortality ratio. *Biostatistics* **5**, S. 61–74, DOI:10.1093/biostatistics/5.1.61.
- Breslow, N. E. und Day, N. E. (1987). *The design and analysis of cohort studies*, Band 2: Statistical methods in cancer research, Lyon: International Agency for Research on Cancer.
- Cain, K. C. und Diehr, P. (1992). Testing the null hypothesis in small area analysis. *Health Service Research* **27**, S. 267–294.
- Carriere, K. C. und Roos, L. L. (1997). A method of comparison for standardized rates of low-incidence events. *Medical Care* **35**, S. 57–69, DOI:10.1097/00005650-199701000-00005.
- Clayton, D. und Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* **43**, S. 671–681, DOI:10.2307/2532003.
- Coory, M. und Gibberd, R. (1998). New measures for reporting the magnitude of small-area variation in rates. *Statistics in Medicine* **17**, S. 2625–2634, DOI:10.1002/(SICI)1097-0258(19981130)17:22<2625::AID-SIM957>3.0.CO;2-4.
- Cressie, N. (1992). Smoothing Regional Maps Using Empirical Bayes Predictors. *Geographical Analysis* **24**, S. 75–92, DOI:10.1111/j.1538-4632.1992.tb00253.x.
- Cressie, N. A. C. (1993). *Statistics for Spatial Data*, New York: John Wiley and Sons.
- Cressie, N. und Chan, N. H. (1989). Spatial Modeling of Regional Variables. *Journal of the American Statistical Association* **84**, S. 393–401, DOI:10.1080/01621459.1989.10478783.
- Datta, G. S., Hall, P. und Mandal, A. (2011). Model Selection by Testing for the Presence of Small-Area Effects, and Application to Area-Level Data. *Journal of the American Statistical Association* **106**, S. 362–374, DOI:10.1198/jasa.2011.tm10036.
- Diehr, P., Cain, K., Connell, F. und Volinn, E. (1990). What is too much variation? The null hypothesis in small-area analysis. *Health Service Research* **24**, S. 741–771.
- Diehr, P., Cain, K. C., Kreuter, W. und Rosenkranz, S. (1992). Can Small-Area Analysis Detect Variation in Surgery Rates? The Power of Small-Area Variation Analysis. *Medical Care* **30**, S. 484–502, DOI:10.1097/00005650-199206000-00003.
- Diehr, P. und Grembowski, D. (1990). A small area simulation approach to determining excess variation in dental procedure rates. *American Journal of Public Health* **80**, S. 1343–1348, DOI:10.2105/AJPH.80.11.1343.
- Diggle, P. J. (2010). Historical Introduction. In: *Handbook of Spatial Statistics* hrsg. v. Gelfand, A. E., Diggle, P. J., Fuentes, M. und Guttorp, P. Boca Raton: CRC Press, S. 3–14, DOI:10.1007/978-1-4899-7603-1\_65.
- Efron, B. und Morris, C. (1972). Limiting the Risk of Bayes and Empirical Bayes Estimators—Part II: The Empirical Bayes Case. *Journal of the American Statistical Association* **67**, S. 130–139, DOI:10.1080/01621459.1972.10481215.
- Efron, B. und Morris, C. (1973). Stein's Estimation Rule and Its Competitors—An Empirical Bayes Approach. *Journal of the American Statistical Association* **68**, S. 117–130, DOI:10.1080/01621459.1973.10481350.

- Efron, B. und Morris, C. (1975). Data Analysis Using Stein's Estimator and its Generalizations. *Journal of the American Statistical Association* **70**, S. 311–319, DOI:10.1080/01621459.1975.10479864.
- Fay, M. P. und Feuer, E. J. (1997a). Confidence intervals for directly standardized rates: A method based on the gamma distribution. *Statistics in Medicine* **16**, S. 791–801, DOI:10.1002/(SICI)1097-0258(19970415)16:7<791::AID-SIM500>3.0.CO;2-%23.
- Fay, M. P. und Feuer, E. J. (1997b). A semi-parametric estimate of extra-Poisson variation for vital rates. *Statistics in Medicine* **16**, S. 2389–2401, DOI:10.1002/(SICI)1097-0258(19971115)16:21<2389::AID-SIM683>3.0.CO;2-J.
- Fay, M. P. und Feuer, E. J. (2007). A semi-parametric estimate of extra-Poisson variation for vital rates – Authors' reply. *Statistics in Medicine* **26**, S. 2133–2135, DOI:10.1002/sim.2745.
- Fleiss, J. L., Levin, B. und Paik, M. C. (2003). *Statistical Methods for Rates and Proportions*, Hoboken (NJ): John Wiley and Sons, 3. Aufl. DOI:10.1002/0471445428.
- Gail, M. (1978). The Analysis of Heterogeneity for Indirect Standardized Mortality Ratios. *Journal of the Royal Statistical Society. Series A* **141**, S. 224–234, DOI:10.2307/2344454.
- Glynn, R. J., Stukel, T. A., Sharp, S. M., Bubolz, T. A., Freeman, J. L. und Fisher, E. S. (1993). Estimating the Variance of Standardized Rates of Recurrent Events, with Application to Hospitalizations among the Elderly in New England. *American Journal of Epidemiology* **137**, S. 776–786, DOI:10.1093/oxfordjournals.aje.a116738.
- Haynes, A. G., Wertli, M. M. und Aujesky, D. (2020). Automated delineation of hospital service areas as a new tool for health care planning. *Health Service Research* **55**, S. 469–475, DOI:10.1111/1475-6773.13275.
- He, Y. und Escobar, M. (2007). A semi-parametric estimate of extra-Poisson variation for vital rates. *Statistics in Medicine* **26**, S. 2131–2133, DOI:10.1002/sim.2743.
- Ibáñez, B., Librero, J., Bernal-Delgado, E., Peiró, S., López-Valcarcel, N. M. B. G. und Aizpuru, F. (2009). Is there much variation in variation? Revisiting statistics of small area variation in health services research. *BMC Health Services Research* **9**, S. 1–12, DOI:10.1186/1472-6963-9-60.
- Jiang, J., Lahiri, P. und Wan, S.-M. (2002). A unified jackknife theory for empirical best prediction with  $M$ -estimation. *The Annals of Statistics* **30**, S. 1782–1810, DOI:10.1214/aos/1043351257.
- Kendall, M. und Gibbons, J. D. (1990). *Rank Correlation Methods*, New York: Oxford University Press, 5. Aufl.
- Lahiri, P. und Maiti, T. (2002). Empirical Bayes Estimation of Relative Risks in Disease Mapping. *Calcutta Statistical Association Bulletin* **53**, S. 213–224, DOI:10.1177/0008068320020304.
- Laird, N. M. und Louis, T. A. (1989). Empirical Bayes Ranking Methods. *Journal of Educational Statistics* **14**, S. 29–46, DOI:10.3102/10769986014001029.
- Lawson, A. B., Biggeri, A. B., Boehning, D., Lesaffre, E., Viel, J. F., Clark, A., Schlattmann, P. und Divino, F. (2000). Disease mapping models: an empirical evaluation. Disease Mapping Collaborative Group. *Statistics in medicine* **19**, S. 2217–2241, DOI:10.1002/1097-0258(20000915/30)19:17/18<&t;2217::aid-sim565>3.0.co;2-e.
- Marshall, R. J. (1991). Mapping disease and mortality rates using empirical Bayes estimators. *Journal of the Royal Statistical Society. Series C* **40**, S. 283–294, DOI:10.2307/2347593.
- Martuzzi, M. und Hills, M. (1995). Estimating the Degree of Heterogeneity between Event Rates Using Likelihood. *American Journal of Epidemiology* **141**, S. 369.–374, DOI:10.1093/aje/141.4.369.
- McPherson, K., Downing, A. und Buirski, D. (1996). Systematic variation in surgical procedures and hospital admission rates: A methodological study. Report II. Technical Report PHP departmental publication no. 23, London School of Hygiene and Tropical Medicine, Dept. of Public Health & Policy, Health Promotion Sciences Unit, London.

- McPherson, K., Wennberg, J. E., Hovind, O. B. und Clifford, P. (1982). Small-Area Variations in the Use of Common Surgical Procedures: An International Comparison of New England, England, and Norway. *New England Journal of Medicine* **307**, S. 1310–1314, DOI:10.1056/NEJM198211183072104.
- Miettinen, O. S. (1972). Standardization of Risk Ratios. *American Journal of Epidemiology* **96**, S. 6, DOI:10.1093/oxfordjournals.aje.a121470.
- Molina, I., Rao, J. N. K. und Datta, G. S. (2015). Small area estimation under a Fay-Herriot model with preliminary testing for the presence of random area effects. *Survey Methodology* **41**, S. 1–19.
- Mollié, A. (1996). Bayesian mapping of disease. In: *Markov Chain Monte Carlo in Practice* hrsg. v. Gilks, W. R., Richardson, S. und Spiegelhalter, D. J. Dordrecht: Springer Science+Business Media, Kap. 20, S. 359–379.
- Morris, C. N. (1983). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association* **78**, S. 47–55, DOI:10.1080/01621459.1983.10477920.
- Ng, H. K. T., Filardo, G. und Zheng, G. (2008). Confidence interval estimating procedures for standardized incidence rates. *Computational Statistics and Data Analysis* **52**, S. 3501–3516, DOI:10.1016/j.csda.2007.11.004.
- Pickle, L. W. und White, A. A. (1995). Effect of the choice of age-adjustment on maps of death rates. *Statistics in Medicine* **14**, S. 615–627, DOI:10.1002/sim.4780140519.
- Prasad, N. G. N. und Rao, J. N. K. (1990). The Estimation of the Mean Squared Error of Small-Area Estimators. *Journal of the American Statistical Association* **85**, S. 163–171, DOI:10.2307/2289539.
- R Development Core Team (2023). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, <http://www.R-project.org/>, ISBN 3-900051-07-0.
- Rao, J. (2003). *Small Area Estimation*, Hoboken (NJ): John Wiley and Sons, DOI:10.1002/0471722189.
- Rao, J. und Molina, I. (2015). *Small Area Estimation*, Hoboken (NJ): John Wiley and Sons, 2. Aufl. DOI:10.1002/9781118735855.
- Shao, J. und Tu, D. (1996). *The Jackknife and Bootstrap*, New York: Springer-Verlag, 2. Aufl.
- Shao, J. (2007). *Mathematical Statistics*, New York: Springer, 2. Aufl.
- Shen, W. und Louis, T. A. (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society: Series B* **60**, S. 455–471, DOI:10.1111/1467-9868.00135.
- Shwartz, M., Ash, A. S., Anderson, J., Iezzoni, L. I., Payne, S. M. C. und Restuccia, J. D. (1994). Small Area Variations in Hospitalization Rates: How Much You See Depends on How You Look. *Medical Care* **32**, S. 189–201, DOI:10.1097/00005650-199403000-00001.
- Swift, M. B. (1995). Simple confidence intervals for standardized rates based on the approximate bootstrap method. *Statistics in Medicine* **14**, S. 1875–1888, DOI:10.1002/sim.4780141704.
- Swift, M. B. (2009). Comparison of confidence intervals for a Poisson mean – further considerations. *Communications in Statistics. Theory and Methods* **38**, S. 748–759, DOI:10.1080/03610920802255856.
- Swift, M. B. (2010). A simulation study comparing methods for calculating confidence intervals for directly standardized rates. *Computational Statistics and Data Analysis* **54**, S. 1103–1108, DOI:10.1016/j.csda.2009.10.021.
- Waller, L. und Carlin, B. (2010). Disease Mapping. In: *Handbook of Spatial Statistics* hrsg. v. Gelfand, A. E., Diggle, P. J., Fuentes, M. und Guttorp, P. Boca Raton: CRC Press, Kap. 14, S. 217–243, DOI:10.1201/9781420072884.
- Waller, L. W. und Gotway, C. A. (2004). *Applied Spatial Statistics for Public Health Data*, Hoboken (NJ): John Wiley and Sons, DOI:10.1002/0471662682.

## A Test-Indikatoren

---

*MedStat: HSA-Topologie: HSA\_2013*

---

ALLA	Alle Hospitalisierungen
APPE	Laparoskopische Appendektomie
CHOL	Laparoskopische Cholezystektomie
EGDB	Ösophagogastroduodenoskopie [EGD] mit geschlossener Biopsie
EGDO	Ösophagogastroduodenoskopie
EHBL	Exzision oder Destruktion von Läsion oder Gewebe der Harnblase, transurethral
HALL	Plastische Rekonstruktion bei Hallux valgus mit Exostosenresektion, Weichteilkorrektur und Osteotomie am Os metatarsale I
HTEP	Totalendoprothese des Hüftgelenks
INGH	Operation Inguinalhernie
KTEP	Totalendoprothese des Kniegelenks
MENK	Meniskektomie am Knie, arthroskopisch
RAFX	Offene Reposition einer distalen Radiusfraktur mit innerer Knochenfixation
ROTM	Plastische Rekonstruktion der Rotatorenmanschette
SEKT <sup>(1)</sup>	Sectio caesarea
TUHS	Transurethrales Entfernen einer Obstruktion von Ureter und Nierenbecken, Entfernung eines Harnsteines
TURP <sup>(2)</sup>	Transurethrale Prostatektomie
VSTR	Crossektomie und Stripping, V. saphena magna

---

*MedStat: HSA-Topologie: HSA\_2013\_spine*

---

DECO	Dekompression am Spinalkanal (ohne gleichzeitige Diskushernie)
FUSI	Fusion (Spondylodese)
KYPH	Kyphoplastik
TDRE	Einsetzen von totaler Diskusprothese
VERT	Vertebroplastik

---

*MedStat: HSA-Topologie: HSA\_2013\_herz*

---

BYPA	Bypass
HKAT	Herzkatheter ohne Stents, arteriographie coronaire
PTCA	Perkutanen Transluminalen Koronarangioplastie
STEN	Stents

---

<sup>(1)</sup> Nur Frauen im Alter 18–50 Jh.; <sup>(2)</sup> nur Männer

## B Methoden

In diesem Anhang werden folgende Themen diskutiert:

- Konstruktion von Vertrauensintervalle für direkt-standardisierte Raten (Kapitel [B.1](#))
- Kendall'sche Konkordanzkoeffizient (Kapitel [B.2](#))

### B.1 Vertrauensintervalle zur direkten Standardisierung

In diesem Abschnitt besprechen wir die Herleitung und Berechnung von Vertrauensintervallen (VI) für Raten, die nach der Methode der direkten Standardisierung berechnet sind.

Wir nehmen im Folgenden an, dass es  $j = 1, \dots, J$  Altersgruppen gibt. Zur Vereinfachung der Diskussion differenzieren wir nicht nach Geschlecht und betrachten nur eine HSA. Die Anzahl der Fälle in der  $j$ -ten Altersgruppen sei mit  $O_j$  bezeichnet, wobei  $O_j$  eine Zufallsvariable ist. Die Realisierungen werden mit  $o_j$  bezeichnet. In der Literatur zur direkten Standardisierung (vgl. bspw. [Breslow und Day, 1987](#); [Swift, 1995](#); [Fay und Feuer, 1997a](#); [Ng et al., 2008](#)) wird – nahezu einhellig – unterstellt, dass gilt

$$O_j \stackrel{\text{ind}}{\sim} \text{Pois}(\theta_j) \quad (j = 1, \dots, J) \quad (\text{B.1})$$

mit  $\theta_j = r_j n_j$ , wobei  $r_j$  die wahre Rate und  $n_j$  die Anzahl der Individuen in der  $j$ -ten Altersgruppe bezeichnen. Die Fallzahlen  $n_j$  sind feste, nicht-stochastische Größen.

**Anmerkung.** Die Annahme, dass die  $O_j$  für  $j = 1, \dots, J$  als stochastisch unabhängige Zufallsvariablen aufgefasst werden, wird in der Literatur nicht infrage gestellt; siehe bspw. [Breslow und Day \(1987, S. 59\)](#). Letztere gehen auf die Tauglichkeit der Unabhängigkeitsannahme detailliert ein. Die Annahme der Poissonverteilung hingegen sollte hinterfragt werden. Wir kommen auf diesen Aspekt im weiteren Verlauf der Besprechung zurück.

Die direkte Standardisierung ist in Bezug auf die Standard- oder Referenzpopulation (z. B. European Standard Population) definiert. Die Anteilswerte der Altersgruppen in der Referenzpopulation seien mit  $s_j$  bezeichnet, so dass gilt  $\sum_{j=1}^J s_j = 1$ . Die  $s_j$  werden als feste, nicht-stochastische Größen aufgefasst. Die nach der direkten Standardisierung berechnete Rate in der Studienpopulation ist nun gegeben durch

$$r^{\text{dir}} = \sum_{j=1}^J s_j \frac{O_j}{n_j} = \sum_{j=1}^J w_j O_j$$

mit Gewicht  $w_j = s_j/n_j$ . Unter der Annahme unabhängig Poisson-verteilter Zufallsvariablen in [\(B.1\)](#) folgt, dass für den Erwartungswert und die Varianz von  $O_j$  gilt  $E(O_j) = \text{Var}(O_j) = \theta_j$ ,  $j =$

$1, \dots, J$ . Daraus folgt umgekehrt, dass  $E(r^{\text{dir}}) = \sum_{j=1}^J w_j \theta_j$  und  $\text{Var}(r^{\text{dir}}) = \sum_{j=1}^J w_j^2 \theta_j$ . Der Schätzer von  $r^{\text{dir}}$  nach der Maximum-Likelihood-Methode ist nun gegeben durch

$$\hat{r}^{\text{dir}} = \sum_{j=1}^J w_j o_j.$$

Einen Schätzer für die Varianz erhält man mit

$$\widehat{\text{Var}}(r^{\text{dir}}) = \sum_{j=1}^J w_j^2 o_j.$$

Beide Schätzer berücksichtigen nicht, dass die Altersgruppen-spezifischen Schätzer  $o_j$  der Raten  $\theta_j$  unterschiedlich präzise sind (Breslow und Day, 1987, S. 59).

Für den Schätzer  $\hat{r}^{\text{dir}}$  sollen nun Methoden zur Konstruktion von Vertrauensintervallen untersucht werden. Die Konstruktion von VI ist nicht trivial, weil nur  $J$  Beobachtungen,  $o_j$ , vorliegen, um die  $J$  Parameter  $\theta_j$  zu bestimmen. Es ist auch diesem Umstand geschuldet, dass in der Literatur eine Vielzahl von Konstruktionsmethoden vorgeschlagen wurde. Einen guten Überblick zu den Konstruktionsmethoden von VI im Kontext der direkten Standardisierung bietet die Arbeit von Ng et al. (2008). Es handelt sich dabei um eine grossangelegte Simulationsstudie, in der 20 Methoden verglichen werden. Swift (2010) hat die Simulationsstudie um die ABC-Methode (approximate bootstrap confidence nach DiCiccio–Efron, siehe Swift, 1995) ergänzt. Insgesamt stehen also 21 Methoden zur Auswahl.

Die Methoden zur Konstruktion von Vertrauensintervallen werden nach den folgenden Kriterien beurteilt (vgl. bspw. Shao, 2007, Kapitel 7): 1) Diskrepanz zwischen tatsächlicher und nomineller Überdeckungswahrscheinlichkeit (coverage probability), 2) Ausdehnung und 3) Invarianz. Wenn die tatsächliche Überdeckungswahrscheinlichkeit grösser als die nominelle ist, dann wird das Intervall, genauer gesagt die Methode als "konservativ" bezeichnet. Bei gleicher Überdeckungswahrscheinlichkeit, wird diejenige Methode bevorzugt, deren Intervall eine kleinere Ausdehnung besitzt (weil sie informativer ist). Das Kriterium der Invarianz ist für den Fall der standardisierten Raten von untergeordneter Bedeutung.<sup>1</sup> Die genannten Aspekte/ Kriterien sind komplementär zu einander: Eine "Verbesserung" in einer der Dimensionen kann nur zulasten der anderen Dimensionen erfolgen.

In der Studie von Ng et al. (2008) war die Methode von Fay und Feuer (1997a) die konservativste Methode (aber auch die Methode mit der grössten VI-Ausdehnung). Aus diesem Grund schlagen wir vor, diese Methode zu verwenden. Sie ist wie folgt definiert. Für ein fest vorgegebenes  $\alpha \in$

<sup>1</sup> In der Literatur werden zusätzliche Kriterien besprochen, wie Symmetrie und das Vorliegen einer mathematisch geschlossenen Form; siehe bspw. Swift (2009). Beide Kriterien sind im Kontext der direkten Standardisierung nicht von zentraler Bedeutung und werden darum auch nicht weiter diskutiert.

$(0, 1)$  ist das  $100(1 - \alpha)\%$  VI nach [Fay und Feuer \(1997a\)](#) definiert als  $(G_u, G_o)$  mit den Grenzen

$$G_u = \frac{\widehat{\text{Var}}(r^{\text{dir}})}{2\widehat{r}^{\text{dir}}} \chi_{\alpha/2}^2 \left( \frac{2(\widehat{r}^{\text{dir}})^2}{\widehat{\text{Var}}(r^{\text{dir}})} \right),$$

$$G_o = \frac{\widehat{\text{Var}}(r^{\text{dir}}) + \max(w_j)^2}{2(\widehat{r}^{\text{dir}} + \max(w_j))} \chi_{1-\alpha/2}^2 \left( \frac{2(\widehat{r}^{\text{dir}} + \max(w_j))^2}{\widehat{\text{Var}}(r^{\text{dir}}) + \max(w_j)^2} \right),$$

wobei  $\chi_q^2(d)$  das  $q$ -Quantil der Chi-Quadratverteilung mit  $d$  Freiheitsgraden und  $\max(w_j)$  das Maximum der Gewichte  $w_j$  bezeichnen. Falls keine Fälle für die Schätzung vorliegen, schlagen wir vor, die Rate und ihr VI nicht anzugeben.

**Anmerkung.** Die Konstruktion der VI beruht auf der Annahme einer Poissonverteilung; siehe [\(B.1\)](#). Diese Verteilung besitzt die Besonderheit, dass die Varianz nicht gesondert geschätzt werden muss, weil sie gleich dem Erwartungswert ist. Bei Vorliegen von Überdispersion (overdispersion oder extra-Poisson variation) gilt dies nicht. [Fay und Feuer \(1997b\)](#) haben einen Vorschlag unterbreitet – siehe auch [He und Escobar \(2007\)](#) und [Fay und Feuer \(2007\)](#) –, um Vertrauensintervalle bei Vorliegen von Überdispersion anzugeben; siehe auch [Glynn et al. \(1993\)](#) und [Carriere und Roos \(1997\)](#). Diese Methode ist jedoch wesentlich komplexer und erfordert die Schätzung der Varianz über mehrere Zeitperioden.

Regionale Variation ist im Kontext von Zählraten fast immer auf Überdispersion zurückzuführen. Es ist darum problematisch, VI zu konstruieren, deren Gültigkeit die Abwesenheit von Überdispersion voraussetzt.

## B.2 Kendall'scher Konkordanzkoeffizient (Kendall's W)

Der Kendall'sche Konkordanzkoeffizient (auch als Kendall'sches W bezeichnet) ist eine nichtparametrische statistische Methode, die in der Praxis z. B. eingesetzt wird, um zu quantifizieren, wie gut die Ränge der Bewertungen mehrerer Beurteiler (rater) übereinstimmen (Kendall und Gibbons, 1990, Kapitel 6). Diese Methode soll nun zur Beurteilung herangezogen werden, ob es – verkürzt formuliert – im Zeitverlauf stets die gleichen Regionen sind, die Höchstwerte bei einem Indikator aufweisen. Als "Bewertungen" fungieren hier die Ränge der Regionen für einen Indikator (z. B. Region 5 mit der höchsten Rate steht auf dem ersten Rang, gefolgt von Region 21 mit der zweithöchsten Rate auf Rang 2 usw.). Anstelle von "Beurteilern" denken wir uns verschiedene, aufeinander folgende Zeitpunkte, z. B. Jahre 2013, 2014 und 2015. Für jedes Jahr wird eine separate Rangstatistik erstellt; siehe Tabelle B.1 mit den Ränge für den Indikator BYPA (2013–2015).

Der Konkordanzkoeffizient gibt nun Aufschluss darüber, wie gut die Rangstatistiken mehrerer Jahre übereinstimmen. Er nimmt Werte zwischen 0 (keine Übereinstimmung) und 1 (vollständige Konkordanz/ Übereinstimmung) an. Der Wert 0 tritt auf, wenn die Bewertungen "essentially random" sind (Kendall und Gibbons, 1990, Kapitel 6). Für die MedStat-Daten der Jahre 2013–2015 treten Werte zwischen 0.6495 (CHOL) und 0.9606 (MENK) auf. Für die Interpretation der Werte kann es hilfreich sein, eine Analogie zum Korrelationskoeffizienten von Pearson zu bilden (und z. B. Werte grösser als 0.9 als "sehr stark" zu bezeichnen).

**Tabelle B.1** Ränge (pro HSA und Jahr) der Raten für den Indikator BYPA. Die HSA mit der grössten Rate hat den Rang 1 (Daten: MedStat 2013–2015)

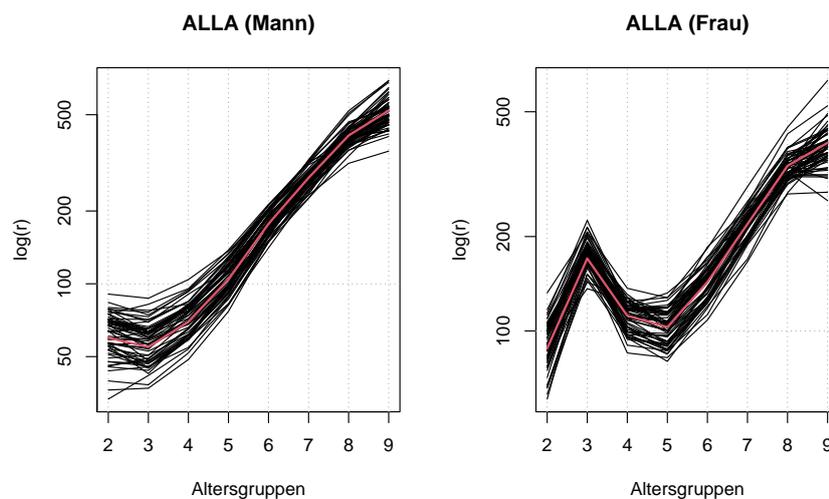
HSA	2013	2014	2015
AG07	17	13	16
BE05	18	16	15
BL16	5	4	4
BS02	16	12	13
FR02	13	19	9
GE05	2	2	3
GE14	1	3	2
GR01	3	1	1
LU02	9	7	7
LU04	11	6	17
SG01	15	17	18
TG16	7	18	11
TI20	19	15	19
VD03	8	9	6
VS17	4	8	8
ZH05	14	10	10
ZH07	10	11	14
ZH19	12	14	12
ZH72	6	5	5

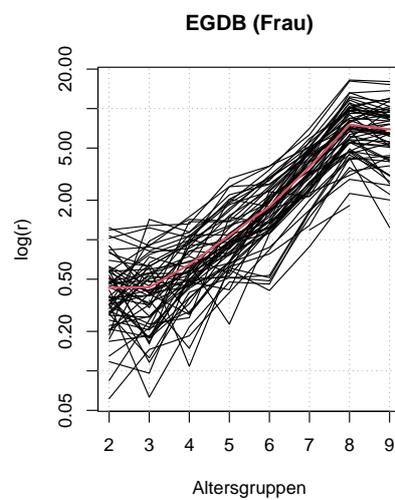
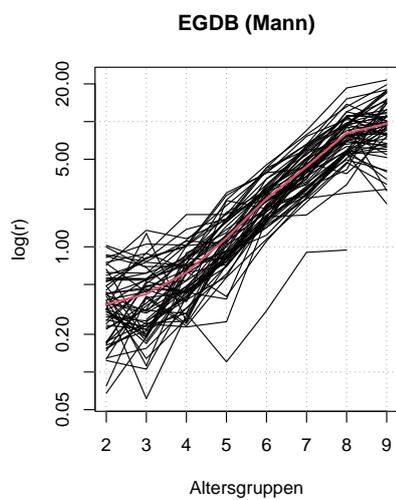
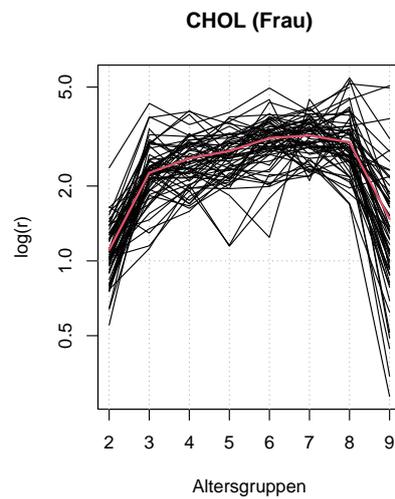
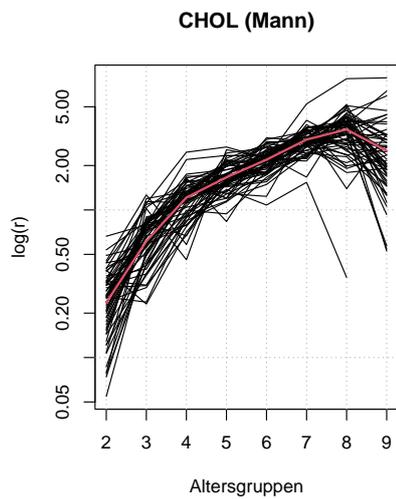
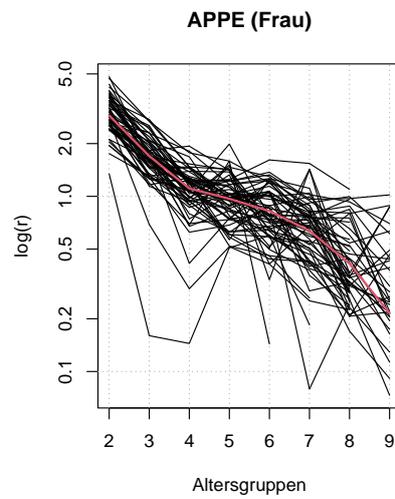
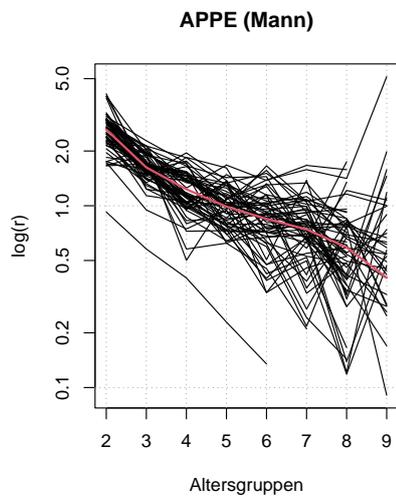
## C Diagnostikplot nach Gail (1978)

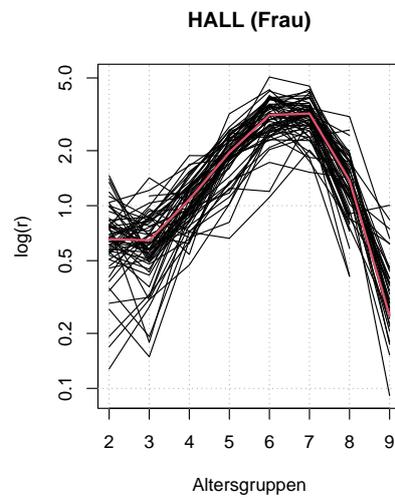
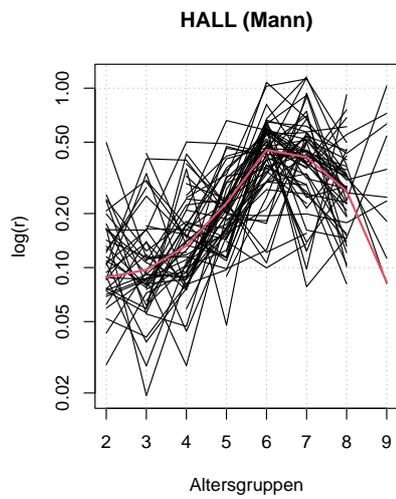
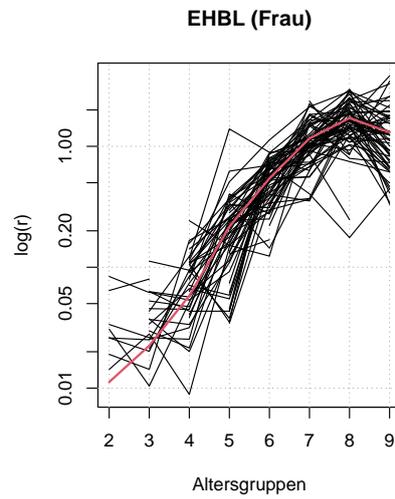
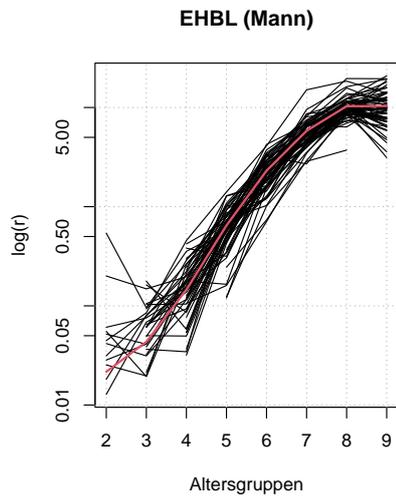
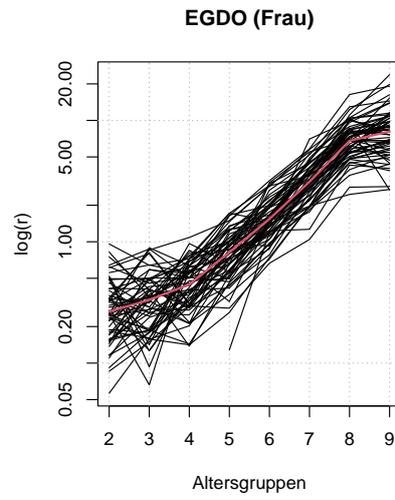
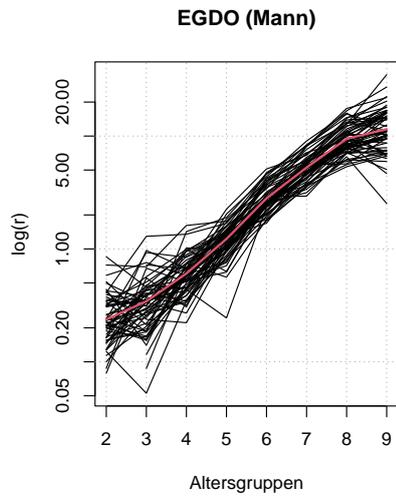
Der Diagnostikplot nach [Gail \(1978\)](#) trägt  $\log(r_{ij})$  (für jede einzelne Region  $i = 1, \dots, n$ ) gegen die Altersgruppe  $j$  auf und verbindet die Punkte einer Region mit einer Linie. Wir bezeichnen die einzelnen Linien als Profile. Die Altersgruppen umfassen jeweils zehn Jahre. Jeder Plot zeigt zusätzlich auch das Profil der Rate für die Schweiz (keine Regionalisierung, rote Linie). Falls alle Linien durch eine vertikale Parallelverschiebungen ineinander überführt werden können, dann ist die Proportionalitätsbedingung nach [Pickle und White \(1995\)](#) erfüllt, so dass die direkt- und indirekt.standardisierten Raten identisch sind.

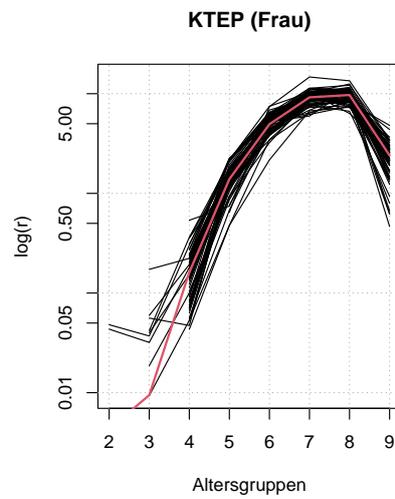
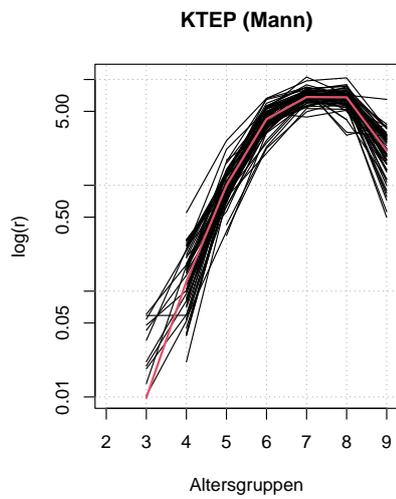
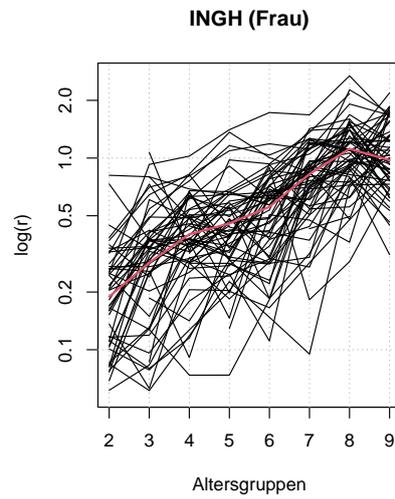
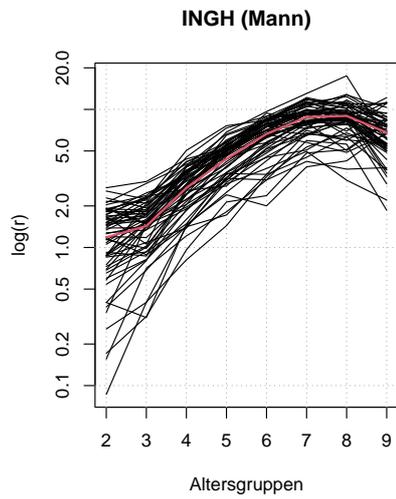
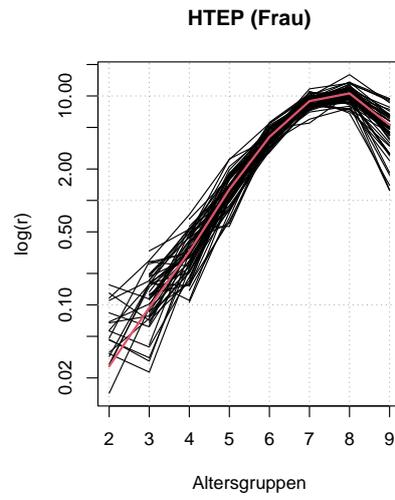
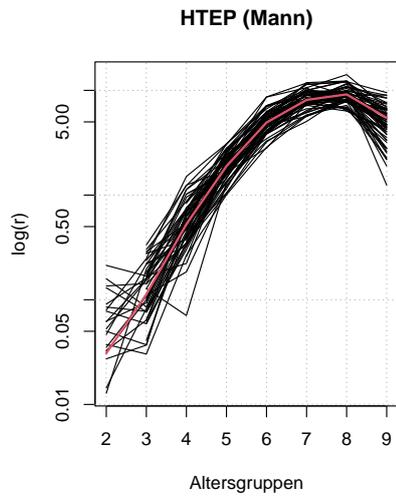
Bei einigen Indikatoren gibt nur sehr wenige Fälle. Als Folge davon sind die Profile durch Zufallsfehler ("Rauschen") überlagert. Dieses Phänomen wird auch bei [Gail \(1978\)](#) besprochen. Es liesse sich durch die Zusammenlegung von Altersklassen abschwächen oder beheben. Wir haben darauf verzichtet, weil es für unsere Zwecke ausreicht, uns einen groben Eindruck zu verschaffen.

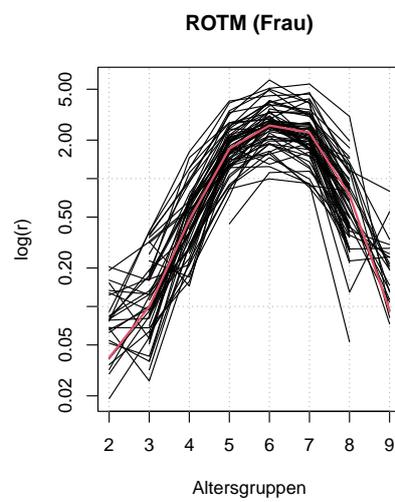
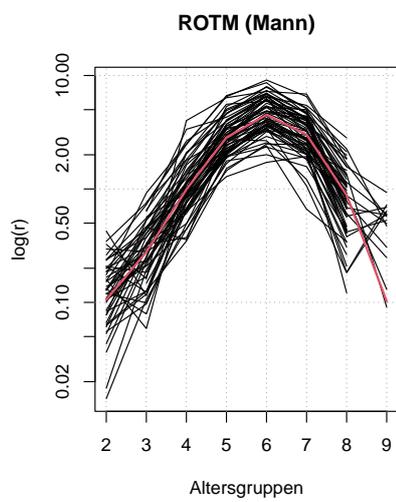
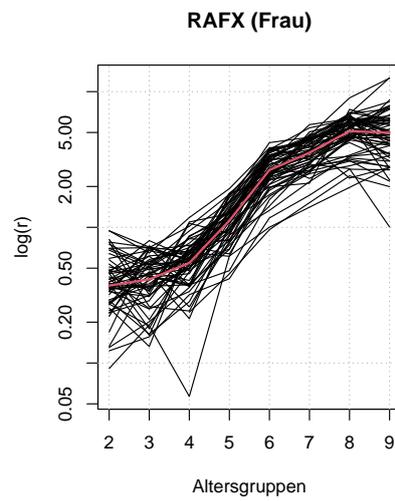
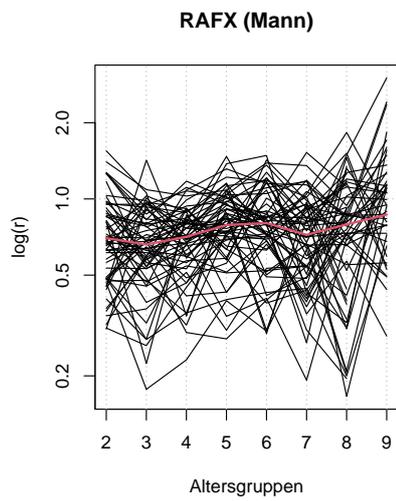
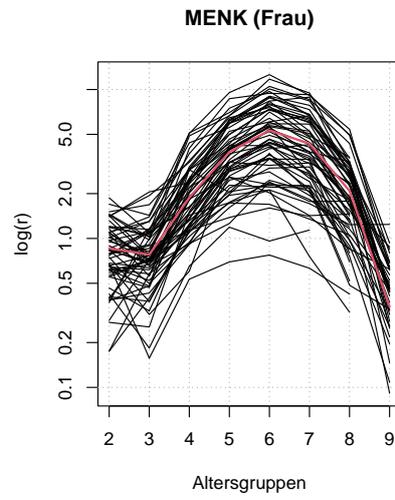
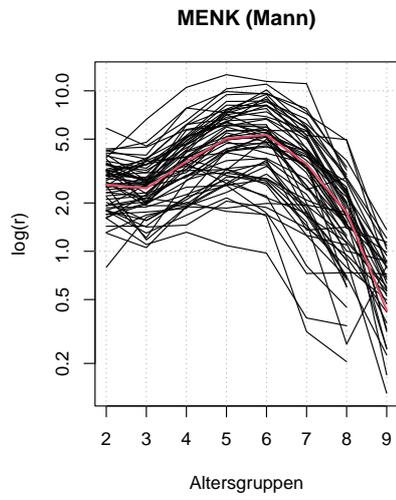
Nachfolgend sind die Diagnostikplots zu allen Indikatoren (ohne weitere Angaben) aufgeführt.



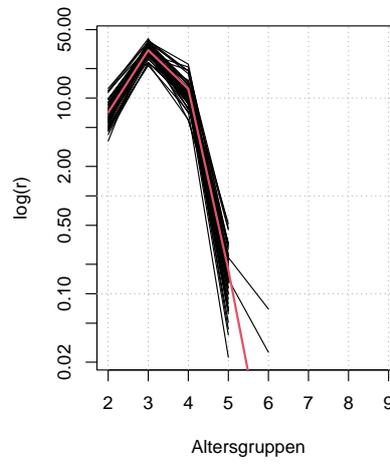




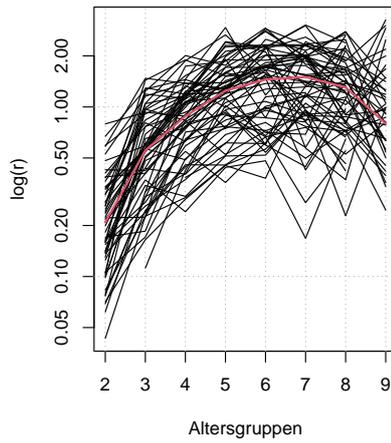




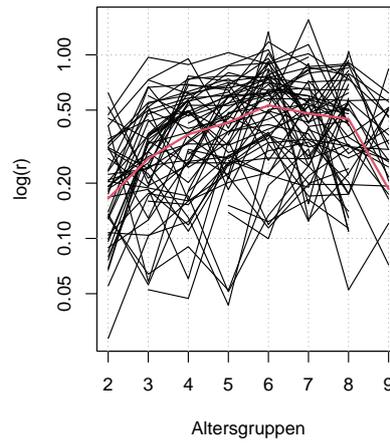
**SEKT (Frau)**



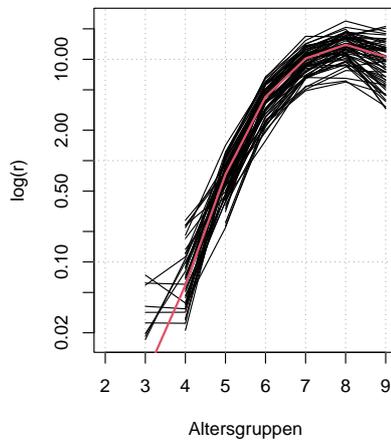
**TUHS (Mann)**



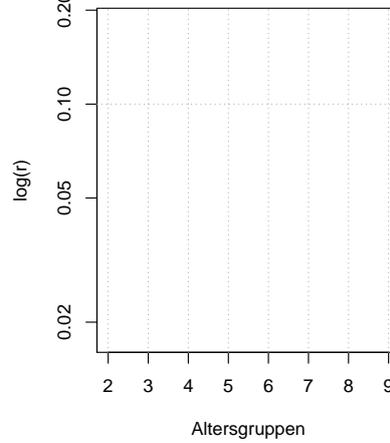
**TUHS (Frau)**

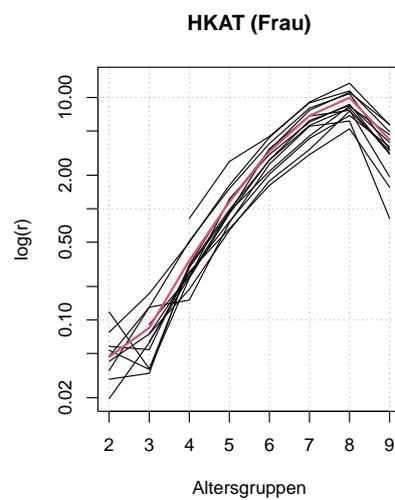
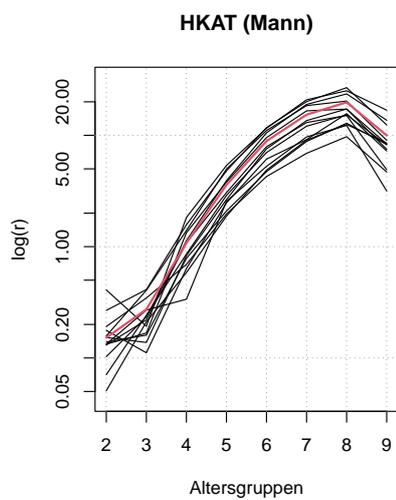
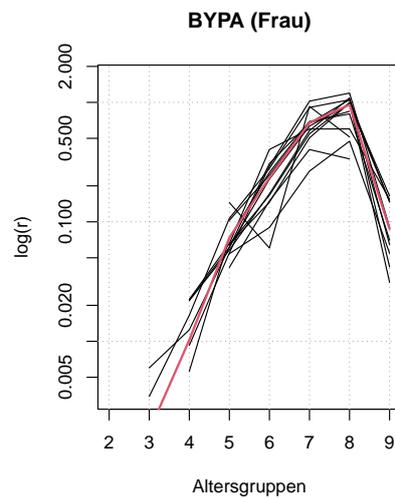
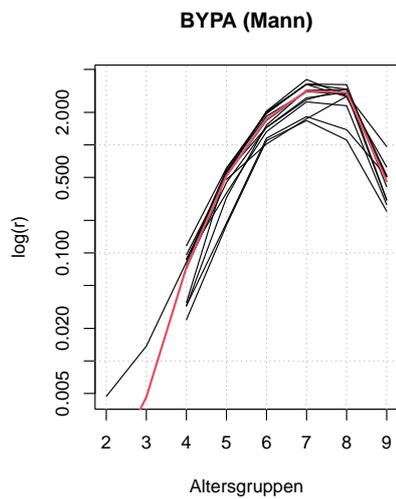
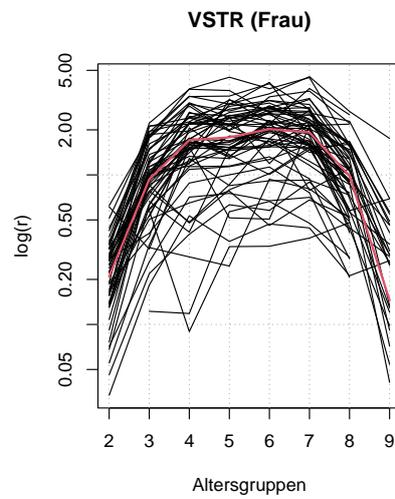
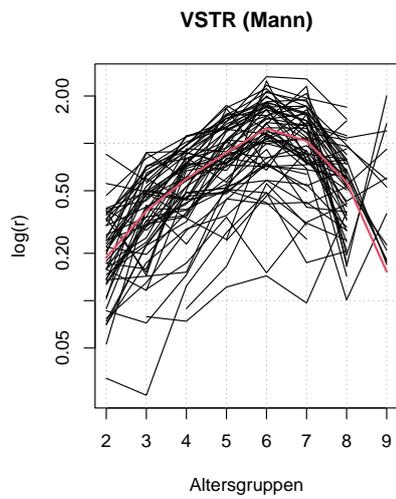


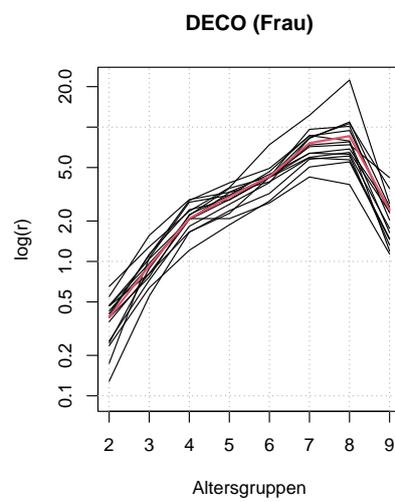
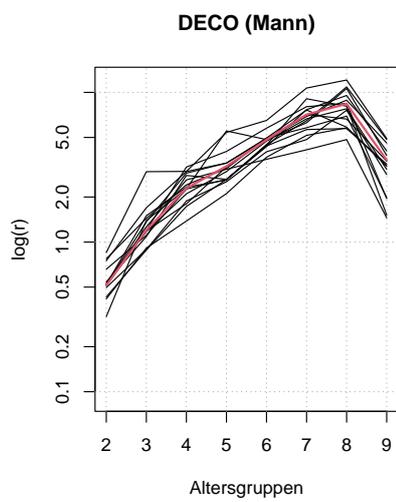
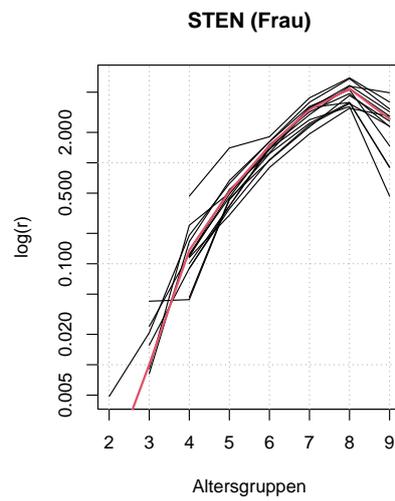
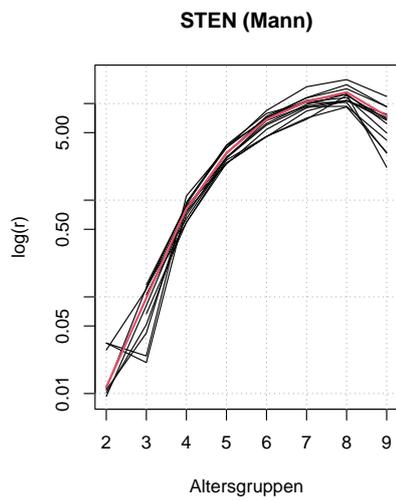
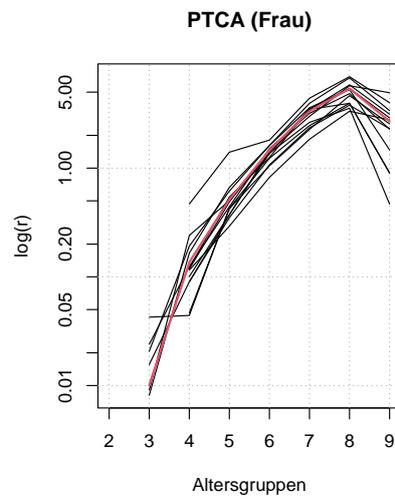
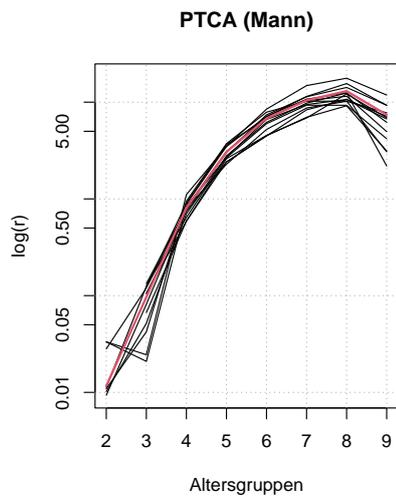
**TURP (Mann)**

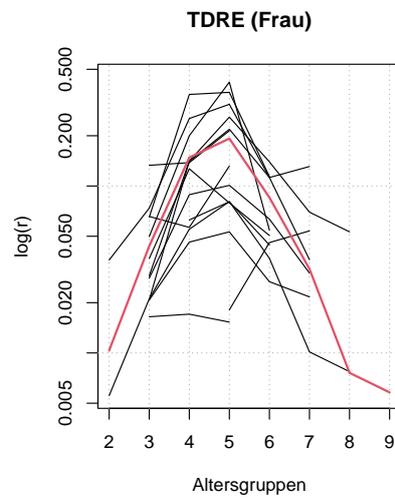
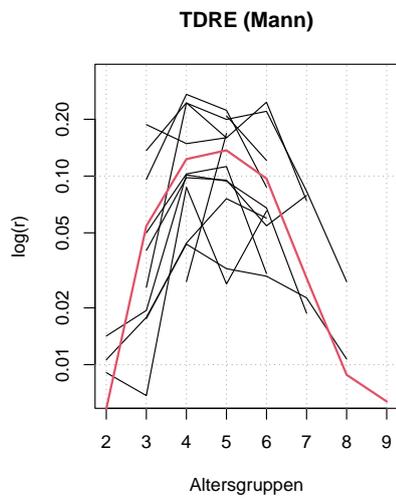
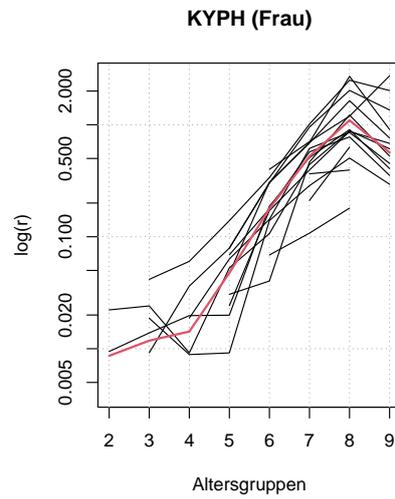
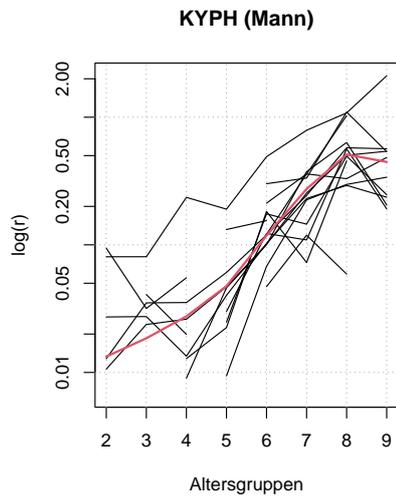
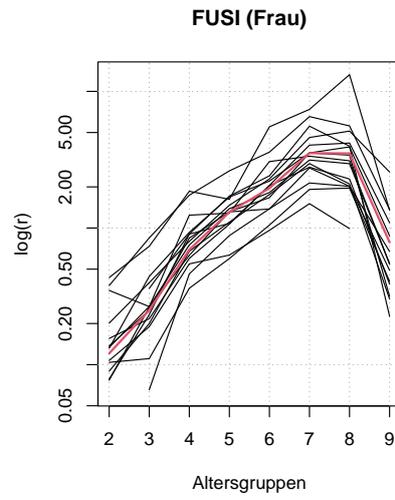
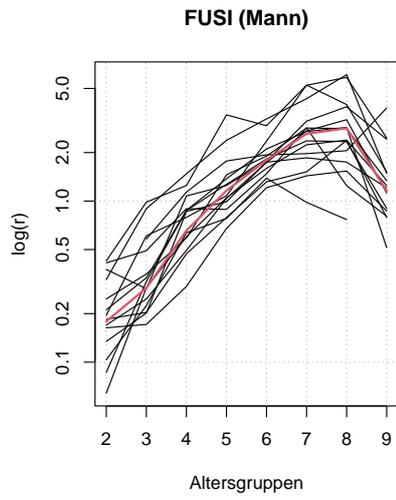


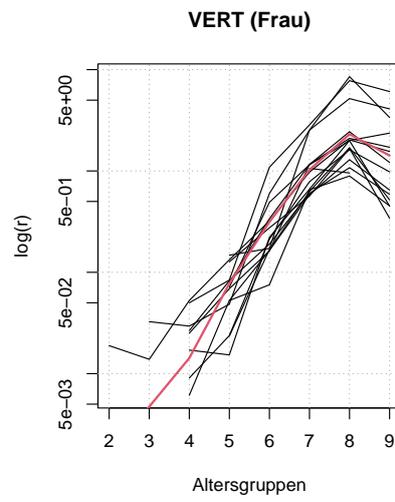
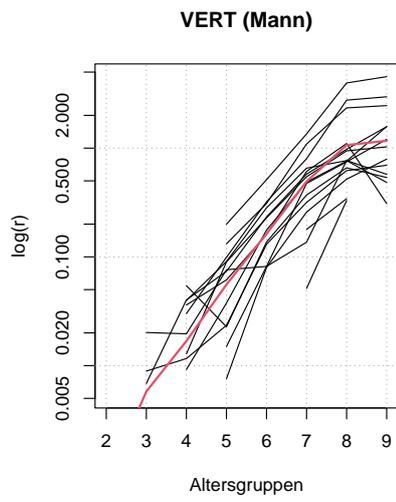
**TURP (Frau)**











## D Verteilung

Die Abbildungen D.1, D.2 und D.3 zeigen für jeden Indikator (mit Ausnahme des Indikators ALLA, der alle Indikatoren zusammenfasst) den Quantil-Quantil-Plot (QQ-Plot) der empirischen Verteilungen von  $(o_i - e_i)/e_i$  gegen die Normalverteilung ( $o_i$  und  $e_i$  sind in Kapitel 2 beschrieben).

Insgesamt werden die empirischen Verteilungen relativ gut durch die Normalverteilung beschrieben. Allerdings zeigen sich bei einigen Indikatoren an den Rändern Abweichungen von der theoretischen Verteilung. Dies ist vor allem bei den folgenden Indikatoren zu beobachten.

- Topologie HSA\_2013\_herz: BYPA und HKAT (Abb. D.1),
- Topologie HSA\_2013\_spine: DECO, FUSI und VERT (Abb. D.2),
- Topologie HSA\_2013: EHBL, EGDO, HTEP, KTEP, MENK, RAFX, ROTM und THUS (Abb. D.3).

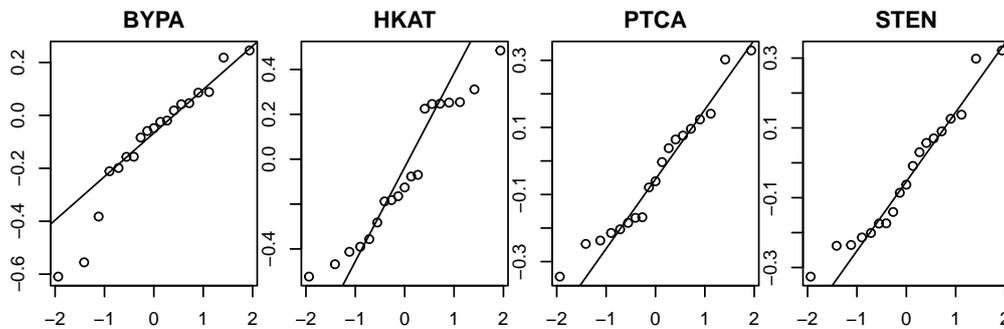


Abbildung D.1 QQ-Plot von  $(o_i - e_i)/e_i$  gegen die theoretische Normalverteilung (Topologie: HSA\_2013\_herz)

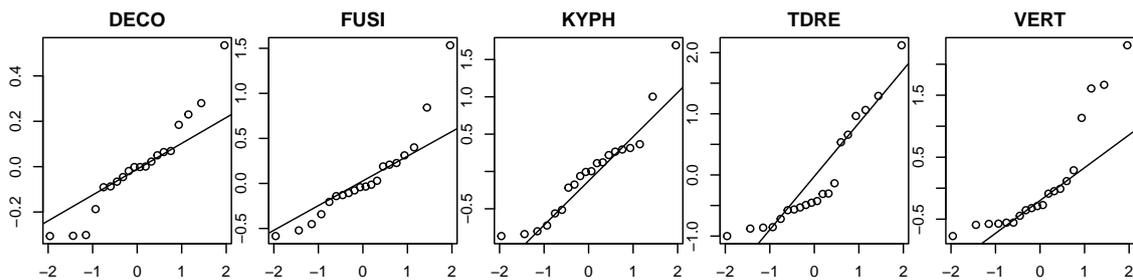


Abbildung D.2 QQ-Plot von  $(o_i - e_i)/e_i$  gegen die theoretische Normalverteilung (Topologie: HSA\_2013\_spine)

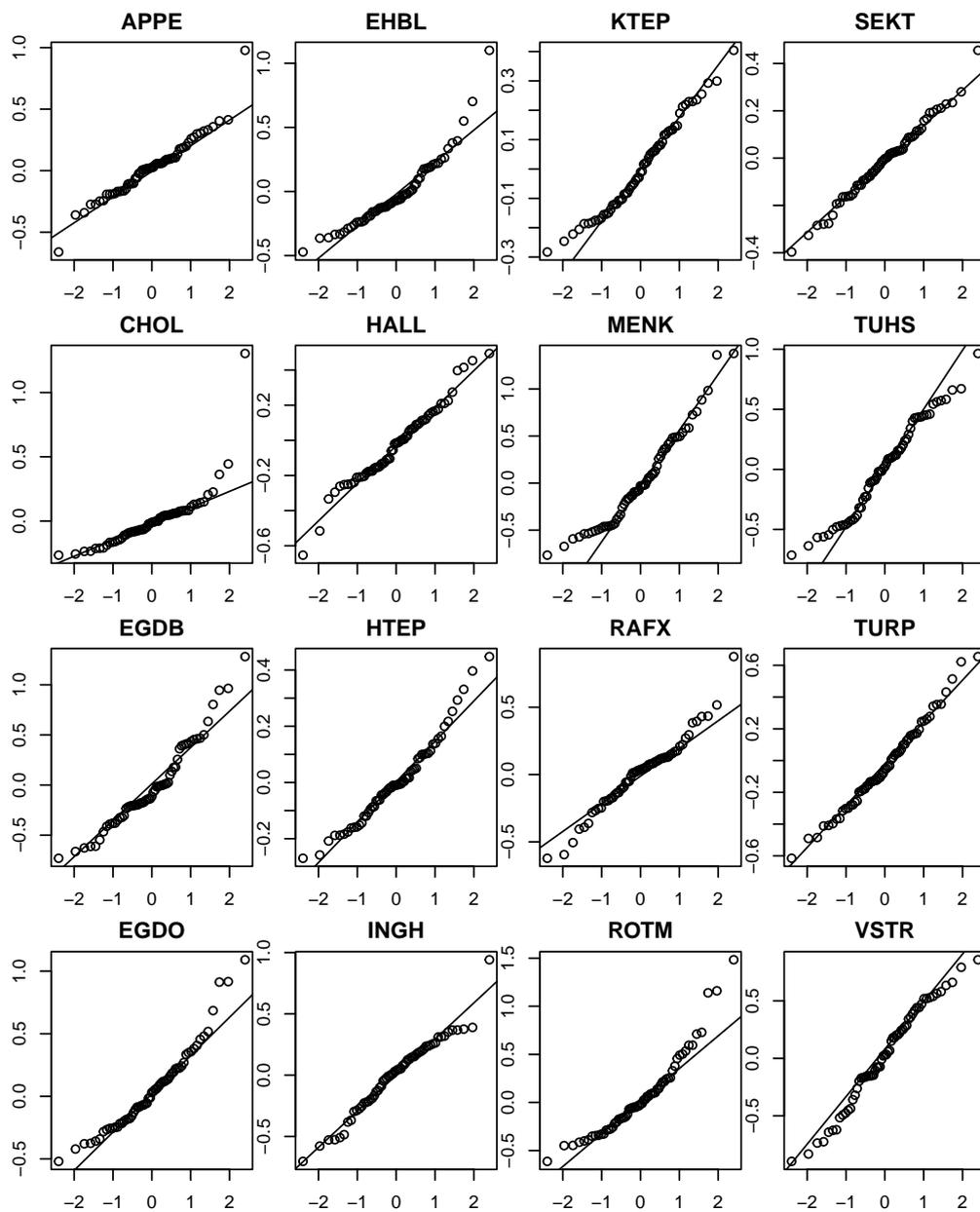


Abbildung D.3 QQ-Plot von  $(o_i - e_i)/e_i$  gegen die theoretische Normalverteilung (Topologie: HSA\_2013)

## E R-Package: sava

Die wichtigsten Funktionen des sava R-Packages sind nachfolgend dokumentiert. Das Code Repository befindet sich auf GitHub und kann unter <https://github.com/tobiasschoch/sava> eingesehen werden.

Dieser Bericht basiert auf der Version 0.4 des Packages. Alle Versionen sind auf GitHub mit einem Tag gekennzeichnet; siehe <https://github.com/tobiasschoch/sava/tags>.

### Installation des Packages

Das Package kann direkt aus GitHub heruntergeladen und installiert werden. Dies sei hier am Beispiel der Version 0.4 (das Argument `ref` bezieht sich auf das Tag "v0.4") illustriert:

```
devtools::install_github("tobiasschoch/sava", ref = "v0.4",
                          build_vignettes = TRUE)
```

### Die wichtigsten Funktionen

In den nachfolgenden Abschnitten besprechen wir den Einsatz der Funktionen im sava Package anhand ausgewählter Beispiele. Im ersten Schritt wird das Package geladen.

```
library("sava")
```

### Direkte Standardisierung und Vertrauensintervalle

Wir verwenden die Daten zum Down-Syndrom (Michigan, 1950–1964), die von [Fay und Feuer \(1997a\)](#) in der Diskussion ihrer Methode eingesetzt wurden. Die beobachteten Zähldaten für die Altersgruppen ("unter 20", 20–24, 25–29, 30–34, 35–39, "40 und über 40" Jahre) sind gegeben durch

```
> oi <- c(0, 8, 63, 112, 262, 295)
```

Die Anzahl der Personen in der Studienpopulation (nach Altersgruppen) sind

```
> ni <- c(327, 30666, 123419, 149919, 104088, 34392)
```

Wir definieren die Anteilswerte  $s_i$  der Altersgruppen an der Referenzpopulation, so dass  $\sum(s_i) = 1$ .

```
> si <- c(0.1133, 0.3296, 0.2784, 0.1728, 0.0842, 0.0217)
```

Die Gewichte (nach Altersgruppen) sind wie folgt definiert

```
> wi <- si / ni
```

Die Schätzung der direkt-standardisierten Rate pro 10'000 (siehe Argument `scale = 10000`) und das 95%-Vertrauensintervall können wie folgt berechnet werden.

```
> DSR(oi, wi, scale = 10000)
```

```
$DSR
```

```
[1] 7.552625
```

```
$conf.int
```

```
[1] 6.770004 18.835611
```

```
attr(,"conf.level")
```

```
[1] 0.95
```

```
$scale
```

```
[1] 10000
```

```
$method
```

```
[1] "Weighted Poisson CI, Fay and Feuer (1997, Stat Med)"
```

Die Konstruktion der Vertrauensintervalls folgt [Fay und Feuer \(1997a\)](#) und basiert auf der Gamma-Verteilung. Wenn nur das Vertrauensintervall benötigt wird, kann die Funktion `DSR()` mit dem Argument `simple = TRUE` aufgerufen werden. Der Rückgabewert ist dann der Vektor mit der unteren und oberen Schranke des Vertrauensintervalls.

```
> DSR(oi, wi, scale = 10000, simple = TRUE)
```

```
[1] 6.770004 18.835611
```

### Systematic component of variation (SCV)

Die Schätzung der SCV wird für Daten zum Indikator TURP (Anzahl der transurethralen Resektionen der Prostata für 61 Spitalversorgungsregionen in der Schweiz, 2014) illustriert. Die ersten sechs Zeilen des Datensatzes sind gegeben durch

```
> data("TURP")
```

```
> head(TURP)
```

```
  HSA      yi      di oi      ei      ni
```

```

1  1  0.03594360 0.016768424 64 61.77942 44119
2  2  0.16609029 0.004183897 325 278.70912 194613
3  3  0.20432349 0.004802633 302 250.76319 174463
4  4 -0.06214409 0.009995156 88 93.83104 69496
5  5  0.61775494 0.018049180 145 89.63039 57359
6  6  0.08737748 0.001525664 775 712.72398 474477

```

wobei  $o_i$  und  $e_i$  die beobachtete bzw. erwartete Anzahl der Eingriffe bezeichnen. Die erwartete Anzahl wurde nach der internal, indirect standardization berechnet. Daraus folgt  $\text{sum}(\text{TURP}\$o_i) = \text{sum}(\text{TURP}\$e_i)$ . Die Variablen  $y_i$  und  $d_i$  sind in diesem Beispiel nicht relevant.

Die Funktion `SCV()` implementiert den Schätzer von [McPherson et al. \(1982\)](#) (Argument `approx = FALSE`) und den leicht modifizierten Schätzer in [Diehr et al. \(1990\)](#) mit dem Argument `approx = TRUE`. Der Schätzer von [Diehr et al. \(1990\)](#) ist unser Standardschätzer für die SCV und wird wie folgt berechnet

```

> s <- SCV(TURP$o_i, TURP$e_i)
> s

```

Systematic component of variation (definition of Diehr et al., 1990)  
Parameter estimate(s):

```

      scv
0.08672

```

Der Rückgabewert von `SCV()` ist ein Objekt/ eine Liste mit den folgenden Slots

```

> names(s)

[1] "method"      "params"      "model"      "converged"  "call"

```

Die Schätzung der SCV kann aus dem Objekt `s` wie folgt extrahiert werden

```

> params(s)

      scv
0.08672256

```

Alle anderen Elemente können ebenfalls mit dem `$`-Operator extrahiert werden.

## Poisson-Gamma EB-Schätzer

[Martuzzi und Hills \(1995\)](#) haben einen EB-Ansatz vorgeschlagen, der auf einem hierarchischen Poisson-Gamma-Modell basiert. Das Modell entspricht demjenigen von [Clayton und Kaldor \(1987\)](#), mit der Ausnahme, dass der Erwartungswert der Gamma-A-priori-Verteilung auf 1.0 festgelegt wird. Die Randverteilung ist die negative Binomialverteilung (NegBin), die mit einem Parameter ( $\alpha$ ) parametrisiert ist. Der Maximum-Likelihood-Schätzer von  $\alpha$  wird wie folgt geschätzt

```
> p <- EBpoisson0(TURP$oi, TURP$ei)
> p
```

```
NegBin MLE estimator (one-parameter)
Parameter estimate(s):
  alpha
10.51737
```

```
Algorithm converged in 16 iterations
```

Die Schätzung für  $\alpha$  kann aus dem Objekt  $p$  wie folgt extrahiert werden

```
> params(p)

  alpha
10.51737
```

Die geschätzte Varianz der A-priori-Verteilung wird als Kehrwert des geschätzten Parameters berechnet.

```
> 1 / params(p)

  alpha
0.09508078
```

Alle anderen Elemente des Objekts  $p$  können ebenfalls mit dem  $\$$ -Operator extrahiert werden.

Anstelle der Crude Rates  $o_i/e_i$  werden die nach der EB-Methode vorhergesagten Raten verwendet. Diese Raten berechnen sich folgendermassen

```
> head(predict(p))
```

```
[1] 1.0307147 1.1600506 1.1960988 0.9441195 1.5528792 1.0861068
```

Die statistische Inferenz für die HSA-spezifischen Raten basiert auf dem geschätzten mittleren quadratischen (Vorhersage-)Fehler (mean square error, MSE). Die Funktion `mse()` implementiert zwei Jackknife-Schätzer für den MSE:

- Methode = "rao" siehe [Rao und Molina \(2015, Kapitel 9.5.1\)](#)
- Methode = "jiang" siehe [Jiang et al. \(2002\)](#)

Die Methode "rao" ist die Standardmethode. Der geschätzte MSE (für die Methode "rao") wird wie folgt berechnet

```
> head(mse(p))
```

```
[1] 0.0134974800 0.0033562406 0.0040232871 0.0082799696 0.0169018269  
[6] 0.0006886123
```

[Coory und Gibberd \(1998\)](#) schlugen vor, das Quintile ratio der vorhergesagten Raten zu berechnen. Das Verhältnis QR(80/20) erhält man mit dem Befehl

```
> QR(p)
```

```
[1] 1.694126
```

In den nachfolgenden Kapiteln sind die wichtigsten Funktionen des `sava` Packages dokumentiert.

## E.1 DSR: Direct Standardized Rates and Confidence Intervals

Direct standardized rates are computed given the observed count data and the weights (by stratum). Confidence interval estimation is based on the method of Fay and Feuer (1997).

### Usage

```
DSR(oi, wi, conf.level = 0.95, scale = 1, simple = FALSE, empty = TRUE)
```

## Arguments

`oi` [numeric vector] observed counts (by stratum).

`wi` [numeric vector] weights (by stratum).

`conf.level` [numeric] confidence level of the interval (default: 0.95).

`scale` [numeric] scale factor, choose e.g., `scale = 1000` to compute rates per 1,000 individuals (default: 1).

`simple` [logical] if TRUE, the return value is the confidence interval, i.e., a vector of size 2; if FALSE, a list is returned (default: FALSE).

`empty` [logical] if TRUE, the confidence interval for vectors with not observations is  $c(0, \text{Inf})$ ; if FALSE, the confidence interval is  $c(\text{NA}, \text{NA})$  (default: TRUE).

## Details

The direct standardized rate (DSR) is computed as the weighted sum of the  $oi$  with weights  $wi$ , which is maximum-likelihood estimator. The confidence intervals are computed under the assumption that the rate can be modeled as weighted sum of independent Poisson random variables. The construction has been proposed by Fay and Feuer (1997). The simulation study of Ng et al. (2008) showed that this construction leads to conservative confidence intervals.

The validity of the confidence intervals relies on the Poisson assumption. If the distribution of the count data shows significant overdispersion, the computed confidence intervals tend to be too short.

## Value

If `simple = FALSE`, the return value is the confidence interval (a numeric vector of size 2), otherwise a list with the following slots is returned

- `DSR` direct standardized rate
- `conf.int` confidence interval, vector (with attributes)
- `scale` scaling factor
- `method` description of method

## Examples

```
# Observed count data (by age group)
oi <- c(0, 8, 63, 112, 262, 295)

# Number of individuals (by age group) in the study population
ni <- c(327, 30666, 123419, 149919, 104088, 34392)

# Shares of the age groups in the standard population (e.g., European
# Standard Population). The shares sum up to 1.
si <- c(0.1133, 0.3296, 0.2784, 0.1728, 0.0842, 0.0217)

# Weights (by age group)
wi <- si / ni

# Direct standardized rate and 95% confidence interval (default)
DSR(oi, wi)
```

## E.2 EBnormal0: Normal–Normal Empirical Bayes Estimator (With Fixed Location)

Empirical Bayes (EB) estimator of the variance in a Normal–Normal mixture (with a priori location fixed at the origin) and the generalized James–Stein estimator/predictor of the area-specific rates

### Usage

```
EBnormal0(oi, ei, maxit = 100, tol = 1e-5)
```

```
## S3 method for class 'genjs0'
predict(object, ...)
## S3 method for class 'genjs0'
mse(object, method = "analytic", ...)
```

### Arguments

oi [numeric vector] area-specific observed counts.

ei [numeric vector] area-specific expected counts.

maxit [integer] maximum number of iterations to use (default: 100).

`tol` [double] numerical tolerance criterion to stop the iterations (default:  $1e-05$ ).

`object` an object of class `sava_scv`.

`method` [character] type of mse estimator, "analytic" or "jackknife" (default: "analytic").

... additional arguments.

## Details

**Workflow** The workflow is as follows: 1) estimate the parameters by `EBnormal0()`, 2) predict the rates by `predict()`, and 3) compute the mean square prediction error using `mse()`

**Parameter estimation** Consider the following 2-stage conditionally independent hierarchical Normal–Normal Bayesian model, where the first-stage model is given by

$$Y_i | \Theta_i = \theta_i \sim N(\theta_i, D_i)$$

with  $Y_i = (o_i - e_i)/e_i$ , where  $o_i$  and  $e_i$  denote, respectively, the observed and expected number of counts in the areas  $i = 1, \dots, n$ . It is assumed that the  $e_i$ 's are computed by internal, indirect standardization (see e.g., Fleiss et al., 2003, Chapter 19). The variances  $D_i$ 's are known quantities, which are defined as  $D_i = o_i/e_i^2$  (Shwartz et al., 1994). The a priori distribution is

$$\Theta_i \sim N(0, A)$$

where the variance  $A$  is unknown; see also Efron and Morris (1975).

`EBnormal0()` estimates  $A$  by maximum likelihood. The estimate of  $A$  is set to NA if the estimate is negative or when the algorithm does not converge.

*Note.* The a priori location is zero. Hence, the EB prediction rule shrinks towards the origin. This is reflected in the name of `EBnormal0` by the suffix  $0$ .

**Prediction** The EB predictor of the rates (which is a generalized James–Stein estimator about the origin, see Efron and Morris, 1975) is computed by `predict()` based on the object estimated by `EBnormal0()`. The implemented predictor applies the bias correction proposed in Morris (1983).

**Uncertainty – Mean square prediction error estimation** The mean square prediction error can be computed by 1) the Prasad and Rao (1990) analytic approximation (set the argument `method = "analytic"`) or 2) the jackknife estimator of Jiang et al. (2002), see `method = "jackknife"`. It is implemented by `mse()`, which takes as an argument an object of class `sava`, i.e., output of `EBnormal0()`.

## Value

`EBnormal0()` returns an object of class `sava` (and `genjs0`). The return value of, respectively, `predict()` and `mse()` are the predicted rates and area-specific estimates of the mean square prediction error.

## Examples

```
data("TURP")

# estimate
est <- EBnormal0(TURP$oi, TURP$ei)
est

# predict rates
predict(est)

# estimates of the mean square prediction error
mse(est)
```

## E.3 EBpoisson0: Poisson–Gamma Empirical Bayes Estimator of Variance (With Fixed Variance)

Empirical Bayes (EB) estimator of the variance in a Poisson–Gamma mixture (with a priori expectation equal to one) and estimator/predictor of the area-specific rates.

### Usage

```
EBpoisson0(oi, ei, interval = c(1e-5, 1e4), maxit = 1000)

## S3 method for class 'negbin0'
predict(object, ...)
## S3 method for class 'negbin0'
mse(object, type = c("rao", "jiang"), ...)
```

### Arguments

`oi` [numeric vector] observed counts.

`ei` [numeric vector] expected counts.

`interval` `numeric` vector of size `to` that defines the boundaries of the interval within which a root is searched.

`maxit` [`integer`] maximum number of iterations to use (default: 1000).

`type` [`character`] type of jackknife estimator, either `type = "rao"` or `type = "jiang"` (default: "rao").

`object` an object of class `sava_scv`.

... additional arguments.

## Details

**Workflow** The workflow is as follows: 1) estimate the parameters by `EBpoisson0()`, 2) predict the rates by `predict()`, and 3) compute the mean square prediction error using `mse()`

**Parameter estimation** Consider the following 2-stage conditionally independent hierarchical Poisson–Gamma Bayesian model, where the first-stage model is given by

$$Y_i \mid \Lambda_i = \lambda_i \sim \text{Pois}(e_i \lambda_i)$$

with  $Y_i = (o_i - e_i)/e_i$ , where  $o_i$  and  $e_i$  denote, respectively, the observed and expected number of counts in the areas  $i = 1, \dots, n$ . It is assumed that the  $e_i$ 's are computed by internal, indirect standardization (see e.g., Fleiss et al., 2003, Chapter 19). The a priori distribution is

$$\Lambda_i \sim \text{Gamma}(\alpha, \alpha),$$

where  $\alpha > 0$  is the rate parameter of the Gamma distribution. This parametrization implies that the  $\Lambda_i$ 's have expected value 1. The marginal distribution of the  $Y_i$ 's is negative binomial in only the parameter  $\alpha$  (Martuzzi and Hills, 1995).

`EBpoisson0()` estimates  $\alpha$  by maximum likelihood. The estimate of  $\alpha$  is set to NA if the estimate is negative or when the algorithm does not converge.

**Prediction** The EB predictor of the rates (under the above model) is computed by `predict()` based on the object estimated by `EBpoisson0()`.

**Uncertainty – Mean square prediction error estimation** The mean square prediction error is estimated by the jackknife estimator. Two implementations are available: "rao": Rao (2003, Chapter 9.5.1) or "jiang": Jiang et al. (2002). It is implemented by `mse()`, which takes as an argument an object of class `sava`, i.e., output of `EBpoisson0()`.

## Value

`EBpoisson0()` returns an object of class `sava` (and `negbin0`). The return value of, respectively, `predict()` and `mse()` are the predicted rates and area-specific estimates of the mean square prediction error.

## Examples

```
data("TURP")

# estimate
est <- EBpoisson0(TURP$oi, TURP$ei)
est

# predict rates
predict(est)

# estimate mean square prediction error (type = "rao")
mse(est)
```

## E.4 kendall: Kendall Coefficient of Concordance (Kendall's W)

### Usage

The Kendall coefficient of concordance, also known as Kendall's W, is a rank-based measure of correlation

```
kendall(x, na.rm = FALSE)
```

### Arguments

`x` ratings, [matrix] of dimension  $(n, p)$ , where the  $p$  columns contain the ratings of the  $p$  observers who rated the  $n$  objects.

`na.rm` [logical] indicating whether NA values should be removed before the computation proceeds (default: FALSE).

## Details

The Kendall coefficient of concordance can be used for assessing the agreement among observers who rated the same objects (Kendall and Gibbons, 1990, Chapter 6). The ratings need not be whole numbers. The observer-specific ratings (columns of matrix  $x$ ) are ranked internally before computing the coefficient.

The Kendall coefficient of concordance ranges from 0 (no agreement) to 1 (complete agreement).

## Value

Kendall coefficient of concordance (numeric)

## Examples

```
# Data without ties from Kendall and Gibbons (1990, Table 6.1) on
# 4 observers who rated 6 objects
x <- structure(c(5, 4, 1, 6, 3, 2, 2, 3, 1, 5, 6, 4, 4, 1, 6, 3, 2,
  5, 4, 3, 2, 5, 1, 6), dim = c(6L, 4L))
x
kendall(x)
```

```
# Data with ties from Kendall and Gibbons (1990, Table 6.2) on
# 3 observers who rated 10 objects
x <- structure(c(1, 4.5, 2, 4.5, 3, 7.5, 6, 9, 7.5, 10, 2.5, 1, 2.5,
  4.5, 4.5, 8, 9, 6.5, 10, 6.5, 2, 1, 4.5, 4.5, 4.5, 4.5, 8, 8, 8,
  10), dim = c(10L, 3L))
x
kendall(x)
```

## E.5 SCV: Systematic Component of Variation

Estimator of the systematic component of variation (SCV) of McPherson et al. (1982)

### Usage

```
SCV(oi, ei, approx = TRUE)
```

```
# Utility functions
## S3 method for class 'sava_scv'
predict(object, ...)
```

## Arguments

`oi` [numeric vector] observed counts.

`ei` [numeric vector] expected counts.

`approx` [logical] if TRUE, then  $\log(o[i] / e[i])$  is approximated by  $(o[i] - e[i]) / e[i]$  (default: FALSE); see Sect. details.

`object` an object of class `sava_scv`.

... additional arguments.

## Details

McPherson et al. (1982, p. 1314) define  $y_i = \log(o_i/e_i)$ , where  $o_i$  and  $e_i$  denote, respectively, the number of observed and expected counts. The number of expected cases must be strictly larger than zero. The SCV is then computed based on  $y_i^2$ . The above definition of  $y_i$  can be problematic if zero cases were observed in one or more areas because  $y_i$  will then be negative infinity. As a consequence, SCV will be negative infinity as well. Diehr et al. (1990, p. 754) define instead  $y_i = (o_i - e_i)/e_i$ , which is an approximation that does not suffer from the issue. This definition will be used if `SCV()` is called with argument `approx = TRUE`.

The SCV can (in principle) be negative if the sum of the  $y_i^2$  is smaller than the sum of the  $1/e_i$ . If this happens, `SCV()` returns NA because the SCV as a measure of variation is not well defined in this case.

## Value

An object of class `sava_scv`

## Examples

```
data("TURP")

# estimate SCV
est <- SCV(TURP$oi, TURP$ei)
est

# predict rates
predict(est)
```

## E.6 Utility Functions for Class sava

Utility functions for instances of class sava

### Usage

```
params(object, ...)  
mse(object, ...)  
jackknife(object, ...)  
EQ(object, q = 0.2)
```

```
# Utility functions  
## S3 method for class 'sava'  
print(x, digits = max(1L, getOption("digits") - 2L), ...)  
## S3 method for class 'sava'  
params(object, ...)
```

### Arguments

x an object of class sava.  
object an object of class sava.  
digits [integer] minimal number of significant digits.  
q [numeric] quantile  
... additional arguments.

### Details

Function `params()` extracts the estimated parameters; the functions `mse()` and `jackknife()` compute, respectively, the mean square prediction error and the jackknife variance of the (variance) parameter.

### Examples

```
data("TURP")  
  
# estimate SCV and extract estimated parameters  
params(SCV(TURP$oi, TURP$ei))  
  
# estimate Poisson-gamma model  
est <- EBpoisson0(TURP$oi, TURP$ei)
```

```
# compute extremal quotient  
EQ(est)
```