Bachelor-Thesis

Anonymisierungsverfahren: Randverteilungen und ihr statistisches Analysepotential

Autor: Christian Kluge

Matrikelnummer: xxxxxxx

Betreuer: Prof. Dr. Augustin

Ludwig-Maximilians-Universität

München, den 23.08.2010

Kurzfassung

Gegenstand dieser Bachelor-Thesis ist die Untersuchung des statistischen Analysepotentials von Randverteilungen eingebettet in das Thema „Anonymisierungsverfahren“.

Dazu werden im ersten Teil die relevanten theoretischen Grundlagen der Kontingenztafel, der linearen Optimierung und der Fréchet Bounds erläutert. Anschließend werden mit diesen Grundlagen beispielhaft aus gegebenen Randverteilungen die Intervalle der Häufigkeitsverteilung einer dreidimensionalen Kontingenztafel berechnet. Danach wird diskutiert, welche Eigenschaften und Unterschiede die Ergebnisse aufweisen.

Im zweiten Teil wird die im ersten Teil erläuterte Theorie an einem medizinischen Kosten-Szenario für Behandlungen von Diabetes-Erkrankungen erneut angewendet. Die resultierenden Intervalle der Behandlungskosten werden anschließend interpretiert und hinterfragt.

Der dritte Teil beinhaltet eine weitere Erläuterung der für die statistische Auswertung relevanten Grundlagen und eine beispielhafte Anwendung, die das Szenario des ersten Teils weiterführt. Danach folgt eine Erörterung, ob und inwiefern Ergebnisse ermittelt werden konnten und wie diese beschaffen sind.

Zum Abschluss wird ein Ausblick gegeben, wie der Ansatz, dass nur Randverteilungen zu Analysezwecken zur Verfügung stehen, erweitert oder in anderer Form angewendet werden kann.

Schlagwörter: dreidimensionale Kontingenztafel, lineare Optimierung, Fréchet Bounds, 1- bzw. 2-dimensionale Randverteilung, Häufigkeitsverteilung, absolute Häufigkeit, Aggregation, Anonymisierung

Abstract

Subject of this bachelor's thesis is to investigate the statistical analysis of potential of marginal distributions embedded in the topic „anonymization processes".

For this purpose in the first section are explained the relevant theoretical foundations of contingency tables, linear optimization and the Fréchet Bounds. Thereupon the intervals of frequency distribution of three-dimensional contingency table are exemplarily calculated with given marginal distributions. Afterwards is discussed which characteristics and differences the results show.

In the second section the already explained theory is reapplied to a medical costs-scenario for treatment of diabetes-diseases. Subsequently the resulting intervals of the treatment expenses are interpreted and analyzed.

The third section contains a further explanation for the statistical analysis of relevant principles and an exemplary application which continues the scenario of the first section. Afterwards follows a discussion of whether and to what extent results could be calculated and whereby they are characterized.

As conclusion is given a forecast of how the approach that only marginal distributions are available for analysis can be extended or applied in other forms.

Keywords: three-dimensional contingency table, linear programming, Fréchet Bounds, 1- or 2-dimensional marginal distribution, frequency distribution, absolute frequency, aggregation, anonymization

Danksagung

Die Entstehung dieser Bachelor-Thesis ist der Unterstützung vieler Menschen, die Ratschläge und neue Ideen eingebracht oder den Text und die Inhalte korrigiert haben, zu verdanken.

Herr Prof. Dr. Thomas Augustin, Frau Andrea Wiencierz und Herr Marco Cattaneo vom Institut für Statistik an der Ludwig-Maximilians-Universität München verdienen ganz besonderen Dank, da sie die Bachelor-Thesis seitens der Ludwig-Maximilians-Universität betreuten und zu jeder Zeit ein offenes Ohr für Diskussionen und Fragen hatten. Das Gelingen dieser Bachelor-Thesis gründet sich maßgeblich auf der immer währenden Unterstützung in allen Belangen der Bachelor-Thesis.

Nicht zu vergessen ist die Unterstützung meiner Eltern, meines Bruders und meiner Freunde, die für mich über drei Monate und durch das gesamte Studium hindurch durch Ihre Toleranz und Ihr Verständnis immer eine große Stütze und Motivation waren.

Deshalb widme ich diese Arbeit, als Abschluss meines Studiums, meinen Eltern, Frau J. Kluge und Herr Dr. med. R. Kluge, die mir dieses Studium ermöglicht haben.

München, den 23.08.2010 Christian Kluge

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

In unserer heutigen Gesellschaft spielt das Thema Datenschutz eine immer größere Rolle. Auf der einen Seite wollen wir alle immer mehr Informationen und auf der anderen Seite möchten wir nicht alles von unseren privaten Daten preisgeben bzw. dass Dritte unsere Daten weitergeben. Das Thema Datenschutz hat damit auch in der Wissenschaft, im Gesundheitswesen sowie in der empirischen Wirtschaftsforschung und Politikberatung sehr an Bedeutung gewonnen. Gerade in diesen Bereichen ist aber die Nachfrage nach Mikrodaten in den letzten Jahren gestiegen. Nur mit Mikrodaten können Verhaltensweisen einzelner Unternehmen und Betriebe sowie einzelner Haushalte und Personen beobachtet, Zusammenhänge und Individualeffekte erklärt werden. Durch Aggregation gehen solche Informationen oft verloren oder sind gar nicht erst möglich. [Rosemann, 2006]

Um der Wissenschaft oder auch freien Anwendern Daten in geeigneter Form weitergeben zu können, benötigt man Anonymisierungsverfahren. Diese sollen den Schutz vor der Reidentifizierung der einzelnen Personen oder der einzelnen Unternehmen, deren Daten erhoben wurden, gewährleisten. Darüber hinaus sollen sie die statistische Analyse der Daten und vor allem bestmögliche und sinnvolle Ergebnisse weiterhin ermöglichen. Nachfolgend werden die Begriffe *Anonymisierung* und *Pseudonymisierung* definiert:

Definition Anonymisierung und Pseudonymisierung, Bundesdatenschutzgesetz 1990, § 3 Abs. 6, 6(a), [Bundesministerium der Justiz, 2009]

§ 3 (6) Anonymisieren ist das Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Personen zugeordnet werden können.

(6a) Pseudonymisierung ist das Ersetzen des Namens und anderer Identifikationsmerkmale durch ein Kennzeichen zu dem Zweck, die Bestimmung des Betroffenen auszuschließen oder wesentlich zu erschweren.

Anonymisierung beschreibt also die tatsächliche Veränderung von personenbezogenen Daten, so dass diese Daten nicht mehr einer Person zugeordnet werden können. Pseudonymisierung

stellt lediglich das Ersetzen von Identifikationsmerkmalen durch ein Pseudonym (auch Code genannt) dar, um die Identifizierung des Betroffenen zu verhindern. In Folgedessen bleiben bei der Pseudonymisierung Bezüge verschiedener Datensätze, die auf dieselbe Art pseudonymisiert wurden, erhalten. Das bedeutet, dass die Pseudonymisierung, unter Zuhilfenahme eines Schlüssels, die Zuordnung von Daten zu einer Person oder zu einem Unternehmen ermöglicht.

Im Bundesstatistikgesetz wird ebenfalls der Umgang mit statistischen Einzeldaten vorgegeben. Dort wird auch die unterschiedliche Intensität der Anonymisierung berücksichtigt.

Anonymität von Einzeldaten, Bundesstatistikgesetz - BStatG 1987, § 16 Geheimhaltung, Abs. 1, 6, [Statistisches Bundesamt, 2008]

§ 16 (1) Einzelangaben über persönliche und sachliche Verhältnisse, die für eine Bundesstatistik gemacht werden, sind von den Amtsträgern und für den öffentlichen Dienst besonders Verpflichteten, die mit der Durchführung von Bundesstatistiken betraut sind, geheimzuhalten, soweit durch besondere Rechtsvorschrift nichts anderes bestimmt ist. Dies gilt nicht für

(...)

4. Einzelangaben, wenn sie dem Befragten oder Betroffenen nicht zuzuordnen sind.

(...)

(6) Für die Durchführung wissenschaftlicher Vorhaben dürfen vom Statistischen Bundesamt und den statistischen Ämtern der Länder Einzelangaben an Hochschulen oder sonstige Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung übermittelt werden, wenn die Einzelangaben nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können und die Empfänger Amtsträger, für den öffentlichen Dienst besonders Verpflichtete oder Verpflichtete nach Absatz 7 sind.

Demnach beinhaltet der § 16 des BStatG die grundsätzliche Geheimhaltung von Einzeldaten bis auf gewisse Ausnahmen. Die Ausnahmen bedingen sich unter anderem durch den Verwendungszweck. Die Einzeldaten, die den Hochschulen oder der wissenschaftlichen Forschung zur Verfügung gestellt werden, sind zum Beispiel nicht *absolut* sondern „nur“ *faktisch* anonymisiert. Man unterscheidet zwischen verschiedenen Intensitätsstufen der Anonymisierung. Das Statistische Bundesamt deklariert eine Einteilung von drei bis vier Stufen.

Nachfolgend werden die drei wichtigsten Stärkegrade erläutert.

Absolute Anonymisierung wird durch Vergröberung oder Entfernung einzelner Merkmale (Aggregation) erreicht. So sollen die Daten so weit verändert werden, dass eine Reidentifizierung der Auskunftgebenden unmöglich wird. Die absolut anonymisierten Mikrodaten werden von der amtlichen Statistik in Form von *Public Use Files* (PUF) angeboten. Solche Datensätze werden von den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder allen interessierten Personen zur Verfügung gestellt.

Faktische Anonymisierung bedeutet, dass die Deanonymisierung der Mikrodaten zwar nicht gänzlich ausgeschlossen werden kann, allerdings die Angaben nur mit *unverhältnismäßig hohem Aufwand an Zeit, Kosten und Arbeitskraft* dem jeweiligen Merkmalsträger zugeordnet werden können. Faktisch anonymisierte Daten dürfen jedoch nur wissenschaftlichen Einrichtungen und hierbei nur zur Durchführung wissenschaftlicher Vorhaben zugänglich gemacht werden. Das Hauptziel dieser eingeschränkten Anonymisierung ist es, die Zuordnungsmöglichkeiten von Merkmalsausprägungen zu den entsprechenden Merkmalsträgern durch vorsichtige Informationsreduktion und -veränderungen zu verringern, aber dennoch den statistischen Informationsgehalt zu erhalten. Die faktisch anonymisierten Datensätze werden als sogenannte *Scientific Use Files* (SUF) zur Verfügung gestellt.

Formal anonymisierte Mikrodaten bieten die Forschungsdatenzentren Datennutzern im Rahmen der kontrollierten Datenfernverarbeitung an. Der Grad der formalen Anonymisierung ist geringer als bei der faktischen Anonymisierung und wird vor allem dann benötigt, wenn z.B. regional und fachlich tief gegliederte Auswertungen vorgenommen werden sollen. Hierfür würden die faktisch anonymisierten Daten nicht ausreichend Information bereitstellen. Zum anderen eignen sich manche Mikrodaten der amtlichen Statistik nicht zur faktischen Anonymisierung oder sind nur bedingt anonymisierbar. Die formale Anonymisierung zeichnet sich durch die Entfernung der direkten Identifikatoren aus. Der Merkmalsumfang und die fachlichen und regionalen Gliederungen bleiben erhalten. Erst die Ergebnisse werden einer Überprüfung auf geheim zu haltende Fälle unterzogen, bevor diese an die Datennutzer wieder übermittelt werden. [FDZ, 2007], [Schoffer, 2008, S. 2]

Die folgende Abbildung 1.1 visualisiert den Zusammenhang zwischen dem Grad der Anonymisierung und dem Grad des Analysepotentials.

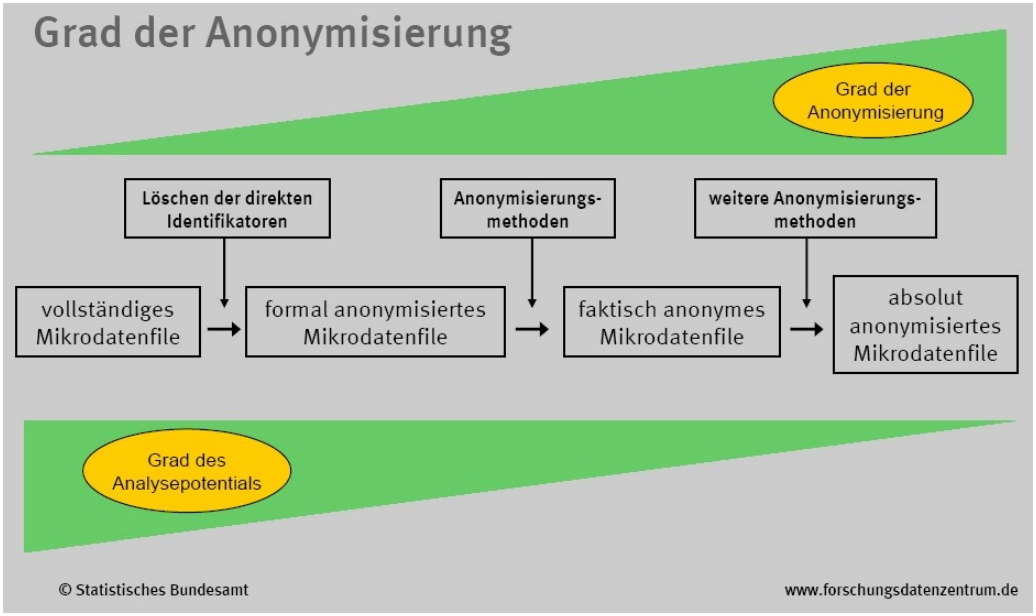

Abbildung 1.1: Grad der Anonymisierung, [FDZ, 2009, S. 5]

Es ist ersichtlich, dass eine stärkere Anonymisierung zu einem geringeren statistischen Analysepotential führt. Der Informationsverlust beeinträchtigt die Aussagekraft und die Güte der Ergebnisse aus den Analysen. Dadurch kann die Realität nicht mehr genau abgebildet werden.

Bei den vorangehenden Erklärungen und gesetzlichen Regelungen ist vor allem zu beachten, dass es wesentliche Unterschiede zwischen personen- und unternehmensbezogenen Mikrodaten gibt. Unternehmen sind Teil wesentlich kleinerer Grundgesamtheiten als Personen und haben einzelne Merkmalsträger mit eventuell sehr extremen Merkmalsausprägungen. Durch die kleineren Grundgesamtheiten sind starke Zusammenhänge zwischen einzelnen Erhebungen (eventuell wurden dieselben Unternehmen öfter betrachtet) zu erwarten, da Unternehmen eventuell gleichzeitig in unterschiedlichen Erhebungen vorzufinden sind. Des Weiteren steht im Bereich der Wirtschaft wesentlich mehr Zusatzwissen zur Verfügung. Die Deanonymisierung wird in den meisten Fällen im wirtschaftlichen Bereich einen deutlich höheren Nutzen, z.B. durch unternehmeninternes Wissen und finanzielle Vorteile, haben als die Deanonymisierung von personenbezogenen Daten. [FDZ, 2009, S. 23]

Die vielen bereits entwickelten Anonymisierungsverfahren müssen je nach Zielsetzung der Anonymisierung sensibel ausgewählt und oft auch sinnvoll kombiniert werden. Es existieren die unterschiedlichsten Unterteilungen in Grafiken, welche die Übersicht der gängigsten Anonymisierungsverfahren darstellen. Allerdings gleichen sich natürlich die dort aufgeführten Verfahren. Anhand zweier Grafiken des Statistischen Bundesamtes wurde von mir deshalb eine zusammenfassende Übersicht erstellt, siehe Abbildung 1.2 Seite 5. [Höhne, 2003, S. 73], [Ronning et al., 2005, S. 98]

Um eine sinnvolle Verfahrensauswahl zu treffen, ist es hilfreich auf verschiedene Kriterien zu achten. Beim Statistischen Bundesamt sind im Forum der Bundesstatistik, Bd. 42/2003 folgende Kriterien genannt:

- Leichte Handhabbarkeit des Verfahrens
  Da die Anonymisierungsverfahren später durch das Personal und die technischen Möglichkeiten durchgeführt werden müssen, ist eine leichte Handhabbarkeit notwendig.

- Erfolgsaussichten des Verfahrens
  Hier muss teilweise auf Verfahrensbewertungen in der Literatur zurückgegriffen werden. Über einige Verfahren sind Nachteile bekannt, die bereits gelöst wurden. Andere Verfahren existieren erst in der Theorie und sind daher noch nicht praktisch einsetzbar.

- Repräsentative Vertretung der Verfahrensgruppen
  In der Verfahrensauswahl sollten möglichst alle Verfahrensgruppen vertreten sein. Da jede Verfahrensgruppe einen anderen Ansatz der Anonymisierung repräsentiert, sollte kein genereller Ausschluss einzelner Verfahrensgruppen vorgenommen werden.

- Abhängigkeit zwischen Verfahren
  Oft ist eine wirkungsvolle Anonymisierung nur durch Verwendung mehrerer Verfahren möglich. Hier sind vor allem untereinander abhängige Verfahren zu beachten.

[Höhne, 2003, S. 73, 74]

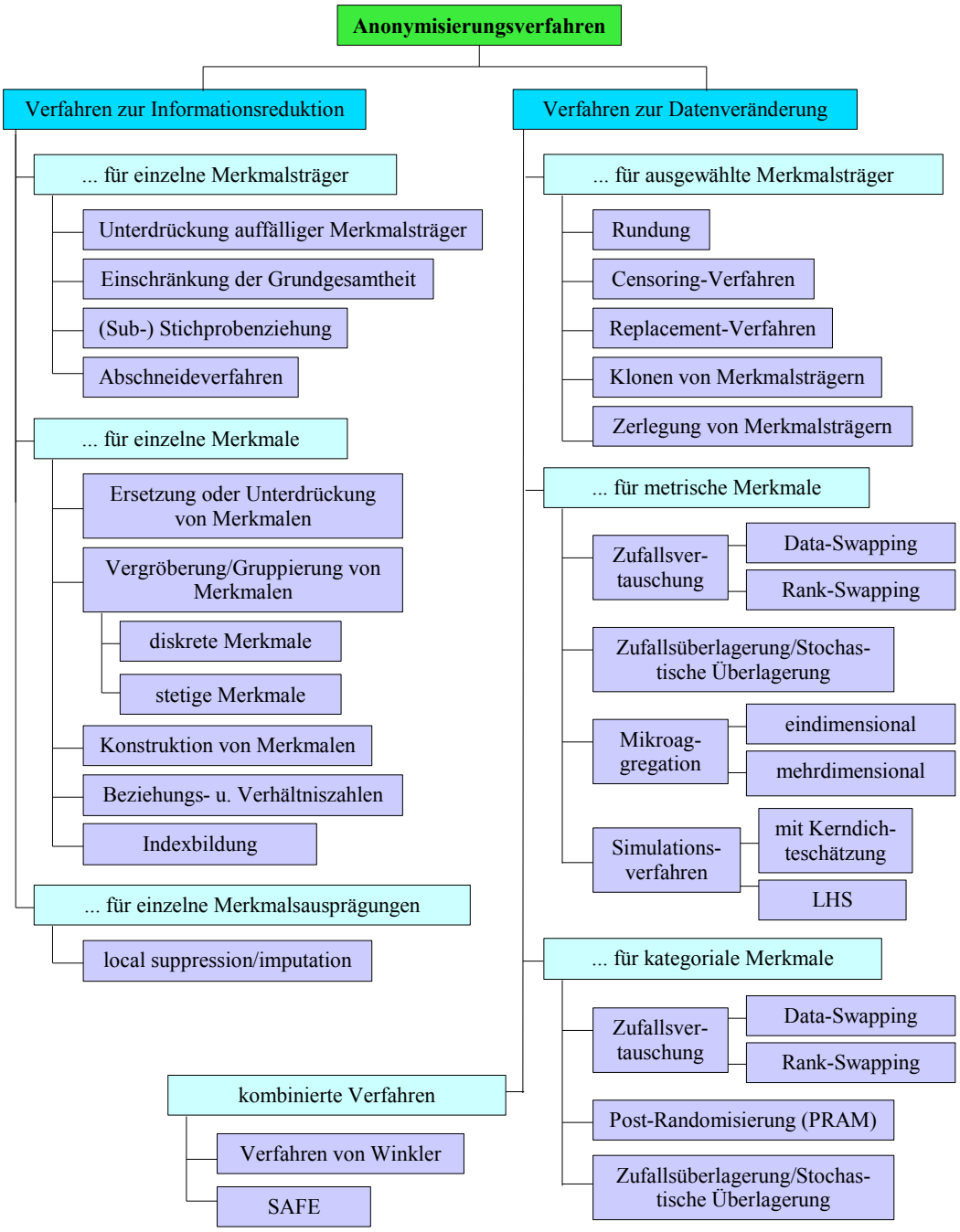

Abbildung 1.2: Anonymisierungsverfahren

Auf die Charakterisierung und Erläuterung der verschiedenen Anonymisierungsverfahren und deren Kombinationsmöglichkeiten wird nicht näher eingegangen, da dies den Rahmen dieser Arbeit übersteigen würde. Nähere Informationen sind unter anderem im „Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten“ (Statistisches Bundesamt, Statistik und Wissenschaft, Bd. 4/2005) zu finden.

1.1 Ziel der Arbeit

Ziel der Arbeit ist es, eine spezielle Form anonymisierter Daten hinsichtlich ihres statistischen Analysepotentials zu untersuchen. Diese spezielle Form der Anonymisierung zeichnet sich dadurch aus, dass keinerlei Mikrodaten, sondern nur Randverteilungen verschiedener Merkmale in Kontingenztafeln zu Analysezwecken zur Verfügung stehen. Dies geschieht vor dem Hintergrund, dass viele öffentlich zugänglich gemachte Daten (z.B. vom Statistischen Bundesamt) sehr aufwendig anonymisiert werden müssen, um den Datenschutz zu gewährleisten. Wenn das Analysepotential von Randverteilungen sehr hoch wäre, könnte man in einigen Bereichen auf aufwendige Anonymisierung verzichten und ausschließlich Randverteilungen öffentlich zugänglich machen.

Es soll untersucht werden, inwieweit statistische Berechnungen nur mit Randverteilungen möglich sind und ob diese Analysen aussagekräftige und sinnvoll interpretierbare Ergebnisse hervorbringen. Die Auswahl der statistischen Analysemethoden spielt hier eine große Rolle, denn ohne Einzeldaten sind einige statistische Analyseverfahren, wie zum Beispiel Regressionen, Varianzanalysen, nicht anwendbar.

Als Methoden zur Durchführung der Berechnungen in R dienen die lineare und konvexe Optimierung. Diese sollen im Laufe der Arbeit in Anlehnung an die Methode der Fréchet Bounds erweitert werden.

Um das Potential der Analysen bewerten zu können, werden die Methoden auf unterschiedliche Szenarien angewendet. Die Ergebnisse dieser Anwendungen sollen zeigen, dass durch statistische Analysen, die sich nur auf gegebene Randverteilungen stützen, sinnvolle Aussagen und Interpretationen möglich sind.

1.2 Aufbau der Arbeit

Die vorliegende Arbeit gliedert sich im Wesentlichen in drei Teile:

- Theorie und Anwendung in $2 \times 2 \times 2$ Kontingenztafeln
- Kostenanwendung in der Medizin anhand einer $2 \times 2 \times 2$ Kontingenztafel
- Theorie und Anwendung der statistischen Auswertung in $2 \times 2 \times 2$ Kontingenztafeln

Im ersten Teil der Arbeit werden zuerst die theoretischen Hintergründe der Kontingenztafeln, der linearen Optimierung und der Fréchet Bounds zur Berechnung der Einzelhäufigkeiten ($\widehat{=}$ Häufigkeitsverteilung) erläutert.

Daraufhin werden diese anhand eines medizinischen Szenarios angewendet. In diesem Szenario handelt es sich um eine Studie über die Wirkung des Placebo-Effekts in Zusammenhang mit den Merkmalen „Therapieerfolg“ und „Geschlecht“.

Im darauf folgenden Kapitel wird an einem beispielhaften Kosten-Szenario aus der Medizin die Anwendung der linearen Optimierung erweitert. Das Szenario beschäftigt sich mit der Kostenberechnung für Behandlungen von Patienten, die an Diabetes erkrankt sind.

Im letzten Teil wird die für die Arbeit relevante Theorie für die statistische Auswertung erläutert und danach auf das Anwendungsbeispiel des ersten Teils angewendet.

Abschließend wird ein kurzer Ausblick über mögliche Erweiterungen und Verbesserungen bezüglich dieser Thematik gegeben.

Alle Anwendungen finden mit der statistischen Software „R“ und in einem modellhaften Rahmen statt. Es werden in jeder Anwendung drei Merkmale mit jeweils 2 Merkmalsausprägungen in Kontingenztafeln kombiniert. Alle Zahlen und Sachverhalte, die in dieser Arbeit auftreten, sind fiktiv gewählt und spiegeln nicht zwangsläufig die Realität wider.

Kapitel 2

Theorie und Anwendung in $2 \times 2 \times 2$ Kontingenztafeln

2.1 Theorie zur Berechnung der Einzelhäufigkeiten

In diesem Abschnitt wird der grundlegende Aufbau einer Kontingenztafel erklärt, die Theorie der linearen Optimierung und der Fréchet-Bounds (angelehnt an eine Arbeit von Lawrence Cox) erläutert.

2.1.1 Kontingenztafeln

Eine Kontingenztafel stellt eine Tabelle dar, die die absoluten oder relativen Häufigkeiten von Kombinationen bestimmter Merkmalsausprägungen enthält. Das Wort *Kontingenz* hat dabei die Bedeutung des gemeinsamen Auftretens von zwei oder mehreren Merkmalen bzw. Variablen. Die Häufigkeiten in der Tabelle werden durch deren Randsummen ergänzt, die die Randhäufigkeitsverteilung bzw. Randverteilung bilden. Den Spezialfall einer 2×2 Kontingenztafel bezeichnet man auch als Vierfeldertafel (im Fall der Kombination zweier Merkmale mit jeweils 2 Ausprägungen).

Tabelle 2.1 zeigt eine Vierfeldertafel mit den Merkmalen X und Y. Merkmal X besitzt die Ausprägungen $i = 1, \ldots, I$, (in diesem Fall ist $I = 2$). Merkmal Y beinhaltet vergleichbar mit X ebenfalls die Merkmalsausprägungen $j = 1, \ldots, J$, (in diesem Fall ist $J = 2$). Die Einträge einer Kontingenztabelle werden immer als *nichtnegativ*, $h_{ij} \geq 0$, angenommen.

		Merkmal Y		
		1	2	$h_{i\cdot}$
Merkmal X	1	h_{11}	h_{12}	$h_{1\cdot}$
	2	h_{21}	h_{22}	$h_{2\cdot}$
	$h_{\cdot j}$	$h_{\cdot 1}$	$h_{\cdot 2}$	$h_{\cdot\cdot}$

Tabelle 2.1: Beispiel zweidimensionale 2×2 Kontingenztafel

Die Dreidimensionalität einer Kontingenztafel entsteht, wenn zu den üblichen zwei Merkmalen ein drittes Merkmal mit einbezogen wird. Das dritte Merkmal repräsentiert somit die dritte Dimension. In den folgenden Tabellen sind ebenfalls die Merkmale X, mit $i = 1, \ldots, I$, und Y, mit $j = 1, \ldots, J$, enthalten. Hinzukommt das dritte Merkmal Z mit den Merkmalsausprägungen $k = 1, \ldots, K$, (in diesem Fall ist auch $K = 2$). Deshalb benötigt die Nomenklatur der Häufigkeiten einen zusätzlichen Dimensionsindex $k : h_{ij\boldsymbol{k}} \geq 0$.
Für die Dreidimensionalität gibt es unter anderem folgende Darstellungsweisen:

- 3D-Darstellung
- Partialtabellen-Darstellung
- Marginaltabellen-Darstellung
- Einfache dreidimensionale Kontingenztafel

3D-Darstellung

Eine 3D-Darstellung (Abbildung 2.1) einer solchen Kontingenztafel ist für die Visualisierung der Problemstellung und der Berechnungen nicht sehr praktikabel, da einige Felder des „Würfels“, z.B. h_{222} und einige Randhäufigkeiten, nicht darstellbar sind. Diese Darstellungsweise kann nur einen groben Überblick geben.

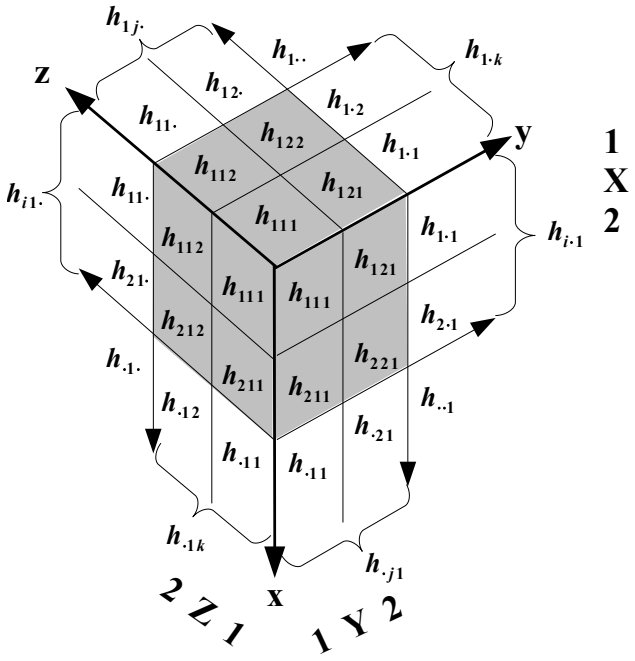

Abbildung 2.1: Kontingenztafel in 3D, Kontingenz-“Würfel“

Stattdessen bietet es sich an, den Kontingenz-„Würfel“ in die verschiedenen Ebenen aufzuspalten.

Partialtabellen-Darstellung

In der Tabelle 2.2 ist der „Würfel“ in die Ebenen $k = 1$ und $k = 2$ aufgeteilt. Die Tabellen dieser Aufspaltungsweise nennt man auch *Partialtabellen* [Mortensen, 2010, S.18].

$k = 1$		Merkmal Y		
		1	2	$h_{i\cdot 1}$
Merkmal X	1	h_{111}	h_{121}	$h_{1\cdot 1}$
	2	h_{211}	h_{221}	$h_{2\cdot 1}$
	$h_{\cdot j1}$	$h_{\cdot 11}$	$h_{\cdot 21}$	$h_{\cdot\cdot 1}$

$k = 2$		Merkmal Y		
		1	2	$h_{i\cdot 2}$
Merkmal X	1	h_{112}	h_{122}	$h_{1\cdot 2}$
	2	h_{212}	h_{222}	$h_{2\cdot 2}$
	$h_{\cdot j2}$	$h_{\cdot 12}$	$h_{\cdot 22}$	$h_{\cdot\cdot 2}$

Tabelle 2.2: *Partialtabellen*-Darstellung der dreidimensionalen $I \times J$ - Kontingenztafel, K Ebenen

Äquivalent dazu kann man den „Würfel“ nach $X = i$ und $Y = j$ aufspalten, siehe Anhang A.1 und A.2, Seite 65. Abhängigkeiten und Zusammenhänge in einer Partialtabelle werden „partielle Assoziationen“ genannt.

Wenn man die Einträge der rechten Tabelle mit dem Kontingenz-“Würfel“ vergleicht, erkennt man, dass diese Einträge der Tabelle im 3-D-“Würfel“ nicht sichtbar sind, weil sie im Hintergrund liegen.

Marginaltabellen-Darstellung

Eine weitere Darstellungsmöglichkeit ist die *Marginaltabellen*-Darstellung (Tabelle 2.3). Hier ist über eines der Merkmale aggregiert, d.h. summiert worden. Dadurch stehen die 2-dimensionalen Randhäufigkeiten innerhalb der Tafel. Abhängigkeiten und Zusammenhänge in einer Marginaltabelle heißen „marginale Assoziationen“. Die Assoziationen in Marginaltabellen können sich sehr von denen in Partialtabellen unterscheiden, dieses Phänomen ist als *Simpsons Paradoxon* bekannt [Mortensen, 2010, S.18]. In Marginaltabellen lässt sich die Gesamthäufigkeit $h_{\cdots}$, die sich aus den Randhäufigkeiten errechnet, deutlicher erkennen als in den Partialtabellen. In diesen muss man erst die 1-dimensionalen Randhäufigkeiten addieren, um die Gesamthäufigkeit zu erhalten.

$\sum_{k=1}^{K=2} k$		Merkmal Y		
		1	2	$h_{i\cdot\cdot}$
Merkmal X	1	$h_{11\cdot} = h_{111} + h_{112}$	$h_{12\cdot} = h_{121} + h_{122}$	$h_{1\cdot\cdot}$
	2	$h_{21\cdot} = h_{211} + h_{212}$	$h_{22\cdot} = h_{221} + h_{222}$	$h_{2\cdot\cdot}$
	$h_{\cdot j\cdot}$	$h_{\cdot 1\cdot}$	$h_{\cdot 2\cdot}$	$h_{\cdots}$

Tabelle 2.3: *Marginaltabellen*-Darstellung der dreidimensionalen $I \times J$ - Kontingenztafel, Summierte k Ebenen

Das **Simpsons Paradoxon** wurde erstmals 1951 von Edward Hugh Simpson systematisch untersucht. Dabei kann die Bewertung des Zusammenhangs zwischen verschiedenen Merkmalen unterschiedlich ausfallen, je nachdem ob die Ergebnisse der Merkmalsausprägungen aufgeteilt (Partialtabellen) oder aggregiert (Marginaltabellen) sind. Die Zusam-

menhänge können nicht nur unterschiedliche (nah bei einander liegende) Werte aufweisen, sondern sogar unterschiedliche Richtungen des Zusammenhangs anzeigen. Für nähere Informationen und Beispiel ist ausreichend Literatur vorhanden. Hier ist eine kleine Auswahl: [Beck-Bornholdt, 2005], [Simpson, 1951], [Wagner, 1982], [Kühne, 2009].

Einfache dreidimensionale Kontingenztafel

Die einfachste und gleichzeitig vollständigste Darstellung, in der alle Einzelhäufigkeiten in **einer** Tabelle vertreten sind, ist die einfache dreidimensionale Kontingenztafel (Tabelle 2.4). Allerdings ist zu beachten, dass hier nicht *alle* 1-dimensionalen Randhäufigkeiten in Erscheinung treten. In folgender Tabelle 2.4 sind zum Beispiel die Randhäufigkeiten $h_{..k}$ nicht direkt erkennbar. Um diese Häufigkeiten in der Tabelle sehen zu können, müsste man, statt das Merkmal Y durch Merkmal Z zu unterteilen, das Merkmal Z durch Merkmal Y unterteilen, siehe Anhang A.3, Seite 66.

		Merkmal Y				
		1		2		
		$Z=1$	$Z=2$	$Z=1$	$Z=2$	$h_{i..}$
Merkmal X	1	h_{111}	h_{112}	h_{121}	h_{122}	$h_{1..}$
	2	h_{211}	h_{212}	h_{221}	h_{222}	$h_{2..}$
	$h_{.j.}$	$h_{.1.}$		$h_{.2.}$		$h_{...}$

Tabelle 2.4: Dreidimensionale $I \times J$ - Kontingenztafeln

Die Einzelhäufigkeiten lassen sich sowohl durch die 1-dimensionalen Randhäufigkeiten als auch durch die 2-dimensionalen Randhäufigkeiten ausdrücken, wie in den Tabellen 2.3 und 2.4 zu erkennen ist.

Gültige Restriktionen:

$$\begin{aligned} h_{...} &= \sum_{i=1}^{2} h_{i..} = \sum_{i=1}^{2}\sum_{j=1}^{2} h_{ij.} = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} h_{ijk} \ , \\ \iff h_{...} &= \sum_{j=1}^{2} h_{.j.} = \sum_{j=1}^{2}\sum_{k=1}^{2} h_{.jk} = \sum_{j=1}^{2}\sum_{k=1}^{2}\sum_{i=1}^{2} h_{ijk} \ , \\ \iff h_{...} &= \sum_{k=1}^{2} h_{..k} = \sum_{k=1}^{2}\sum_{i=1}^{2} h_{i\cdot k} = \sum_{k=1}^{2}\sum_{i=1}^{2}\sum_{j=1}^{2} h_{ijk} \ . \end{aligned} \tag{2.1}$$

mit $\forall h_{ijk} \geq 0$ und $\forall h_{ijk} = \left[\underline{h_{ijk}} \ , \ldots , \ \overline{h_{ijk}}\right]$,
$h_{i..}$, $h_{.j.}$, $h_{..k} \mathrel{\widehat{=}}$ 1-dimensionale Randhäufigkeiten,
$h_{ij.}$, $h_{.jk}$, $h_{i\cdot k} \mathrel{\widehat{=}}$ 2-dimensionale Randhäufigkeiten.

2.1.2 Lineare Optimierung

Die *Lineare Optimierung* (oder *Lineare Programmierung*) ist eines der Hauptverfahren in der Unternehmens- und Planungsforschung (Operations Research). Das Ziel ist die Optimierung linearer Funktionen über einer Menge, die durch lineare Bedingungen (Gleichungen und Ungleichungen) eingeschränkt ist. Es wird also nach Extremstellen (Minimum, Maximum) einer Funktion f, die meist als *Zielfunktion* bezeichnet wird, gesucht.

Die lineare Optimierung ist ein Spezialfall der verallgemeinerten *konvexen Optimierung*, die Grundlage für viele Lösungsverfahren in der ganzzahligen linearen und der nichtlinearen Optimierung ist. Wenn die Anzahl der Variablen und Restriktionen endlich ist, kann die Menge der Variablen auch grafisch als *konvexer Polyeder* dargestellt werden.

Definition Standard-Maximum-Problem
$A \in \mathbb{R}^{m,n}$, $b \in \mathbb{R}^{m,1}$, $c \in \mathbb{R}^{n,1}$. Eine *zulässige Lösung* ist ein Vektor $w \in \mathbb{R}^{n,1}$, der die Nebenbedingungen

$$\begin{array}{ccccc} A & \cdot & w & \leq & b \\ (m \times n) & & (n \times 1) & & (m \times 1) \\ & & w & \geq & 0 \end{array} \tag{2.2}$$

erfüllt.
Die Gleichung

$$\begin{array}{ccccc} c^T & \cdot & w & \rightarrow & \max \\ (1 \times n) & & (n \times 1) & & \end{array} \tag{2.3}$$

ist die Zielfunktion und stellt mit den Restriktionen das Standard-Maximum-Problem dar.

Definition Standard-Minimum-Problem
$A \in \mathbb{R}^{m,n}$, $b \in \mathbb{R}^{m,1}$, $c \in \mathbb{R}^{n,1}$. Eine *zulässige Lösung* ist ein Vektor $w \in \mathbb{R}^{n,1}$, der die Nebenbedingungen

$$\begin{array}{ccccc} A & \cdot & w & \geq & b \\ (m \times n) & & (n \times 1) & & (m \times 1) \\ & & w & \geq & 0 \end{array} \tag{2.4}$$

erfüllt.

Die Gleichung

$$\begin{array}{c} c^T \cdot w \rightarrow \min \\ (1 \times n)\ (n \times 1) \end{array} \tag{2.5}$$

ist die Zielfunktion und stellt mit den Restriktionen das Standard-Minimum-Problem dar.

Umformung des Standard-Minimum-Problems

Eine andere Möglichkeit, ein Standard-Minimum-Problem zu erzeugen, besteht darin, die Zielfunktionsvektor c mit (-1) zu multiplizieren und die Zielfunktion zu maximieren:

$$\begin{array}{cccccc} c^T & \cdot & w & \rightarrow & \min & \\ (1 \times n) & & (n \times 1) & & & \end{array} \iff \begin{array}{cccccc} -c^T & \cdot & w & \rightarrow & \max & . \\ (1 \times n) & & (n \times 1) & & & \end{array} \tag{2.6}$$

Umformung von Ungleichheits-Bedingungen

Um Gleichheitsbedingungen $A \cdot w = b$ aus $\leq$-Bedingungen und $\geq$-Bedingungen zu erzeugen, multipliziert man A und b ebenfalls mit (-1).
Dies lässt sich ausdrücken durch

$$A \cdot w \geq b \iff -A \cdot w \leq -b\,. \tag{2.7}$$

Daraus folgt

$$\begin{array}{ccccc} A & \cdot & w & \leq & b \\ \wedge & & & & \\ -A & \cdot & w & \leq & -b \end{array} \iff \begin{array}{ccccc} A & \cdot & w & = & b\,. \end{array} \tag{2.8}$$

Diese Umformungen sind notwendig, damit sowohl das Maximum- als auch das Minimum-Problem in **einem** Gleichungssystem untergebracht werden kann, denn für die Berechnungen der Intervalle für die Einzelhäufigkeiten in den Kontingenztabellen braucht man sowohl das Standard-Maximum-Problem für die *obere Schranke* der Intervalle als auch das Standard-Minimum-Problem für die *untere Schranke* der Intervalle.

Dualität

In der linearen Optimierung gibt es außerdem noch den Begriff der *Dualität*. Diese wird hier erläutert, da sie ein Bestandteil der linearen Optimierung ist. Allerdings hat sie für diese Arbeit keine große Bedeutung.

Jedem linearen Optimierungsproblem in Standardform wird ein zugehöriges *duales Problem* zugeordnet. Man geht hierbei von einem Maximum-Problem über zu einem Minimum-Problem (oder umgekehrt). Für jede Bedingung des ursprünglichen Problems führt man eine

neue Variabel ein. Die ursprüngliche Bedingungsmatrix A wird transponiert und wird so zur Bedingungsmatrix A^T des dualen Problems. Des Weiteren werden die Rollen von c und b vertauscht (sie werden ebenfalls transponiert): c wird zum Spaltenvektor (Bedingungsvektor) und b^T wird zum Zielfunktionsvektor. Es gibt die sogenannten *Dualitätssätze*, in denen die engen Zusammenhänge zwischen *primalem* und *dualem* Problem beschrieben werden. Nachfolgend werden die wichtigsten Eigenschaften genannt:

→ Das duale Standard-Maximum-Problem zu dem dualen Standard-Minimum-Problem ist wieder das ursprüngliche primale Standard-Maximum-Problem (oder umgekehrt).

→ Dualitätssätze:

- schwache Dualität: für alle zulässigen Lösungen w und u des primalen bzw. dualen Problems gilt $c^T \cdot w \ \leq \ b^T \cdot u$, d.h. der Wert jeder dualen Lösung ist größer gleich der Wert der primalen Lösung,
- starke Dualität: w und u sind Optimallösungen, wenn in der Gleichung der *schwachen Dualität* **Gleichheit** herrscht.

→ Für das primale Standard-Maximum-Problem existiert genau dann eine Optimallösung w_{opt} , wenn bereits für das duale Standard-Minimum-Problem eine Optimallösung u_{opt} exisitiert.

[Augustin, 2009]

Weitere Informationen sowohl über die lineare, nichtlineare und konvexe Optimierung als auch über die Dualität sind zum Beispiel in der Quelle [Grötschel, 2004] zu finden.

Darstellung der Matrizen und Vektoren

Nun folgt die Darstellung der benötigten Matrizen und Vektoren. Die Matrizen und Vektoren werden gleichzeitig sowohl für das Standard-Maximum-Problem als auch für das Standard-Minimum-Problem konzipiert.

Seien $m = 12$, $n = 8$.

Die Einzelhäufigkeiten aus der Kontingenztafel sind im Vektor $w = \begin{pmatrix} h_{111} \\ h_{112} \\ h_{121} \\ h_{122} \\ h_{211} \\ h_{212} \\ h_{221} \\ h_{222} \end{pmatrix} \in \mathbb{Z}^{n,1}$

enthalten. Der *Gesamt*-Bedingungsvektor $b_G = \begin{pmatrix} b \\ -b \end{pmatrix} = \begin{pmatrix} h_{1..} \\ h_{2..} \\ h_{.1.} \\ h_{.2.} \\ h_{..1} \\ h_{..2} \\ -h_{1..} \\ -h_{2..} \\ -h_{.1.} \\ -h_{.2.} \\ -h_{..1} \\ -h_{..2} \end{pmatrix} \in \mathbb{Z}^{m,1}$ und die

Gesamt-Bed.-matrix $A_G = \begin{pmatrix} A \\ -A \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 \\ -1 & -1 & 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 & -1 & -1 \\ -1 & 0 & -1 & 0 & -1 & 0 & -1 & 0 \\ 0 & -1 & 0 & -1 & 0 & -1 & 0 & -1 \end{pmatrix} \in \mathbb{N}^{m,n}$

stellen die Operatoren für die Nebenbedingungen mit den Randhäufigkeiten dar.

Damit sind das Standard-Maximum- und Standard-Minimum-Problem in den Bedingungen

$$A_G \cdot w = \begin{pmatrix} A \\ -A \end{pmatrix} \cdot w \leq \begin{pmatrix} b \\ -b \end{pmatrix} = b_G \tag{2.9}$$

vereint.
Für das Standard-Maximum-Problem stellt

$$\overline{h_{111}}\, c^T = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,n} \tag{2.10}$$

den Zielfunktionsvektor für h_{111} dar,
für das Standard-Minimum-Problem lautet dieser

$$\underline{h_{111}}\, c^T = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,n}. \tag{2.11}$$

Damit das Ergebnis für den Vektor $\underline{h_{111}}\, c^T$ richtig ist, muss die vorherige Negation (2.6) durch nochmalige Multiplikation des Funktionswertes mit (-1) rückgängig gemacht werden.
Als Ergebnisse kommen dann die Grenzen der Intervalle für jedes h_{ijk} in der Form

$$h_{ijk} = [\min\{h_{ijk}\}, \ \ldots, \ \max\{h_{ijk}\}] = \left[\underline{h_{ijk}}, \ \ldots, \ \overline{h_{ijk}}\right]$$

heraus.

2.1.3 Fréchet Bounds

Mit Fréchet Bounds können Ober- und Untergrenzen für Zellhäufigkeiten (= Einzelhäufigkeiten) aus Randhäufigkeiten errechnet werden. Dies ist bei der Aggregation/Anonymisierung von Daten, vor allem bei Zellunterdrückung (cell suppression) oder Datenüberlagerung (data swapping), interessant. Durch die einzelnen Aggregationsverfahren wird versucht, Daten zusammenzufassen und so höher Zellhäufigkeiten zu erreichen. Dies hat dann eine geringere Chance der Reidentifizierung zur Folge.

Die Aggregation von Daten bedeutet für Kontingenztafeln, dass Kategorien bzw. Merkmalsausprägungen zusammengefasst werden. Die Randverteilungen und die Zellhäufigkeiten innerhalb der Kontingenztafeln ändern sich dadurch. Die Randverteilung, die bei der Aggregation entsteht, wird fixiert, also festgehalten. Um die Anonymisierung noch zu erweitern, versucht man in den einzelnen Zellen nicht nur eine Zahl pro Zelle, sondern ein Intervall pro Zelle anzugeben.

Die Methode der Fréchet Bounds kann generell für Kontingenztafeln eingesetzt werden. Beispielsweise können bei (2×2-Kontingenztafeln) die 1-dimensionalen Fréchet Bounds angewendet werden. „1-dimensional“ bedeutet in diesem Fall, welche Randhäufigkeiten in die Berechnung der Fréchet Bounds einfließen. Im Folgenden ist die allgemeine Formel der Fréchet Bounds für 2×2-Kontingenztafeln dargestellt.

$$\max\{h_{i\cdot} + h_{\cdot j} - h_{\cdot\cdot}\ ,\ 0\} \ \leq\ h_{ij}\ \leq\ \min\{h_{i\cdot}\ , h_{\cdot j}\}$$

Bei 2×2-Kontingenztafeln gibt es nur 1-dimensionale Randhäufigkeiten. Bei den $2 \times 2 \times 2$-Kontingenztafeln, um die es in dieser Arbeit hauptsächlich geht, gibt es, wie bereits ausführlich erläutert, 1-dimensionale und 2-dimensionale Randhäufigkeiten. Demzufolge können die 1-dimensionalen Fréchet Bounds oder die 2-dimensionalen Fréchet Bounds angewendet werden. Je mehr Dimensionen die Fréchet Bounds haben, desto enger werden die Intervalle, die in den Zellen berechnet werden. [Fienberg, NA, S. 1-3]

2.1.3.1 2-dimensionale Fréchet Bounds

Wie schon in den Gleichungen (2.1) zu sehen war, summieren sich die Einzelhäufigkeiten zu den 2-dimensionalen Randhäufigkeiten und diese wiederum zu den 1-dimensionalen Randhäufigkeiten auf.

Bei den F-Bounds nach Lawrence Cox [Cox, 2002, S. 21-23] handelt es sich ebenfalls um Fréchet Bounds. Lawrence Cox hat die 2-dimensionalen Fréchet Bounds mit den 2-dimensionalen Randhäufigkeiten $h_{ij\cdot}$, $h_{\cdot jk}$ und $h_{i\cdot k}$ verwendet. Das bedeutet die 2-dimensionalen Randhäufigkeiten sind fest und bekannt.

Die jeweils zugehörige Summation der Einzelhäufigkeiten sieht folgendermaßen aus:

$$h_{ij\cdot} = \sum_{k=1}^{K} h_{ijk} \ , \ h_{\cdot jk} = \sum_{i=1}^{I} h_{ijk} \ , \ h_{i\cdot k} = \sum_{j=1}^{J} h_{ijk} \ , \tag{2.12}$$

$\forall\ h_{ijk} \ \geq \ 0 \ , \ \forall\ h_{\cdot jk}, h_{i\cdot k}, h_{ij\cdot} \ \geq \ 0 \ .$

Aus den zuvor genannten Gleichungen und den Gleichungen (2.1) lassen sich weitere Beziehungen zwischen den Randhäufigkeiten entwickeln, die sogenannten *Konsistenzbedingungen*:

$$\begin{aligned}
\sum_{k=1}^{K} h_{i\cdot k} &= \sum_{j=1}^{J} h_{ij\cdot} = h_{i\cdot\cdot} \ , \\
\sum_{i=1}^{I} h_{ij\cdot} &= \sum_{k=1}^{K} h_{\cdot jk} = h_{\cdot j\cdot} \ , \\
\sum_{i=1}^{I} h_{i\cdot k} &= \sum_{j=1}^{J} h_{\cdot jk} = h_{\cdot\cdot k} \ .
\end{aligned} \tag{2.13}$$

Die Gesamthäufigkeit $h_{\cdot\cdot\cdot}$ errechnet sich, wie ebenfalls schon erwähnt, durch

$$h_{\cdot\cdot\cdot} = \sum_{i=1}^{I} h_{i\cdot\cdot} = \sum_{j=1}^{J} h_{\cdot j\cdot} = \sum_{k=1}^{K} h_{\cdot\cdot k} \ . \tag{2.14}$$

Zur Bestimmung der ganzzahligen *unteren* und *oberen Schranken* für jeden Eintrag h_{ijk} muss den Forderungen (2.12) - (2.13) genüge getan werden.

Eine exakte Abgrenzung ist durch das Intervall $[\min\{h_{ijk}\}, \ \ldots, \ \max\{h_{ijk}\}]$ für alle ganzzahligen zulässigen Lösungen $h^* = \{h^*_{ijk}\}$ von (2.12) - (2.13) bestimmt.

In einer dreidimensionalen Kontingenztafel werden die *unteren* und *oberen Schranken* für jeden Eintrag h_{ijk} mit der Formel

$$\max \left\{ \begin{array}{l} 0 \\ h_{ij\cdot} + h_{i\cdot k} - h_{i\cdot\cdot} \\ h_{ij\cdot} + h_{\cdot jk} - h_{\cdot j\cdot} \\ h_{i\cdot k} + h_{\cdot jk} - h_{\cdot\cdot k} \end{array} \right\} \ \leq \ h_{ijk} \ \leq \ \min\{h_{ij\cdot}\ , h_{\cdot jk}\ , h_{i\cdot k}\} \tag{2.15}$$

berechnet.
Das bedeutet

$$\underline{h_{ijk}} \ = \ \min\{h_{ijk}\} \ = \ \max \left\{ \begin{array}{l} 0 \\ h_{ij\cdot} + h_{i\cdot k} - h_{i\cdot\cdot} \\ h_{ij\cdot} + h_{\cdot jk} - h_{\cdot j\cdot} \\ h_{i\cdot k} + h_{\cdot jk} - h_{\cdot\cdot k} \end{array} \right\}$$

und

$$\overline{h_{ijk}} \;=\; \max\{h_{ijk}\} \;=\; \min\{h_{ij\cdot}\,, h_{\cdot jk}\,, h_{i\cdot k}\}$$

ergeben das Intervall für jeden Eintrag

$$h_{ijk} \;=\; [\min\{h_{ijk}\},\; \ldots,\; \max\{h_{ijk}\}] \;=\; \left[\underline{h_{ijk}}\;,\; \ldots,\; \overline{h_{ijk}}\right] \;.$$

[Cox, 2002, S. 21-23]

2.1.3.2 Implementierung der 2-dimensionalen Fréchet Bounds in die Lineare Optimierung

Zur Implementierung der 2-dimensionalen Fréchet Bounds werden die zusätzlich bekannten und festen 2-dimensionalen Randhäufigkeiten als weitere lineare Bedingungen in die Formeln der Linearen Optimierung mit aufgenommen. Diese linearen Bedingungen werden aus den Gleichungen (2.12) generiert. Hier sieht man die direkte Verbindung bzw. Äquivalenz von den 2-dimensionalen F-Bounds und der linearen Optimierung.

Das bedeutet letztendlich, dass die Bedingungsmatrix A_G und der Bedingungsvektor b_G um die 2-dimensionalen Randhäufigkeiten erweitert werden.

Um die *Gleichheits*bedingungen zu generieren, müssen natürlich die negierten 2-dimensionalen Randhäufigkeiten ebenfalls aufgenommen werden.

Die Darstellung der benötigten Matrizen und Vektoren ist äquivalent zu Kapitel 2.1.2 und wird im Anhang A.2.2.1 auf Seite 72 und auf Seite 73 bereits mit Zahlenwerten visualisiert. Der Vektor w und die Zielfunktionsvektoren $\frac{}{h_{ijk}}c^T$ verändern sich nicht, siehe Seite 21, 22 und Anhang A.2.1.1.

Durch die Implementierung der 2-dimensionalen Fréchet Bounds in die lineare Optimierung sollen Intervalle berechnet werden, die sich von der Berechnung mit nur 1-dimensionalen Randhäufigkeiten unterscheiden.

2.2 Berechnung der Einzelhäufigkeiten in R

2.2.1 Rahmenbedingungen

Für die Anwendung der zuvor erläuterten Theorie (Kapitel 2.1) wird nun folgender Sachverhalt zugrunde gelegt.

In medizinischen Studien wird häufig untersucht, wie bzw. ob Medikamente wirken. Dabei wurde bereits festgestellt, dass der Placebo-Effekt oft eine Rolle spielt. Das heißt zum Beispiel, dass selbst Tabletten ohne Wirkstoff wirken, weil die behandelten Patienten daran glauben, ein vollwertiges Medikament bekommen zu haben.

Das Merkmal X wird in dieser Anwendung „Medikament“ sein und die Ausprägungen „Placebo $= 0$“ und „Wirkstoff $= 1$“ haben. Als zweites Merkmal Y wird der „Erfolg der Therapie“ mit „nein $= 0$“ und „ja $= 1$“ ausgedrückt. Als dritte Dimension wäre das Geschlecht interessant. Reagieren Frauen $(= 1)$ und Männer $(= 0)$ unterschiedlich auf Placebo bzw. Wirkstoff?

Zusammengefasst, bilden nun folgende Informationen den Ausgangspunkt:

- alle drei Merkmale, die hier verwendet werden, sind qualitativ bzw. kategorial,
- Merkmal $X \mathrel{\widehat{=}}$ „Medikament“ mit den Ausprägungen $x_{i=1} = 1 \mathrel{\widehat{=}}$ „Wirkstoff“ und $x_{i=2} = 0 \mathrel{\widehat{=}}$ „Placebo“, $(i = 1, 2)$,
- Merkmal $Y \mathrel{\widehat{=}}$ „Erfolg der Therapie“ mit den Ausprägungen $y_{j=1} = 1 \mathrel{\widehat{=}}$ „ja“ und $y_{j=2} = 0 \mathrel{\widehat{=}}$ „nein“, $(j = 1, 2)$,
- Merkmal $Z \mathrel{\widehat{=}}$ „Geschlecht“ mit den Ausprägungen $z_{k=1} = 1 \mathrel{\widehat{=}}$ „weiblich“ und $z_{k=2} = 0 \mathrel{\widehat{=}}$ „männlich“, $(k = 1, 2)$.

Merkmal	**Bezeichnung**	**Ausprägungen**	**Kodierung**
X	Medikament	$i = 1$	„Wirkstoff“ $\mathrel{\widehat{=}} 1$
		$i = 2$	„Placebo“ $\mathrel{\widehat{=}} 0$
Y	Erfolg der Therapie	$j = 1$	„ja“ $\mathrel{\widehat{=}} 1$
		$j = 2$	„nein“ $\mathrel{\widehat{=}} 0$
Z	Geschlecht	$k = 1$	„weiblich“ $\mathrel{\widehat{=}} 1$
		$k = 2$	„männlich“ $\mathrel{\widehat{=}} 0$

Tabelle 2.5: Rahmenbedingungen

Des Weiteren sind, wie schon gesagt, nur die 1-dimensionalen Randhäufigkeiten und die Gesamthäufigkeit gegeben. Wie in der Einleitung schon erwähnt, sind alle Zahlen fiktiv und willkürlich gewählt.

Seien

- Gesamthäufigkeit $h_{...} = 100$ an der Studie beteiligte Patienten,
- 1-dimensionale Randhäufigkeiten:

$$h_{1..} = 50 \ , \ h_{2..} = 50 \ ,$$
$$h_{.1.} = 69 \ , \ h_{.2.} = 31 \ ,$$
$$h_{..1} = 45 \ , \ h_{..2} = 55$$

Dabei sind logischerweise die Restriktionen wie in Formel (2.1) auf Seite 11 erfüllt.

2.2.2 Anwendung der Linearen Optimierung

Für die Implementierung in das Lineare Optimierungsverfahren müssen nun die einzelnen Vektoren und Matrizen zusammengestellt werden. Der Vektor der Einzelhäufigkeiten

$w = \begin{pmatrix} h_{111} \\ h_{112} \\ h_{121} \\ h_{122} \\ h_{211} \\ h_{212} \\ h_{221} \\ h_{222} \end{pmatrix} \in \mathbb{Z}^{8,1}$ bleibt unverändert, da dieser mit c^T maximiert bzw. minimiert werden soll. Der Gesamt-Bedingungsvektor $b_G = \begin{pmatrix} b \\ -b \end{pmatrix} = \begin{pmatrix} 50 \\ 50 \\ 69 \\ 31 \\ 45 \\ 55 \\ -50 \\ -50 \\ -69 \\ -31 \\ -45 \\ -55 \end{pmatrix} \in \mathbb{Z}^{12,1}$ beinhaltet die 1-dimensionalen Randhäufigkeiten, die wir eben festgelegt haben.

Die Bedingungsmatrix A_G beinhaltet die Zahlen $-1,\ 0$ oder 1 , um die einzelnen Bedingungsgleichungen hervorzurufen und ändert sich deshalb nicht im Vergleich zum theoretischen Abschnitt:

$$A_G = \begin{pmatrix} A \\ -A \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 \\ -1 & -1 & 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 & -1 & -1 \\ -1 & 0 & -1 & 0 & -1 & 0 & -1 & 0 \\ 0 & -1 & 0 & -1 & 0 & -1 & 0 & -1 \end{pmatrix} \in \mathbb{N}^{12,8} .$$

Die Zielfunktionsvektoren c^T der einzelnen Häufigkeiten sehen beispielhaft für h_{111} und h_{112} folgendermaßen aus: (vgl. (2.10) und (2.11))

Für h_{111} :

$$\overline{h_{111}}\, c^T = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} ,$$

$$\underline{h_{111}}\, c^T = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} .$$

Für h_{112} :

$$\overline{h_{112}}\, c^T = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} ,$$

$$\underline{h_{112}}\, c^T = \begin{pmatrix} 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} ,$$

$\vdots$

Die Zielfunktionsvektoren für die weiteren Einzelhäufigkeiten, von h_{121} bis h_{222} sind im Anhang unter A.2.1.1 zu finden.

Programmierung und Ergebnisse in R

Der folgende Code stellt für h_{111} die Programmierung der Berechnung der Ober- und Untergrenze des Intervalls beispielhaft dar. Die Programmierung für die anderen Einzelhäufigkeiten (h_{112} bis h_{222}) ist äquivalent und deshalb im Anhang A.2.1.2 bzw. auf der Daten-CD in ausführlicher Form zu finden. Dies ist auch bei allen weiteren Anwendungen der Fall.

Um eine lineare Optimierung in R durchführen zu können, muss zu Beginn ein entsprechendes Paket installiert und geladen werden. Das Paket „lpSolve“ ist ein Paket zur Anwendung der linearen Optimierung mit integer-Optionen und wird hier deshalb verwendet. Eine integer-Option lautet `int.vec=1:8` und wird in den Algorithmus des Paketes „lpSolve“ mit

aufgenommen. Diese Option wird bei allen Berechnungen mit linearer Optimierung verwendet, da durch die Verwendung absoluter Häufigkeiten auch nur ganze Zahlen als Ergebnisse errechnet werden sollen.

```
> install.packages("lpSolve")
> library(lpSolve)
```

Nun müssen die für das linearen Optimierungsproblem notwendigen Vektoren und Matrizen definiert werden. Da die Definierung sehr umfangreich und an sich für die Ausarbeitung dieser Arbeit nicht relevant ist, wird hier ebenfalls auf den Anhang A.2.1.2 bzw. die Daten-CD verwiesen.

Im Anschluss werden die einzelnen Operatoren für die lineare Optimierung zugewiesen.

```
> # Operatoren für lpSolve
> f.con <- A
> f.dir <- c("<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=")
> f.rhs <- B
```

Nun wird für das Maximum des Funktionswertes von h_{111} der Zielfunktionsvektor $\overline{h_{111}}c^T$ definiert und dem Paket lpSolve als Objekt zugewiesen.

```
# Zielfunktionsvektor c^T
> C_h111max <- c( 1, 0, 0, 0, 0, 0, 0, 0)
> C_h111max_t <- t(C_h111max)
> # Objekt für lpSolve
> f.obj1 <- C_h111max_t
> # Lin. Opt.-problem
> h_111_max <- lp ("max", f.obj1, f.con, f.dir, f.rhs, int.vec=1:8)
```

In der letzten Befehls-Zeile ist dann der Befehl für das lineare Optimierungsproblem von `h_111_max` mit den einzelnen Operatoren zu sehen.

Das Ergebnis dieses Maximierungsproblems lautet nun wie folgt:

```
> h_111_max$objval
[1] 45
> h_111_max$solution
[1] 45  5  0  0  0 19  0 31
```

Der maximale Funktionswert $\overline{h_{111}}c^T \cdot w = 45$ stellt die Obergrenze des Intervalls für h_{111} dar. Das wären in unserem Beispiel maximal 45 Frauen, die den echten Wirkstoff verabreicht bekommen haben und die Therapie erfolgreich beendet haben.

Das zweite Ergebnis von dem Befehl `h_111_max$solution` zeigt die errechnete Verteilung, mit der der maximale Funktionswert, also die Obergrenze für das Intervall für h_{111}, errechnet werden konnte. Es ist eine Verteilung von acht Zahlen und kennzeichnet die acht verschiedenen Einzelhäufigkeiten. Die Reihenfolge orientiert sich an dem Vektor w, siehe Seite 21.

Für das Minimum des Funktionswertes $\underline{h_{111}}\, c^T \cdot w$ von h_{111} wird analog die lineare Optimierung durchgeführt. Zum Minimieren wird die negative Funktion maximiert, vgl. (2.6), Seite 13.

```
# Zielfunktionsvektor c^T
> C_h111min <- c( -1, 0, 0, 0, 0, 0, 0, 0)
> C_h111min_t <- t(C_h111min)
> # Objekt für lpSolve
> f.obj2 <- C_h111min_t
> # Lin. Opt.-problem
> h_111_min <- lp ("max", f.obj2, f.con, f.dir, f.rhs, int.vec=1:8)
```

Hier muss der Hinweis beachtet werden, dass die Negation des Zielfunktionsvektors rückgängig gemacht werden muss, wie schon auf Seite 16 beschrieben. Dies wurde mit einem Bedingungs-Befehl („if“ ...„else“), der im Anhang A.2.1.2, Seite 68, zu finden ist, gelöst.

Folgendes Ergebnis für die Untergrenze des Intervalls von h_{111} wurde errechnet:

```
> h_111_min$objval
[1] 0
> h_111_min$solution
[1]  0 37 13  0 32  0  0 18
```

Das bedeutet, der Funktionswert von $\underline{h_{111}}\, c^T \cdot w$, also die Untergrenze des Intervalls für h_{111}, lautet 0. Das wären im Beispiel minimal 0 Frauen, die den echten Wirkstoff verabreicht bekommen haben, und die Therapie erfolgreich beendet haben. Das zweite Ergebnis zeigt wiederum die Verteilung, die zu dem errechneten Minimum von h_{111} geführt hat.

Wie bereits angedeutet, wurden die Berechnungen der anderen Einzelhäufigkeiten äquivalent dazu durchgeführt (R-Code, siehe im Anhang A.2.1.2 Seite 68 bzw. auf der Daten-CD). Im folgenden Output sind noch die Ergebnisse der Berechnungen für die weiteren Einzelhäufigkeiten aufgelistet:

```
> # Liste der Intervalle
> list (h_111 , h_112 , h_121 , h_122 , h_211 , h_212 , h_221 , h_222)
[[1]]
[1] "h_111 = [ 0 , ..., 45 ]"
[[2]]
[1] "h_112 = [ 0 , ..., 50 ]"
[[3]]
[1] "h_121 = [ 0 , ..., 31 ]"
[[4]]
[1] "h_122 = [ 0 , ..., 31 ]"
[[5]]
[1] "h_211 = [ 0 , ..., 45 ]"
[[6]]
[1] "h_212 = [ 0 , ..., 50 ]"
[[7]]
[1] "h_221 = [ 0 , ..., 31 ]"
[[8]]
[1] "h_222 = [ 0 , ..., 31 ]"
```

Die errechneten Intervalle der Einzelhäufigkeiten werden nun in einer dreidimensionalen $I \times J$-Kontingenztafel visualisiert, vgl. auch Tabelle 2.4 auf Seite 11.

		Y				
		$= 1$		$= 2$		
		$Z = 1$	$Z = 2$	$Z = 1$	$Z = 2$	$h_{i..}$
X	1	$h_{111} = [0, \ldots, 45]$	$h_{112} = [0, \ldots, 50]$	$h_{121} = [0, \ldots, 31]$	$h_{122} = [0, \ldots, 31]$	$h_{1..} = 50$
	2	$h_{211} = [0, \ldots, 45]$	$h_{212} = [0, \ldots, 50]$	$h_{221} = [0, \ldots, 31]$	$h_{222} = [0, \ldots, 31]$	$h_{2..} = 50$
	$h_{.j.}$	$h_{.1.} = 69$		$h_{.2.} = 31$		$h_{...} = 100$

Tabelle 2.6: Ergebnisse in dreidimensionaler $I \times J$ - Kontingenztafel

Hier ist zu erkennen, dass die einzelnen Intervalle unterschiedliche Breite aufweisen und verhältnismäßig (zur Gesamthäufigkeit $h_{...} = 100$) groß sind. Es ist darüber hinaus festzuhalten, dass zwischen der Berechnung durch lineare Optimierung und der Berechnung durch Fréchet Bounds gewisse Äquivalenz besteht, siehe auch Kapitel 2.1.3. Durch die Abhängigkeiten zwischen den Randhäufigkeiten untereinander und Einzelhäufigkeiten kann das Maximum nur den Wert der kleinsten Randhäufigkeit betragen. Das Minimum ist ebenfalls durch die Abhängigkeiten zwischen den Einzelhäufigkeiten logisch definiert. Die lineare Optimierung verfährt direkter als die Berechnungsweise der Fréchet Bounds, da die lineare Optimierung „einfach“ die Funktion maximiert und minimiert.

Um eventuell engere Intervalle für die Einzelhäufigkeiten zu erzielen, werden die Eigenschaften der 2-dimensionalen Fréchet Bounds nach Lawrence Cox die lineare Optimierung erweitern.

2.2.3 Anwendung der 2-dimensionalen Fréchet Bounds

Hier werden nun die Konsistenzbedingungen (2.13) von Seite 18 der 2-dimensionalen Fréchet Bounds in die lineare Optimierung als zusätzliche Bedingungen sowohl in die Gesamt-Bedingungsmatrix A_G als auch in den Gesamt-Bedingungsvektor b_G mit aufgenommen.

Der Vektor b_G sieht vergleichbar zu der vorhergehenden Anwendung aus, ist allerdings mit 36 Zeilen sehr umfangreich. Folgende 2-dimensionalen Randhäufigkeiten werden zusätzlich zu den 1-dimensionalen Randhäufigkeiten mit aufgenommen:

$h_{1\cdot 1} = 25$	$h_{\cdot 11} = 30$	$h_{11\cdot} = 34$
$h_{1\cdot 2} = 25$	$h_{\cdot 12} = 39$	$h_{12\cdot} = 16$
$h_{2\cdot 1} = 20$	$h_{\cdot 21} = 15$	$h_{21\cdot} = 35$
$h_{2\cdot 2} = 30$	$h_{\cdot 22} = 16$	$h_{22\cdot} = 15$

Tabelle 2.7: Werte der 2-dimensionalen Randhäufigkeiten (Kap. 2.2.3)

Die Gesamt-Bedingungsmatrix A_G sieht ebenfalls sehr ähnlich zu der Gesamt-Bedingungsmatrix ohne den Bedingungen der 2-dimensionalen F-Bounds aus. Da die Darstellungen von b_G und A_G sehr groß sind, sind diese im Anhang A.2.2.1 auf Seite 72 und 73 zu finden.

Der Vektor w und die Zielfunktionsvektoren $\frac{}{h_{ijk}} c^T$ verändern sich nicht, siehe Seite 21, 22 und Anhang A.2.1.1.

Programmierung und Ergebnisse in R

Der folgende Code stellt für h_{111} in Ausschnitten die Programmierung der Berechnung der Ober- und Untergrenze des Intervalls beispielhaft dar. Die Programmierung für die anderen Einzelhäufigkeiten (h_{112} bis h_{222}) ist äquivalent und deshalb im Anhang A.2.2.2 bzw. auf der Daten-CD ausführlich zu finden.

Nach dem Laden des lpSolve-Pakets müssen die notwendigen Vektoren und Matrizen definiert werden. Auch die einzelnen Operatoren für die lineare Optimierung werden vergleichbar zum Vorangegangen zugewiesen, siehe Anhang A.2.2.2 bzw. die Daten-CD.

Nun wird für die Obergrenze von h_{111} der Zielfunktionsvektor c^T definiert.

```
> # Zielfunktionsvektor c^T
> C_h111max <- c( 1, 0, 0, 0, 0, 0, 0, 0)
> C_h111max_t <- t(C_h111max)
> # Objekt für lpSolve
> f.obj1 <- C_h111max_t
> # Lin. Opt.-problem
> h_111_max <- lp ("max", f.obj1, f.con2, f.dir2, f.rhs2, int.vec=1:8)
```

In der letzten Befehls-Zeile ist dann der Algorithmus für das lineare Optimierungsproblem von `h_111_max` mit den einzelnen Operatoren zu sehen.

Das Ergebnis dieses Maximierungsproblems lautet nun wie folgt:

```
> h_111_max$objval
[1] 25
> h_111_max$solution
[1] 25  9  0 16  5 30 15  0
```

Das Maximum der Funktion $\overline{h_{111}} c^T \cdot w$, also die Obergrenze des Intervalls für h_{111}, lautet 25. Das wären in unserem Beispiel maximal 25 Frauen, die den echten Wirkstoff verabreicht bekommen haben, und die Therapie erfolgreich beendet haben. Dem Anhang ist weiterhin zu entnehmen, dass das Maximum der Funktion $\overline{h_{121}} c^T \cdot w$ bei 15 liegt. Es bezeichnet den Sachverhalt, dass maximal 15 Frauen, die den echten Wirkstoff verabreicht bekommen haben, die Therapie **nicht** erfolgreich beendet haben.

Für das Minimum der Funktion $\underline{h_{111}} c^T \cdot w$, also der Untergrenze des Intervalls von h_{111}, wird analog die lineare Optimierung durchgeführt. Zum Minimieren wird die negative Funktion maximiert.

```
> # Zielfunktionsvektor c^T
> C_h111min <- c( -1, 0, 0, 0, 0, 0, 0, 0)
> C_h111min_t <- t(C_h111min)
> # Objekt für lpSolve
> f.obj2 <- C_h111min_t
> # Lin. Opt.-problem
> h_111_min <- lp ("max", f.obj2, f.con2, f.dir2, f.rhs2, int.vec=1:8)
```

Das vorläufige Ergebnis für das Minimum von h_{111} lautet folgendermaßen:

```
> h_111_min$objval
[1] -10
> h_111_min$solution
[1] 10 24 15  1 20 15  0 15
```

Die Negation des Zielfunktionsvektors wird wiederum mit dem Bedingungs-Befehl („if“ . . .„else“) rückgängig gemacht, wie schon auf Seite 16 beschrieben.

```
> if ( h_111_min$objval < 0 ) { h_111_min_w <- h_111_min$objval*(-1) } else {
+ h_111_min_w <- h_111_min$objval }
```

Dann lautet das endgültige Ergebnis folgendermaßen:

```
> h_111_min_w
[1] 10
```

Die Untergrenze des Intervalls für h_{111} lautet nun 10. Das wären im Beispiel minimal 10 Frauen, die den echten Wirkstoff verabreicht bekommen haben, und die Therapie erfolgreich beendet haben.

Wie bereits angedeutet, wurden die Berechnungen der anderen Einzelhäufigkeiten äquivalent dazu durchgeführt (R-Code, siehe im Anhang A.2.2.2 bzw. auf der Daten-CD).

Im folgenden Output sind noch die Ergebnisse der Berechnungen für die weiteren Einzelhäufigkeiten aufgelistet:

```
> list (h_111 , h_112 , h_121 , h_122 , h_211 , h_212 , h_221 , h_222)
[[1]]
[1] "h_111 = [ 10 , ..., 25 ]"
[[2]]
[1] "h_112 = [ 9 , ..., 24 ]"
[[3]]
[1] "h_121 = [ 0 , ..., 15 ]"
[[4]]
[1] "h_122 = [ 1 , ..., 16 ]"
[[5]]
[1] "h_211 = [ 5 , ..., 20 ]"
[[6]]
[1] "h_212 = [ 15 , ..., 30 ]"
[[7]]
[1] "h_221 = [ 0 , ..., 15 ]"
[[8]]
[1] "h_222 = [ 0 , ..., 15 ]"
```

Die errechneten Intervalle der Einzelhäufigkeiten werden nun in einer Partialtabellen-Darstellung visualisiert, vgl. Tabelle 2.2 auf Seite 10.

$k=1$		Y		
		1	2	$h_{i\cdot 1}$
X	1	$h_{111} = [10,\ldots,25]$	$h_{121} = [0,\ldots,15]$	$h_{1\cdot 1} = 25$
	2	$h_{211} = [5,\ldots,20]$	$h_{221} = [0,\ldots,15]$	$h_{2\cdot 1} = 20$
	$h_{\cdot j1}$	$h_{\cdot 11} = 30$	$h_{\cdot 21} = 15$	$h_{\cdot\cdot 1} = 45$

$k=2$		Y		
		1	2	$h_{i\cdot 2}$
X	1	$h_{112} = [9,\ldots,24]$	$h_{122} = [1,\ldots,16]$	$h_{1\cdot 2} = 25$
	2	$h_{212} = [15,\ldots,30]$	$h_{222} = [0,\ldots,15]$	$h_{2\cdot 2} = 30$
	$h_{\cdot j2}$	$h_{\cdot 12} = 39$	$h_{\cdot 22} = 16$	$h_{\cdot\cdot 2} = 55$

Tabelle 2.8: Ergebnisse in *Partialtabellen*-Darstellung der dreidimensionalen $I \times J$ - Kontingenztafel, K Ebenen

Diese Darstellungsform, und nicht z.B. die Marginaltabellen-Darstellung, ist in diesem Fall die sinnvollste, da hier natürlich auch die 2-dimensionalen Randhäufigkeiten zu sehen sein sollen.
Vergleichbar dazu könnte man die Ergebnisse, analog zu Tabelle 2.8, auch in den Partialtabellen darstellen, die nach j und i aufgespalten sind, siehe Anhang A.1 und A.2, Seite 65.

2.3 Interpretation und Diskussion der Ergebnisse

In der Anwendung der linearen Optimierung (Kap. 2.2.2), anhand des Szenarios „Wirkstoff-Placebo-Effekt“, sind nur 1-dimensionalen Randhäufigkeiten als Restriktionen aufgenommen worden. Die Einzelhäufigkeiten sind durch die Randhäufigkeiten bedingt. Der Optimierungsalgorithmus errechnet für jedes Maximum und Minimum (Extrempunkte) der jeweiligen Häufigkeit eine Verteilung aller Einzelhäufigkeiten. Das hat zur Folge, dass unterschiedlichste Verteilungen der Einzelhäufigkeiten generiert werden. Siehe Anhang A.2.1.2 und Daten-CD, Datei `Kap2_Lin_Opt_CodePlus Ergebnisse_CK_20100813`.

Mit den errechneten Intervallen der Einzelhäufigkeiten kann nicht ohne Weiteres weiter gerechnet werden. Zum Beispiel können mit den Intervallen nicht direkt Odds Ratios berechnet werden, da für jeden Eintrag des zuvor berechneten Intervalls, z.B. von h_{111}, mehrere Verteilungen der korrespondierenden Einzelhäufigkeiten existieren. Das bedeutet, dass beispielsweise mehrere Verteilungen der korrespondierenden Einzelhäufigkeiten für konstantes $h_{111} = 15$ (willkürlich aus errechnetem Intervall gewählt, siehe Seite 24) gebildet werden können und natürlich viele weitere für andere Werte von h_{111}. Man kann auch das Intervall einer anderen Einzelhäufigkeit hernehmen und verfährt ebenso.

In der Anwendung der 2-dimensionalen Fréchet Bounds (Kap. 2.2.3) ist in den Lösungsalgorithmus der linearen Optimierung die 2-dimensionale Randverteilung zusätzlich zu der 1-dimensionalen Randverteilung mit eingebunden worden.

Die Ergebnisse der linearen Optimierung mit den zusätzlichen 2-dimensionalen Randhäufigkeiten ergaben in diesem Fall wesentlich kleinere Intervalle für jede Einzelhäufigkeit. Dies war zu erwarten, da die zusätzlichen Bedingungen einschließlich der Konsistenzbedingungen die Ergebnismöglichkeiten der linearen Optimierung von vornherein einschränken. Man stellt fest, dass die berechneten Intervalle **Teil**intervalle, oder Teilmengen, der vorherigen Anwendung (Kap. 2.2.2) sind. Die Einzelhäufigkeiten sind sozusagen „genauer“ geworden.

Darüber hinaus haben alle Intervalle die **gleiche** Breite, d.h. es sind in allen Intervallen gleich viel Elemente enthalten, im Gegensatz zu der Berechnung ohne zusätzliche 2-dimensionale Randverteilung. Es können gerade diese zusätzlichen Bedingungen (Randhäufigkeiten) als Ursache für die Verkleinerung und die gleiche Breite der Intervalle interpretiert werden. Denn wenn man sich den 3D-„Würfel“ auf Seite 9 und die Partialtabellen auf Seite 10 ansieht, fällt auf, dass die 1-dimensionalen Randhäufigkeiten „nur“ in je einer Ecke des „Würfels“ bzw. der Partialtabellen stehen und die 2-dimensionalen Randhäufigkeiten einen „direkteren“ Einfluss auf die Einträge der Kontingenztafel haben. Die 1-dimensionalen Randhäufigkeiten haben demnach weniger „direkten“ Einfluss auf die Einzelhäufigkeiten, sondern eher „nur“ durch die Gesamtgleichungen der Einzelhäufigkeiten, siehe Formeln (2.1) ohne Summation der 2-dimensionalen Randhäufigkeiten. Dadurch sind die nur durch 1-dimensionale Randverteilung errechneten Einzelhäufigkeiten der Intervalle „flexibler“ aber natürlich auch ungenauer.

Die 2-dimensionalen Randhäufigkeiten summieren sich nur je aus zwei Einzelhäufigkeiten

(dies ändert sich natürlich mit mehr Kategorien pro Merkmal), wobei natürlich trotzdem zwischen den 2-dimensionalen Randhäufigkeiten ebenfalls Abhängigkeiten bestehen, da die einzelnen Häufigkeiten Teil mehrerer 2-dimensionaler Randhäufigkeiten sind.

Insgesamt ist festzustellen, dass aus mehr Restriktionen eine höhere Genauigkeit und dadurch weniger Flexibilität, also starrere und kleiner Intervalle, resultieren. Dieses Phänomen findet man in der Physik, Mathematik und Statistik häufiger: je mehr und je stärker die Annahmen sind, desto genauer scheinen/werden die Ergebnisse. Da durch viele Annahmen aber die Realität oft außer Acht gelassen wird, sind die Resultate oft auch nicht näher an der Realität.

Auf diese Arbeit bezogen, dürfte die Realitätsnähe, die willkürliche Wahl der gegebenen Randverteilungen außen vor gelassen, bzgl. mehr oder weniger Bedingungen nicht beeinträchtigt sein, da die Bedingungen schon in der angegebenen Gesamthäufigkeit gegeben ist und nur darüber variiert wird, welche Bedingungen man verwendet oder nicht.

Kapitel 3

Kostenanwendung in der Medizin anhand einer $2 \times 2 \times 2$ Kontingenztafel

3.1 Kosten bei Diabetes-Erkrankungen

In diesem Kapitel wird ein medizinisches Kosten-Szenario vorgestellt und Berechnungen von Kosten durchgeführt. Es werden Kostenintervalle anhand von Randverteilungen durch lineare Optimierung berechnet. Die Sachverhalte sind modellhaft und haben nicht den Anspruch, die Realität abzubilden. Alle Zahlen sind willkürlich gewählt. Das Kapitel soll vor allem die Verwendung der linearen Optimierung und die Möglichkeiten, die sich daraus ergeben, verdeutlichen.

Im Folgenden wird eine kurze Definition von Diabetes gegeben, frei nach dem Artikel „Diabetes-Arten“ Lilly Deutschland GmbH [2009].

Diabetes mellitus oder **Zuckerkrankheit** bezeichnet eine Gruppe von Stoffwechselerkrankungen. Diabetes wird in unterschiedliche Typen unterteilt. In dieser Arbeit sind nur die für das Szenario benötigten Typen aufgeführt und beschrieben.

Diabetes mellitus Typ I wurde früher auch als juveniler Diabetes-Typ bezeichnet, der oft genetische Ursachen hat und bereits vor dem 40. Lebensjahr beginnt. Es wird bei den betroffenen Personen nicht genügend oder gar keine Insulin produziert. Die Ursache wird auf eine Autoimmunerkrankung zurückgeführt. Wichtige körpereigene Antigene werden fälschlicherweise vom Immunsystem als „fremd“ erkannt. Sie lösen wiederum eine Autoimmunreaktion aus, welche sich gegen die insulinproduzierenden Zellen richtet.

Der **Diabetes mellitus Typ II** wurde früher als „Altersdiabetes“ bezeichnet. Der Körper produziert in der Regel zwar genug Insulin, aber er kann das Insulin nicht richtig nutzen. Man nennt dies auch Insulinresistenz. Mit der Zeit produziert der Körper weniger Insulin und dies führt letztendlich auch zu einem Insulinmangel. Im Gegesatz zum Diabetes Typ I tritt der Diabetes Typ II meist bei übergewichtigen Menschen auf und ist weltweit am häufigsten verbreitet (ca. 90% der an Diabetes Erkrankten). Ursachen für

Diabetes Typ II sind familiäre Veranlagung, aber auch Übergewicht und Bewegungsmangel. Weitere Folgen im Alter sind beim Typ II Begleit- und Folgeerkrankungen, wie z.B. Bluthochdruck, Gefäßerkrankungen und Erkrankungen der Augen, Nieren und Nerven.

3.1.1 Rahmenbedingungen

In dieser (fiktiven) medizinischen Studie werden die Behandlungskosten (pro Jahr) für Diabetes-Patienten gegeben die Randverteilung ihrer Merkmale wiederum mit der linearen Optimierung bestimmt. Dies könnte für Prognosekalkulationen der Jahreskosten in Krankenhäusern beispielsweise sinnvoll sein, da eventuell nicht alle Daten aller Patienten bekannt sind oder bekannt gemacht werden dürfen.

Als Merkmale sind folgende vorgesehen: Merkmal X repräsentiert das mögliche „Übergewicht“ von teilnehmenden Patienten mit Ausprägungen „nein“ (Kodierung = 0) und „ja“ (Kodierung = 1). In Merkmal Y ist die „Diabetes“-Erkrankung festgehalten mit dem „Typ I“ (Kodierung = 0) und „Typ II“ (Kodierung = 1). Das dritte Merkmal Z bezeichnet das „Alter“ mit den gruppierten Ausprägungen „< 60“ Jahre (Kodierung = 0) und „≥ 60“ Jahre (Kodierung = 1).

Um einen guten Überblick zu schaffen, werden die Eigenschaften für die einzelnen Merkmale im Folgenden tabellarisch dargestellt.

Merkmal	**Bezeichnung**	**Ausprägungen**	**Kodierung**
X	Übergewicht	$i = 1$	„nein“ $\widehat{=} 0$
		$i = 2$	„ja“ $\widehat{=} 1$
Y	Diabetes	$j = 1$	„Typ I“ $\widehat{=} 0$
		$j = 2$	„Typ II“ $\widehat{=} 1$
Z	Alter	$k = 1$	„< 60“ $\widehat{=} 0$
	(in Jahren)	$k = 2$	„≥ 60“ $\widehat{=} 1$

Tabelle 3.1: Rahmenbedingungen für Kostenberechnung

Die Kostenverteilung wird nun willkürlich festgelegt. Es wird allerdings versucht, Effekte in der Realität nachzuahmen, z.B. dass durch Übergewicht die Kosten von Diabetes steigen, da Begleit- bzw. Folgeerkrankungen auftreten können oder dass im Alter ebenfalls die Kosten wegen Begleiterkrankungen steigen. Dies ist aber nur eine Anlehnung und hier nicht wissenschaftlich untermauert. Es gilt weiterhin, dass die Berechnungen modellhaft durchgeführt werden und der Einsatz der linearen Optimierung (zur Berechnung der Kosten“-intervalle“) ein Ziel dieser Arbeit ist.

Die Behandlungskosten (pro Jahr in €) für Patienten mit den einzelnen Merkmalskombinationen, c_{ijk} mit $i = 1, 2,\ j = 1, 2$ und $k = 1, 2$, werden in der Tabelle 3.2 festgelegt (in Orientierung an [Köster et al., 2005]). Mit diesen Kosten c_{ijk} werden die neuen Zielfunktions-

c_{111}	$= 1200,-$	c_{211}	$= 1500,-$
c_{112}	$= 3100,-$	c_{212}	$= 3460,-$
c_{121}	$= 1550,-$	c_{221}	$= 1870,-$
c_{122}	$= 3340,-$	c_{222}	$= 3645,-$

Tabelle 3.2: Jährliche Kosten pro Patient mit Merkmalskombination, in €

vektoren $\overline{h_{ijk}}c^T$ gebildet. Beispielsweise sieht der Zielfunktionsvektor für c_{111} folgendermaßen aus:

$$\overline{h_{111}}c^T = \begin{pmatrix} c_{111} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1200 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8}.$$

Die Zielfunktionsvektoren der anderen Behandllungskosten sind im Anhang B.1.1.1 dargestellt. Des Weiteren sind, wie schon gesagt, nur die 1-dimensionalen Randhäufigkeiten und die Gesamthäufigkeit gegeben.
Seien

- Gesamthäufigkeit $h_{...} = 100$ an der Studie beteiligte Patienten,
- 1-dimensionale Randhäufigkeiten:

$$h_{1..} = 48\ ,\ h_{2..} = 52\ ,$$
$$h_{.1.} = 25\ ,\ h_{.2.} = 75\ ,$$
$$h_{..1} = 40\ ,\ h_{..2} = 60$$

Daraus ergibt sich der Gesamt-Bedingungsvektor b_G nach Seite 15 mit

$$b_G = \begin{pmatrix} b \\ -b \end{pmatrix} = \begin{pmatrix} 48 \\ 52 \\ 25 \\ 75 \\ 40 \\ 60 \\ -48 \\ -52 \\ -25 \\ -75 \\ -40 \\ -60 \end{pmatrix} \in \mathbb{Z}^{m,1}.$$

Die Gesamt-Bedingungsmatrix A_G und der Vektor w ändern ihr Aussehen nicht im Vergleich zu Seite 15.

3.1.2 Kostenberechnung mit 1-dimensionalen Randhäufigkeiten

Programmierung und Ergebnisse in R

Die Kostenberechnung wird, wie in Kapitel 2.2 die Intervalle der Einzelhäufigkeiten berechnet wurden, mit der linearen Optimierung durchgeführt. Dazu ist es, wie schon gezeigt, notwendig, in den Zielfunktionsvektoren die Behandlungskosten für jede Merkmalskombination aufzunehmen. Die Berechnungen werden hier beispielhaft für die Kosten für nicht übergewichtige Patienten unter 60 Jahren mit einer Diabetes Typ I-Erkrankung (Merkmalskombination (1 1 1)) durchgeführt.

In der Programmierung des Sachverhaltes werden zuerst die 1-dimensionalen Randhäufigkeiten angegeben bzw. definiert. Aus diesen werden äquivalent zu der Programmierung in Kapitel 2.2.2 der Bedingungsvektor b_G, die Bedingungsmatrix A_G und die für den linearen Optimierungsalgorithmus notwendigen Operatoren definiert und zugewiesen. Spätestens jetzt sollte das Paket „lpSolve“ geladen werden, wie schon in Kapitel 2.2.2 erläutert. Siehe Anhang B.1.2.1 bzw. Daten-CD.

Im ersten Abschnitt des folgenden Codes wird der Zielfunktionsvektor für die Obergrenze der Behandlungskosten der Merkmalskombination (1 1 1) zugewiesen.

```
> C_cost_h111max <- c( 1200, 0, 0, 0, 0, 0, 0, 0)
> C_cost_h111max_t <- t(C_cost_h111max)       # Zielfunktionsvektor c^T
> f.obj1 <- C_cost_h111max_t
> # Lin. Opt.-problem
> cost_h_111_max <- lp ("max", f.obj1, f.con, f.dir, f.rhs, int.vec=1:8)
```

Der Ausdruck der letzten Zeile des Code-Abschnitts stellt den Algorithmus der linearen Optimierung dar, vergleichbar mit den Codes in Kapitel 2.2 . Damit das Ergebnis der Verteilung, siehe nächste Code-Abschnitt, ganzzahlig bleibt, muss der Ausdruck `int.vec=1:8` in dem Algorithmus ergänzt werden. Durch ihn kann bestimmt werden, wieviele Variablen ganzzahlig sein sollen.

Der Funktionswert lautet hier 30000, was den Kosten (in €) für die Behandlung von maximal 25 nicht übergewichtigen Patienten unter 60 Jahren mit einer Diabetes-Erkrankung des Typ I entspricht.

```
> cost_h_111_max$objval
[1] 30000
> cost_h_111_max$solution
[1] 25  0 15  8  0  0  0 52
```

In den zwei letzten Zeilen des Abschnitts ist die Verteilung der 100 Patienten dargestellt, mit der die Funktion $\frac{}{h_{111}} c^T \cdot w$ für h_{111} maximiert wurde. Die Reihenfolge der Zahlen orientiert sich am Vektor w, siehe Seite 15.

Im nächsten Abschnitt wird zur Berechnung der Kostenintervalle für Patienten mit der Merkmalskombination (1 1 1) wieder der Zielfunktionsvektor definiert und dem Algorithmus

der linearen Optimierung zugewiesen. Wie im Theorieteil zur linearen Optimierung, Formel (2.6), erläutert, ist nun der Wert im Zielfunktionsvektor negativ.

```
> C_cost_h111min <- c( -1200, 0, 0, 0, 0, 0, 0, 0)
> C_cost_h111min_t <- t(C_cost_h111min)
> f.obj2 <- C_cost_h111min_t
>
> cost_h_111_min <- lp ("max", f.obj2, f.con, f.dir, f.rhs, int.vec=1:8)
```

In der letzten Zeile des Code-Abschnitts ist wieder der Algorithmus des linearen Optimierungsproblems angegeben.

```
> cost_h_111_min$objval
[1] 0
> cost_h_111_min$solution
[1]  0 17 31  0  8  0  1 43
```

Der errechnete Funktionswert entspricht dem Ergebnis für die Untergrenze der Behandlungskosten für Patienten mit diese Merkmalskombination und lautet 0, da die Verteilung der Häufigkeiten vom Lösungsalgorithmus so bestimmt wurde, dass die Zelle von h_{111} nicht besetzt ist.

Um die Negation, die weiter oben erwähnt wurde, rückgängig zu machen, muss das errechnete Ergebnis wieder mit (-1) multipliziert werden.

```
> if ( cost_h_111_min$objval < 0 ) { cost_h_111_min_w <- cost_h_111_min$objval*(-1)
+  } else { cost_h_111_min_w <- cost_h_111_min$objval }
> cost_h_111_min_w
[1] 0
```

Am Ergebnis ändert sich in diesem Fall dadurch natürlich nichts.

Durch den nächsten Programmierungsbefehl werden nun die errechneten Ergebnisse zu einem Intervall für die jährlichen Behandlungskosten der Patienten mit der Merkmalskombination (1 1 1) mit Ober- und Untergrenze zusammengesetzt.

```
> cost_h_111 <- paste("Kosten_h_111 = [",cost_h_111_min_w,", ...,",
+ cost_h_111_max$objval,"]")
```

Die Programmierung der Berechnung der Behandlungskosten von Patienten mit den anderen Merkmalskombinationen werden äquivalent fortgeführt, siehe Anhang B.1.2.1 bzw. Daten-CD.

Im folgenden Code-Abschnitt sind alle Ergebnisse der Kostenberechnung aufgeführt. Für die jährlichen Behandlungskosten von Patienten wurde für jede Merkmalskombination $(i\ j\ k)$ ein Intervall errechnet.

```
> # Liste der Intervalle
```

```
> list (cost_h_111 , cost_h_112 , cost_h_121 , cost_h_122 , cost_h_211 ,
+ cost_h_212 , cost_h_221 , cost_h_222)
[[1]]
[1] "Kosten_h_111 = [ 0 , ..., 30000 ]"
[[2]]
[1] "Kosten_h_112 = [ 0 , ..., 77500 ]"
[[3]]
[1] "Kosten_h_121 = [ 0 , ..., 62000 ]"
[[4]]
[1] "Kosten_h_122 = [ 0 , ..., 160320 ]"
[[5]]
[1] "Kosten_h_211 = [ 0 , ..., 37500 ]"
[[6]]
[1] "Kosten_h_212 = [ 0 , ..., 86500 ]"
[[7]]
[1] "Kosten_h_221 = [ 0 , ..., 74800 ]"
[[8]]
[1] "Kosten_h_222 = [ 0 , ..., 189540 ]"
```

Die höchsten möglichen Behandlungskosten sind bei den Patienten mit der Merkmalskombination (2 2 2) festzustellen.

```
> cost_h_222_max$objval
[1] 189540
> cost_h_222_max$solution
[1] 25  0 15  8  0  0  0 52
```

Das bedeutet, unter der vom Lösungsalgorithmus errechneten Verteilung, dass maximal 52 übergewichtige Patienten mit einer Diabetes Typ II-Erkrankung und einem Alter von über 60 Jahren Behandlungskosten von 189540,- € im Jahr verursachen.

3.1.3 Kostenberechnung mit 1- und 2-dimensionalen Randhäufigkeiten

Wie in Kapitel 2.2.3 werden hier der Bedingungsvektor b_G und die Bedingungsmatrix A_G um die 2-dimensionalen Randhäufigkeiten, sowohl als positive als auch als negierte Werte, erweitert. Die Darstellungen der Matrix A_G unterscheidet sich nicht von der Darstellung aus Kapitel 2.2.3 und ist im Anhang A.2.2.1 auf Seite 73 zu sehen. Der Vektor b_G ähnelt auch dem aus Kapitel 2.2.3, hat allerdings andere Zahlenwerte.

Folgende 2-dimensionalen Randhäufigkeiten werden zusätzlich zu den 1-dimensionalen in den Vektor b_G aufgenommen:

$h_{1\cdot 1} = 19$	$h_{\cdot 11} = 15$	$h_{11\cdot} = 14$
$h_{1\cdot 2} = 29$	$h_{\cdot 12} = 10$	$h_{12\cdot} = 34$
$h_{2\cdot 1} = 21$	$h_{\cdot 21} = 25$	$h_{21\cdot} = 11$
$h_{2\cdot 2} = 31$	$h_{\cdot 22} = 50$	$h_{22\cdot} = 41$

Tabelle 3.3: Werte der 2-dimensionalen Randhäufigkeiten (Kap. 3.1.3)

Die Darstellung ist im Anhang B.1.3.1 auf Seite 82 zu finden.

Programmierung und Ergebnisse in R

Nun werden hier beispielhaft die Intervalle der jährlichen Kosten für übergewichtige Patienten unter 60 Jahren mit einer Diabetes Typ II-Erkrankung (Merkmalskombination (2 2 1)) errechnet.

Im folgenden Abschnitt des Codes wird der Zielfunktionsvektor für die Obergrenze des Intervalls der Behandlungskosten der Mermalskombination (2 2 1) definiert und als Objekt für die lineare Optimierung zugewiesen. (2 2 1) beschreibt die Merkmalskombination „Übergewicht = ja“, „Diabetes = Typ II“ und „Alter < 60“.

```
> C_cost_h221max <- c( 0, 0, 0, 0, 0, 0, 1870, 0)
> C_cost_h221max_t <- t(C_cost_h221max)
> f.obj13 <- C_cost_h221max_t
>
> cost_h_221_max <- lp ("max", f.obj13, f.con2, f.dir2, f.rhs2, int.vec=1:8)
```

Der Ausdruck der letzten Zeile des Code-Abschnitts stellt den Algorithmus der linearen Optimierung dar. Damit das Ergebnis der Verteilung, siehe nächste Code-Abschnitt, ganzzahlig bleibt, muss wiederum der Ausdruck `int.vec=1:8` in dem Algorithmus ergänzt werden. Durch ihn kann bestimmt werden, wieviele Variablen ganzzahlig sein sollen.

Der Algorithmus der linearen Optimierung errechnet einen Funktionswert von 37400, der hier dem Ergebnis von 37400,- €.

```
> cost_h_221_max$objval
[1] 37400
> cost_h_221_max$solution
[1] 14  0  5 29  1 10 20 21
```

Die Verteilung der 100 Patienten, mit der die Kosten für die Merkmalskombination (2 2 1) maximiert wurden, zeigt die maximale Anzahl (= 20) der übergewichtigen Patienten, die jährliche Behandlungskosten von 37400,- € verursachen.

Die Untergrenze für das Kostenintervall wird im folgenden Code-Abschnitt berechnet. Natürlich muss die Negation wieder rückgängig gemacht werden, da zur Minimierung die negative Funktion maximiert wird, wie schon zuvor durch den Bedingungs-Befehl („if“ ...„else“).

```
> cost_h_221_min <- lp ("max", f.obj14, f.con2, f.dir2, f.rhs2, int.vec=1:8)
>
> cost_h_221_min$objval
[1] -18700
> cost_h_221_min$solution
[1]  4 10 15 19 11  0 10 31
>
> if ( cost_h_221_min$objval < 0 ) { cost_h_221_min_w <- cost_h_221_min$objval*(-1)
+ } else { cost_h_221_min_w <- cost_h_221_min$objval }
> cost_h_221_min_w
[1] 18700
>
> cost_h_221 <- paste("Kosten_h_221 = [",cost_h_221_min_w,", ...,",
+ cost_h_221_max$objval,"]")
```

Als Ergebnis errechnet sich eine Untergrenze der Kosten für die Behandlungen von 10 übergewichtigen Patienten unter 60 Jahren mit einer Diabetes-Erkrankung des Typ II von jährlich 18700,- €. Zum Ende des Abschnitts werden errechnete Ober- und Untergrenze zu einem Intervall zusammengefasst.

Nachfolgend sind alle Ergebnisse der Kostenberechnung aufgeführt. Für jede Merkmalskombination (i j k) wurde ein Intervall für die jährlichen Behandlungskosten errechnet.

```
> # Liste der Intervalle
> list (cost_h_111 , cost_h_112 , cost_h_121 , cost_h_122 , cost_h_211 ,
+ cost_h_212 , cost_h_221 , cost_h_222)
[[1]]
[1] "Kosten_h_111 = [ 4800 , ..., 16800 ]"
[[2]]
[1] "Kosten_h_112 = [ 0 , ..., 31000 ]"
[[3]]
[1] "Kosten_h_121 = [ 7750 , ..., 23250 ]"
[[4]]
[1] "Kosten_h_122 = [ 63460 , ..., 96860 ]"
[[5]]
[1] "Kosten_h_211 = [ 1500 , ..., 16500 ]"
[[6]]
```

```
[1] "Kosten_h_212 = [ 0 , ..., 86500 ]"
[[7]]
[1] "Kosten_h_221 = [ 18700 , ..., 37400 ]"
[[8]]
[1] "Kosten_h_222 = [ 76545 , ..., 112995 ]"
```

Die Intervalle der Behandlungskosten der jeweiligen Merkmalskombinationen sind im Vergleich zu Kapitel 3.1.2 wesentlich schmäler geworden. Als Ursache sind wieder die zusätzlichen 2-dimensionalen Randhäufigkeiten anzuführen. Die Intensität der Verkleinerung der Intervalle hängt auch unmittelbar von der Beschaffenheit der 1- und 2-dimensionalen Randverteilung ab. Hier wurden diese direkt beeinflusst, da sie willkürlich festgelegt wurden.

3.2 Gesamtkosten bei Diabetes-Erkrankungen

In diesem Kapitel wird die Berechnung der Intervalle für die jährlichen Gesamtkosten der Diabetes-Erkrankten durchgeführt.

Hierzu ist es im Unterschied zum Kapitel 3.1 notwendig, in den Zielfunktionsvektoren `C_cost_...` alle angesetzten Kosten an die entsprechende Position zu setzen. Damit werden die Intervalle für die jährlichen Gesamtkosten („Gesamt-“ $\widehat{=}$ alle Merkmalskombinationen) berechnet.

3.2.1 Gesamtkostenberechnung mit 1-dimensionalen Randhäufigkeiten

Die Rahmenbedingungen sind gleich den Rahmenbedingungen des vorherigen Kapitels 3.1.1.

Programmierung und Ergebnisse in R

Zu Beginn werden die schon erläuterten Vorbereitungen, wie das Laden des „lpSolve“-Pakets, die Definition des Bedingungsvektors, der Bedingungsmatrix und die Objektzuweisung, programmiert.

Anschließend wird der eben erwähnte Zielfunktionsvektor definiert und dem Algorithmus für die lineare Optimierung zugewiesen.

```
> # Max und Min cost
> # Max cost
> C_cost_max <- c( 1200, 3100, 1550, 3340, 1500, 3460, 1870, 3645)
> C_cost_max_t <- t(C_cost_max)       # Zielfunktionsvektor c^T
> f.objG1 <- C_cost_max_t
> # Lin. Opt.-problem
> cost_max <- lp ("max", f.objG1, f.con, f.dir, f.rhs, int.vec=1:8)
```

In der letzten Zeile dieses Code-Abschnitts ist wieder der Lösungsalgorithmus zu implementieren.

Die Obergrenze des Intervalls für die jährlichen insgesamten Behandlungskosten des dargelegten Sachverhalts beträgt hier 274040,- €, (Funktionswert ist 274040).

```
> cost_max$objval
[1] 274040
> cost_max$solution
[1]  0  0 13 35  0 25 27  0
```

In der Zeile darunter ist die vom Algorithmus ermittelte Verteilung der Häufigkeiten zu sehen.

Mit dem negierten Zielfunktionsvektor und dem Lösungsalgorithmus wird nun die Untergrenze des Intervalls für die jährlichen Gesamtkosten der Behandlungen berechnet.

```
> # Min cost
> C_cost_min <- c( -1200, -3100, -1550, -3340, -1500, -3460, -1870, -3645)
> C_cost_min_t <- t(C_cost_min)
> f.objG2 <- C_cost_min_t
>
> cost_min <- lp ("max", f.objG2, f.con, f.dir, f.rhs, int.vec=1:8)
```

Der Funktionswert lautet für das Minimum -269385.

```
> cost_min$objval
[1] -269385
> cost_min$solution
[1]  0  0 15 33 25  0  0 27
```

Dieser Wert wird im nächsten Code-Abschnitt wieder mit (-1) multipliziert, um die bereits bekannte Negation rückgängig zu machen.

Das Ergebnis der Untergrenze des Intervalls für die jährlichen Gesamtkosten des dargelegten Sachverhalts beträgt nun 269385,- €.

```
> if ( cost_min$objval < 0 ) {cost_min_w <- cost_min$objval*(-1)
+  } else { cost_min_w <- cost_min$objval }
> cost_min_w
[1] 269385
```

Die errechneten Werte werden nun wieder zu einem Intervall zusammengefügt, das wie folgt aussieht.

```
> cost <- paste("Kosten = [",cost_min_w,", ...,", cost_max$objval,"]")
> print ( cost )
[1] "Kosten = [ 269385 , ..., 274040 ]"
```

Es ist zu erkennen, dass das Intervall nicht ganz so breit ist, wie die vorher ausgerechneten Kostenintervalle. Eine Interpretationsmöglichkeit ist, dass bei dieser Gesamtkostenberechnung sowohl bei der Ober- als auch Untergrenze sehr ähnliche Häufigkeitsverteilungen vom Algorithmus ermittelt wurden. Mit ähnlichen Häufigkeitsverteilungen kommen auch ähnliche bzw. nicht allzu weit von einander entfernte Ergebnisse zustande. Die Differenz der Ober- und Untergrenze des vorliegenden Intervalls liegt hier bei 4655,- €. Im Vergleich zu den immensen Kosten ist das nicht sehr viel.

Bei der nachfolgenden Berechnung mit zusätzlichen 2-dimensionalen Randhäufigkeiten als Bedingungen ist zu erwarten, dass das errechnete Intervall noch enger wird.

3.2.2 Gesamtkostenberechnung mit 1- und 2-dimensionalen Randhäufigkeiten

Wie in Kapitel 3.1.3 werden hier äquivalent die 2-dimensionalen Randhäufigkeiten in die Restriktionsgleichungen, also in den Bedingungsvektor (b_G siehe Anhang B.1.3.1) und die Bedingungsmatrix A_G aufgenommen. Die 2-dimensionale Randverteilung aus Kapitel 3.1.3 ist exakt die gleiche wie hier und wird deshalb nicht nochmal dargestellt.

Programmierung und Ergebnisse in R

Zuerst werden die üblichen Vorbereitungen, wie das Laden des „lpSolve"-Pakets, die Definition des Bedingungsvektors, der Bedingungsmatrix und die Objektzuweisung, vorgenommen.

Anschließend wird der Zielfunktionsvektor definiert und dem Algorithmus für die lineare Optimierung zugewiesen.

```
> # Max und Min cost
> # Max cost
> C_cost_max <- c( 1200, 3100, 1550, 3340, 1500, 3460, 1870, 3645)
> C_cost_max_t <- t(C_cost_max)       # Zielfunktionsvektor c^T
> f.objG1 <- C_cost_max_t
> # Lin. Opt.-problem
> cost_max <- lp ("max", f.objG1, f.con2, f.dir2, f.rhs2, int.vec=1:8)
```

In der letzten Zeile dieses Code-Abschnitts stellt wieder den Lösungsalgorithmus dar.

Die Obergrenze des Intervalls für die jährlichen Gesamtkosten für Behandlungen des dargelegten Sachverhalts beträgt hier 271455,- €, (Funktionswert ist 271455).

```
> cost_max$objval
[1] 271455
> cost_max$solution
[1] 14  0  5 29  1 10 20 21
```

In der letzten Zeile dieses Code-Abschnitts ist die vom Algorithmus ermittelte Verteilung der Häufigkeiten zu sehen.

```
> # Min cost
> C_cost_min <- c( -1200, -3100, -1550, -3340, -1500, -3460, -1870, -3645)
> C_cost_min_t <- t(C_cost_min)
> f.objG2 <- C_cost_min_t
>
> cost_min <- lp ("max", f.objG2, f.con2, f.dir2, f.rhs2, int.vec=1:8)
```

Mit dem negierten Zielfunktionsvektor und dem Lösungsalgorithmus wird nun die Untergrenze des Intervalls für die jährlichen Gesamtkosten berechnet.

Der Funktionswert für das Minimum lautet hier -270705. In der letzten Zeile ist wieder die ermittelten Häufigkeitsverteilungen angezeigt.

```
> cost_min$objval
[1] -270705
> cost_min$solution
[1]  4 10 15 19 11  0 10 31
```

Der errechnete Wert wird im nächsten Code-Abschnitt wieder mit (-1) multipliziert, um die bereits bekannte Negation rückgängig zu machen.

```
> if ( cost_min$objval < 0 ) {cost_min_w <- cost_min$objval*(-1)
+  } else { cost_min_w <- cost_min$objval }
> cost_min_w
[1] 270705
```

Nun beträgt das Ergebnis der Untergrenze des Intervalls für die jährlichen Gesamtkosten für die Behandlungen des dargelegten Sachverhalts 270705,- €.

Nach dem Zusammenfügen der errechneten Ober- unter Untergrenze entsteht nachfolgendes Intervall.

```
> cost <- paste("Kosten = [",cost_min_w,", ...,", cost_max$objval,"]")
> print ( cost )
[1] "Kosten = [ 270705 , ..., 271455 ]"
```

Die Differenz der Grenzen des Intervalls beträgt nun 750,- €. Das sind nur 16.11 % der Differenz aus den Berechnungen nur mit der 1-dimensionalen Randverteilung (siehe Seite 41). Die Breite des Intervalls hat sich also wesentlich verkleinert.

3.3 Interpretation und Diskussion der Ergebnisse

Dem Kapitel 3 lag ein weiteres medizinisches Szenario zugrunde. Für die Diabetes Typ I und Typ II-Erkrankungen wurden die verursachten Behandlungskosten modellhaft berechnet. Als weitere Merkmale wurden das Alter und das Körpergewicht miteinbezogen, da bereits Zusammenhänge zu Diabetes-Erkrankungen in anderen Abhandlungen nachgewiesen wurden. Das Alter wurde in zwei Klassen unterteilt, da nach Erfahrungen in der Medizin vor allem ab einem Alter von 60 Jahren durch Begleit- und Folgeerkrankungen, aufgrund der Diabetes-Erkrankung, erheblich höhere Kosten anfallen. In der ersten Anwendung (Kap. 3.1.2) waren die Gesamthäufigkeit und die 1-dimensionale Randverteilung bekannt, in der zweiten Anwendung (Kap. 3.1.3) zusätzlich die 2-dimensionale Randverteilung.

Die für die Behandlungen der einzelnen Merkmalskombinationen angesetzten Kosten wurden in dieser Arbeit willkürlich gewählt. Eine alternative Methode, die Kosten anzusetzen, wäre eine Kostenfestsetzung für jede Merkmalsausprägung und nicht für jede Merkmalskombination.

Zum Beispiel könnte man die Kosten in Abhängigkeit zum Alter als Basiskosten festlegen. Die Kosten für Diabetes und Übergewicht könnten als Zusatzkosten definiert werden. Nun würden die Kosten der einzelnen Merkmalskombinationen c_{ijk} berechnet indem die einzelnen Kosten summiert werden.

Dies alleine würde allerdings bedeuten, dass man implizit von Unabhängigkeit (der Merkmale) ausgeht. Als Korrektur könnte ein zusätzlicher Faktor ν_{ijk} als Gewichtung für jede Merkmalskombination eingeführt werden. Diese Gewichtungsfaktoren müssten mit den jeweiligen vorläufigen Kosten c_{ijk} multipliziert werden. Ergebnis wäre c^*_{ijk} als festzulegende Kosten für die Behandlung jeder Merkmalskombination. Der Gewichtungsfaktor wäre für jede Merkmalskombination $(i\ j\ k)$ unterschiedlich anzusetzen. Zusammenhangsmaße könnten eine Orientierung bzgl. der Zusammenhänge zwischen den Merkmalen geben und die Gewichtungsfaktoren eventuell geschätzt werden. Mit den festgelegten Kosten c^*_{ijk} könnten dieselben Berechnungen wie in diesem Kapitel (mit absolut willkürlich angesetzten Kosten) durchgeführt werden.

Die Methode der Gewichtungsfaktoren wurde hier nicht angewendet, da die Gewichtungsfaktoren sich im besten Fall auch auf statistische Erkenntnisse gründen sollte. Dazu müssten Zusammenhangsanalysen der einzelnen Merkmale als eine Art Rechengrundlage vorliegen.

Für den Aufbau des Algorithmuses der linearen Optimierung wurde ebenso verfahren wie in Kapitel 2.2 . Einzig in den Zielfunktionsvektoren wurden statt dem Wert „1“ die angesetzten Kosten aufgenommen, vgl. Seite 22 und Anhang A.2.1.1. In den ersten beiden Anwendungen (Kap. 3.1.2 und 3.1.3) des 3. Kapitels wurden die Intervalle der jährlichen Behandlungskosten der Patienten mit den jeweiligen Merkmalskombinationen berechnet, zum einen nur mit der 1-dimensionalen Randverteilung und zum anderen zusätzlich mit der 2-dimensionalen Randverteilung.

Ergebnis waren die Intervalle der jährlichen Kosten der Behandlungen für die Patienten mit der jeweiligen Merkmalskombination und die dazu bestimmte jeweilige Verteilung der

Einzelhäufigkeiten (der Patienten). Hier ist, wie in Kapitel 2.2, zu beachten, das verschiedenste Verteilungen der Einzelhäufigkeiten generiert wurden. Da das willkürliche Ansetzen der Kosten an die Erfahrung der Medizin angelehnt war, können die errechneten Kostenintervalle durchaus interpretiert werden.

Aus den Ergebnissen geht hervor, dass das Alter ab 60 (Jahren) oder das Übergewicht oder beides in Kombination sehr hohe jährliche Behandlungskosten verursachen. Allerdings ist beim Vergleich der Intervalle zu beachten, dass die verschiedenen Verteilungen der Einzelhäufigekeiten zugrunde liegen und so die Werte nicht direkt statistisch vergleichbar sind.

Zum Beispiel sind in der ersten Anwendung (Kap. 3.1.2) die Obergrenze der (jährlichen) Behandlungskosten für die Patienten mit den Merkmalskombinationen (1 1 1) und (2 2 2), siehe Anhang B.1.2.1, nicht direkt zu vergleichen. Bei (1 1 1) wurden nur die Kosten für die Behandlung von 25 Patienten und bei (2 2 2) die Kosten für die Behandlung von 52 Patienten berechnet. Also kann man sie direkt nicht vergleichen, man müsste zu erst anpassen, indem man z.B. fiktiv die Grundkosten von 3000,- € c_{111} mal 52 nimmt. (Ergebnis wäre 62400,- €). Nun könnte man die Kosten mit denen von (2 2 2) (Ergebnis war 189540,- €) vergleichen.

Die errechneten Kostenintervalle kann man jedoch insofern vergleichen, dass bei (1 1 1) beispielsweise nicht mehr als 25 Patienten in der Erhebung sind. Das heißt für das Gesundheitswesen, da die Kosten bezahlt werden müssen, spielt es keine Rolle, ob die Höhe der Behandlungskosten untereinander „statistisch richtig“ verglichen werden können, sondern einfach nur die Höhe der Kosten, die von diesen übernommen werden müssen. Dann steht fest, für die übergewichtigen Patienten über 60 Jahre mit Diabetes Typ II-Erkrankung muss mehr Geld ausgegeben werden als für nicht übergewichtige Patienten unter 60 mit Diabetes Typ I-Erkrankung. Zum einen, weil es bei den als zweiten genannten weniger Patienten gibt, weil die Einzelkosten ebenfalls niedriger sind und die grundsetzlichen Behandlungskosten nicht durch zusätzliche Begleiterkrankungen etc. in die Höhe getrieben werden.

In der zweiten Anwendung dieses Kapitels (Kap. 3.1.3) wurden dieselben Berechnungen durchgeführt wie in Kapitel 3.1.2. Allerdings wurden hier die 2-dimensionale Randverteilung als zusätzliche Restriktion in die lineare Optimierung mit aufgenommen. Vergleichbar zu der Anwendung in Kapitel 2.2.3 ergaben die Berechnungen wesentlich kleinere Intervalle der jährlichen Behandlungskosten als die Berechnungen mit der 1-dimensionalen Randverteilung als Restriktion. Die Intervalle sind allerdings weiterhin unterschiedlich breit.

Es lässt sich weiterhin feststellen, dass die Kostenintervalle für die Behandlung von Patienten mancher Merkmalskombinationen nicht bei 0 beginnen, sondern bei einem höheren Wert. Dafür ist wiederum die jeweilige ermittelte Verteilung der Einzelhäufigkeiten verantwortlich.

Auch hier ruft die Merkmalskombination „Alter höher als 60“ und „Übergewicht = ja“ die maximal höchsten jährlichen Kosten hervor.

Im Kapitel 3.2 wurden in den beiden Anwendungen nicht die Intervalle der jährlichen Kosten für die Behandlung der Patienten mit den einzelnen Merkmalskombinatonen berechnet, sondern die Intervalle der jährlichen Gesamtkosten für die Behandlungen aller Patienten zu-

sammen. Dazu musste in den Zielfunktionsvektor alle pro Merkmalskombination angesetzten Kosten an die entsprechende Position des Vektors gesetzt werden, zu sehen in den R-Codes der Anwendung. Demzufolge wurde nur ein Intervall der jährlichen Gesamtkosten berechnet.

In der ersten Anwendung (Kap. 3.2.1) sehen die dazu vom Lösungsalgorithmus ermittelten Verteilungen der Einzelhäufigkeiten relativ ähnlich aus. Darauf beruht vermutlich, dass das Ergebnis ein nicht sehr breites Intervall ist. Bei insgesamt möglichen Kosten von rund 274000,- € hat das Intervall „nur“ ein Breite von 4655,- €. Das arithmetische Mittel des Intervalls (Die Ober- und Untergrenze summieren und durch zwei geteilt.) beträgt 271712,50 €. Die Breite des Intervalls von 4655,- € stellen etwa 1.71 % des arithmetischen Mittels dar.

In der zweiten Anwendung des Kapitels 3.2 wurden wiederholt die 2-dimensionalen Randhäufigkeiten mit in den Lösungsalgorithmus mit einbezogen, in der Erwartung, dass die Breite des Gesamtkostenintervalls sich noch verringert. Die Obergrenze des errechneten Intervalls lag diesmal bei 271455,- €, also etwa 3500,- € unter der Obergrenze der letzten Anwendung (Kap. 3.2.1) mit nur 1-dimensionalen Randhäufigkeiten als Restriktionen. Die Untergrenze liegt hier bei 270705,- €, etwa 1300,- € über der Untergrenze der letzten Anwendung (Kap. 3.2.1). Das heißt, das Intervall ist noch kleiner, also „genauer“, geworden und hat nur noch eine Breite von 750,- €. Das entspricht 16.11 % der Breite des vorherigen Intervalls. Der Unterschied zu den maximalen erreichbaren Gesamtkosten ist noch wesentlich größer. Die Breite des Intervalls stellt nur 0.2763 % des maximal zu erreichenden Wertes dar. Zum arithmetischen Mittel des Intervalls $((270705 + 271455)/2 = 271080)$ beträgt die Breite des Intervalls 0.2766 %.

Die Tatsache, dass das Verhältnis der Breite der Intervalle zur maximal (oder im arithmetischen Mittel) erreichbaren Höhe der Kosten relativ gering ist, könnte für die Kostenkalkulationen des Gesundheitswesens von Vorteil sein bzw. ein Anhaltspunkt. Es kann im Gesundheitswesen, z. B. in Krankenhäusern, durchaus sein, dass keine Einzeldaten zur Verfügung stehen, sei es aus Dokumentations- oder Datenschutzgründen. Kostenintervalle wären in diesem Fall eine Möglichkeit, um Kostenplanungen etc. zu erarbeiten.

Kapitel 4

Theorie und Anwendung der statistischen Auswertung in $2 \times 2 \times 2$ Kontingenztafeln

In diesem Kapitel werden die für die statistische Auswertung notwendigen Formeln erklärt und in der Anwendung an dem medizinischen Beispiel aus Kapitel 2.2 ausgeführt.

Es ist in Erinnerung zu rufen, dass weiterhin nur Randverteilungen bekannt sind. Das bedeutet, es wird hier versucht, die Formeln der jeweiligen statistischen Auswertung in die *konvexe Optimierung* zu implementieren, um wiederum Ober- und Untergrenzen für die Intervalle der einzelnen statistischen Maßzahlen oder Tests zu erhalten. Die Ergebnisse werden im Anschluss interpretiert.

Alle Formeln werden beispielhaft nach Z bedingt dargestellt und erläutert. Äquivalent dazu könnten die Formeln auch nach X oder Y bedingt werden. Die Formeln und die Interpretationen würden sich dementsprechend ändern.

4.1 Theorie zur Durchführung der statistischen Auswertung

Zur statistischen Auswertung in 2×2 Kontingenztafeln gibt es einige Zusammenhangsmaße und statistische Tests, die sich durch die Bedingung auf ein Merkmal auf 3 Dimensionen ausweiten lassen. Hier werden die wichtigsten von ihnen theoretisch abgehandelt.

Darunter wird der *Odds Ratio*, der χ^2-*Koeffizient* (und dessen Normierungen: *Kontingenzkoeffient* und *korrigierter Kotingenzkoeffizient*) und der ϕ-*Koeffizient* sowie der χ^2-*Vierfeldertest* und der *Exakte Test nach Fisher* sein. Das *Loglineare Modell*, ein multivariates Analyseverfahren, wird nicht behandelt.

4.1.1 Odds Ratio

Der *Odds Ratio* stellt ein Chancenverhältnis dar und ist ein Assoziationsmaß. Das heißt, diese statistische Maßzahl sagt etwas über die Stärke eines Zusammenhangs zweier Merkmale aus. Es werden zwei Odds, bzw. Chancen, miteinander verglichen. Man nennt den Odds Ratio auch relative Chance, Quotenverhältnis oder Kreuzproduktverhältnis.

		Merkmal Y		
		1	2	$h_{i\cdot}$
Merkmal X	1	h_{11}	h_{12}	$h_{1\cdot}$
	2	h_{21}	h_{22}	$h_{2\cdot}$
	$h_{\cdot j}$	$h_{\cdot 1}$	$h_{\cdot 2}$	$h_{\cdot\cdot}$

Tabelle 4.1: Beispiel zweidimensionale 2×2 Kontingenztafel

Der Odds Ratio wird mit den Werten aus einer zweidimensionalen Kontingenztafel (Tabelle 4.1) folgendermaßen berechnet:

$$\gamma = \frac{h_{11}/h_{12}}{h_{21}/h_{22}} = \frac{h_{11} \cdot h_{22}}{h_{21} \cdot h_{12}} \tag{4.1}$$

Für die Interpretation des Quotenverhältnisses gilt:

$\gamma = 1$ Chancen in beiden Populationen gleich

$\gamma > 1$ Chancen in Populationen $X = 1$ besser als in Population $X = 2$

$\gamma < 1$ Chancen in Populationen $X = 1$ schlechter als in Population $X = 2$.

Population $\widehat{=}$ Merkmal

Das Chancenverhältnis gibt somit an, welche der Populationen die besseren Chancen besitzt und um wieviel besser diese Chancen sind.

Um den Odds Ratio auf drei Merkmale (hier oft als Populationen bezeichnet) auszuweiten, bedingt man die Verhältnisse. Man sagt also etwas über die Chancen zweier Populationen aus unter der Bedingung, dass das dritte Merkmal z.B. $k = 1$ konstant ist. Die Darstellung der Kontingenztafeln in diesem Zusammenhang haben wir schon in der Tabelle 2.2 auf der Seite 10 gesehen. Es sind die Partialtabellen, die hier das Verständnis unterstützen.

Die Formel des bedingten Odds Ratio wird nun beispielhaft für das festgehaltene Merkmal $Z = k$ gezeigt. Es wird also das Chancenverhältnis von X und Y gegeben $Z = k$ betrachtet.

Dann hat das Chancenverhältnis die Form

$$\gamma\left(X,Y|Z=k\right) \;=\; \frac{P(X=1,Y=1|Z=k)/P(X=1,Y=2|Z=k)}{P(X=2,Y=1|Z=k)/P(X=2,Y=2|Z=k)} \;=$$

$$= \;\frac{f_{11|k}/f_{12|k}}{f_{21|k}/f_{22|k}} = \frac{\frac{h_{11|k}}{n_k}/\frac{h_{12|k}}{n_k}}{\frac{h_{21|k}}{n_k}/\frac{h_{22|k}}{n_k}} \;=\; \frac{h_{11|k}/h_{12|k}}{h_{21|k}/h_{22|k}} \;=\; \frac{h_{11|k}\cdot h_{22|k}}{h_{21|k}\cdot h_{12|k}} \qquad (4.2)$$

mit $k=1,2$ und $h_{..k}=n_k$. [Tutz, 2009]

P stellt die theoretische Wahrscheinlichkeit dar. Das heißt, $P(X=i,Y=j|Z=k)$ kann als die relative Häufigkeit $f_{ij|k}$ gesehen werden. Da sich die 1-dimensionalen Randhäufigkeiten $h_{..k}=n_k$ durch den Bruch kürzen, können auch die absoluten Häufigkeiten verwendet werden. Da in dieser Arbeit die Einzelhäufigkeiten Intervalle darstellen, werden die Odds Ratio für alle drei konstanten Merkmale/Populationen auch Intervalle als Wertebereiche haben. [Fahrmeir et al., 2004]

4.1.2 χ^2-Koeffizient, Kontingenzkoeffizient (Pearson), korrigierter Kontingenzkoeffizient

χ^2-Koeffizient

Der χ^2-Koeffizient ist ebenfalls ein statistisches Assoziationsmaß. Er misst die „Stärke“ des Zusammenhangs jedoch nicht die „Richtung“ des Zusammenhangs. Die Richtungsinformation verliert der Korrelationskoeffizient durch das Quadrieren.

Die Obergrenze, d.h. der Wert, den der χ^2-Koeffizient bei vollkommener Abhängigkeit der betrachteten Merkmale annimmt, ist abhängig von der Anzahl der Ausprägungen der Merkmale und der Stichprobengröße, d.h. der Gesamthäufigkeit $h_{...}=n$. Dadurch ist die Aussagekraft des χ^2-Koeffizienten relativ gering. Eine Vergleichbarkeit von Werten des χ^2-Koeffizienten über verschiedene Kontingenztabellen und Stichprobengrößen ist daher meist nicht möglich. Durch die Quadrierung bei steigender Gesamthäufigkeit, steigender Anzahl der Ausprägungen und steigender Stärke des Zusammenhangs wird der χ^2-Wert größer. Das heißt die Stichprobengröße und die Anzahl der Ausprägungen können den χ^2-Wert verfälschen. Bei völliger Unabhängigkeit der Merkmale ist $\chi^2=0$.

Durch die Dreidimensionalität in dieser Arbeit, wird auch hier der χ^2-Koeffizient bedingt durch Z betrachtet.

Für den χ^2-Koeffizienten werden zusätzlich zu den tatsächlich *beobachteten Häufigkeiten* die *zu erwartenden Häufigkeiten* (wenn Unabhängigkeit vorliegt) benötigt.

Anders ausgedrückt, der quadratische Unterschied wird zwischen tatsächlichen und zu erwartenden Häufigkeiten berechnet. Die zu erwartenden Häufigkeiten errechnen sich durch

das *Postulat der empirischen Unabhängigkeit*:

$$\begin{aligned} \frac{\tilde{h}_{ij|k}}{h_{i\cdot|k}} &= \frac{h_{\cdot j|k}}{n_k} \\ \Leftrightarrow \qquad \tilde{h}_{ij|k} &= \frac{h_{i\cdot|k} \cdot h_{\cdot j|k}}{n_k} \ . \end{aligned}$$

Somit ist der χ^2-Koeffizient bestimmt durch

$$\chi^2_k \ = \ \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(h_{ij|k} - \tilde{h}_{ij|k}\right)^2}{\tilde{h}_{ij|k}} \ , \qquad \chi^2_k \in [0, \infty[\ .$$

In einer $2 \times 2 \times 2$ Kontingenztafel lässt sich die Formel für den χ^2-Koeffizienten einfacher ausdrücken, vgl. Tabelle (2.2):

$$\chi^2_k \ = \ \frac{n_k \cdot \left(h_{11|k} \cdot h_{22|k} - h_{12|k} \cdot h_{21|k}\right)^2}{h_{1\cdot|k} \cdot h_{\cdot 1|k} \cdot h_{\cdot 2|k} \cdot h_{2\cdot|k}} \ , \qquad \chi^2_k \in [0, \infty[\ .$$

Die χ^2-Werte für die jeweils anderen „festgehaltenen“ Merkmale berechnen sich äquivalent. [Fahrmeir et al., 2004]

Kontingenzkoeffizient nach Pearson

Um die Abhängigkeit des Wertebereichs von der Dimension der Kontingenztabelle und der Stichprobengröße zu eliminieren, müssen zwei Normierungsschritte durchgeführt werden.
Durch den ersten Normierungsschritt erhält man den *Kontingenzkoeffizienten* (nach Pearson)

$$K_k \ = \ \sqrt{\frac{\chi^2_k}{n_k + \chi^2_k}}$$

mit dem Wertebereich $K_k \in \left[0, \sqrt{\frac{M-1}{M}}\ \right]$, wobei $M = \ \min\{i, j\}$.

Korrigierter Kontingenzkoeffizient

Der zweite Normierungsschritt führt zum *korrigierten Kontingenzkoeffizienten*

$$K^*_k \ = \ K_k / \sqrt{\frac{M-1}{M}} \ ,$$

mit dem Wertebereich $K^*_k \in [0, 1]$.

Wenn der K^*-Wert nahe 0 liegt, deutet dies auf Unabhängigkeit hin, bei einem Wert nahe 1 auf Abhängigkeit der Merkmale. [Fahrmeir et al., 2004]

4.1.3 ϕ-Koeffizient

Der ϕ-Koeffizient, auch *Bravais-Pearson Korrelationskoeffizient* genannt, ist eigentlich nur für metrische Merkmale geeignet, allerdings stellen dichotome und binäre Merkmale eine Ausnahme dar. Für die Ausprägungen der jeweiligen Merkmale wählt man die Kodierung „0“ und „1“. Dies ist bei dem Anwendungsbeispiel dieser Arbeit schon von Anfang an definiert, siehe Seite 20.

Es wird wieder von der Annahme ausgegangen, dass durch Merkmal Z bedingt wird. Die durch X bzw. Y bedingten Berechnungen des ϕ-Koeffizienten sind äquivalent.

Im Unterschied zum χ^2-Koeffizienten gibt der ϕ-Koeffizient auch die Richtung des Zusammenhangs an, da hier **keine** Quadrierung stattfindet.

Die Formel lautet

$$\phi_k \;=\; \frac{h_{11|k} \cdot h_{22|k} - h_{12|k} \cdot h_{21|k}}{\sqrt{h_{1\cdot|k} \cdot h_{2\cdot|k} \cdot h_{\cdot 1|k} h_{\cdot 2|k}}} \;.$$

mit dem Wertebereich $\phi_k \in [-1, 1]$.

Der ϕ-Koeffizient ist eng verwandt mit dem χ^2-Koeffizienten, es gilt $\phi^2 \;=\; \frac{\chi^2}{n}$. Das bedeutet für die in dieser Arbeit nach Z bedingten ϕ-Koeffizienten, dass folgende Beziehung gilt:

$$\phi_k^2 \;=\; \frac{\chi_k^2}{n_k} \;.$$

Bei der Interpretation der resultierenden Werte muss man die Kodierung der Merkmale beachten. Ein negativer Wert deutet auf einen *gegensinnigen* Zusammenhang hin, ein positiver Wert auf einen *gleichsinnigen* Zusammenhang. [Fahrmeir et al., 2004]

4.1.4 χ^2-Vierfeldertest

Der χ^2-*Vierfeldertest* ist ein Spezialfall des χ^2-Tests. Generell untersucht man mit dem χ^2-Test verschiedene Verteilungseigenschaften einer statistischen Grundgesamtheit.
Folgende Tests unterscheidet man unter anderem:

- *Anpassungstest* oder *Verteilungstest*: Hier wird getestet, ob die vorliegenden Daten auf eine bestimmte Weise (vorgegeben Verteilung) verteilt sind.
  Man spricht auch von *Goodness-of-fit-Tests.*
- *Unabhängigkeitstest*: Hier wird getestet, ob zwei Merkmale stochastisch unabhängig sind.
- *Homogenitätstest*: Hier wird getestet, ob bei zwei oder mehr Stichproben die jeweiligen Verteilungen identisch sind.
- *Vierfeldertest*: Der Vierfeldertest ist ein Spezialfall des χ^2-Tests für 2×2 Kontingenztafeln.

In diesem Kapitel wird nun untersucht, ob zwei dichotome Merkmale, die durch ein drittes Merkmal bedingt sind, stochastisch unabhängig voneinander sind bzw. ob die Verteilung eines dichotomen Merkmals in zwei Gruppen identisch ist.

Es ist wiederholt zu beachten, dass nur die Randverteilung gegeben ist und die Einzelhäufigkeiten in Intervall-Form in die Berechnungen mit einfließen. Das bedeutet, die Formeln der Tests werden in einer, diesmal nicht linearen, sondern konvexen Optimierung maximiert. Die Optimierung ist nicht mehr linear, da die Formeln, die zu maximieren sind, nicht mehr linear, sondern quadratisch, fraktionell und somit konvex sind.

Zunächst ist zu beachten, dass die zu erwartenden Zellhäufigkeiten aller vier Felder mindestens 5 betragen sollten (Faustregel).

Die zu erwartenden Zellhäufigkeiten berechnet man mit

$$\frac{\text{Zeilensumme} \cdot \text{Spaltensumme}}{\text{Gesamtsumme}} .$$

Bedingt nach $Z = k$ lautet die Formel dann

$$\frac{h_{i\cdot|k} \cdot h_{\cdot j|k}}{h_{\cdot\cdot|k}} ,$$

mit $i = 1,2$, $j = 1,2$.

Wenn der Erwartungswert kleiner 5 ist, empfiehlt sich der *Exakte Test nach Fisher*, der im nächsten Kapitel beschrieben wird.

Der χ^2-Vierfelder Test sieht nun folgendermaßen aus:

χ^2-Vierfeldertest

Annahmen: Zwei unabhängige dichotome Merkmale $X,\ i = 1, 2$ und $Y,\ j = 1, 2$ bedingt durch ein drittes Merkmal $Z,\ k = 1, 2$, gruppiert in einer $(2 \times 2 \times 2)$-Kontingenztafel

Hypothese: H_0 :
$P(X = i, Y = j|Z = k) = P(X = i|Z = k) \cdot P(Y = j|Z = k)$,
$\forall i, j$

H_1 :
$P(X = i, Y = j|Z = k) \neq P(X = i|Z = k) \cdot P(Y = j|Z = k)$,
für mindestens ein Paar (i, j)

Teststatistik:

$$\chi^2_k = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(h_{ij|k} - \tilde{h}_{ij|k}\right)^2}{\tilde{h}_{ij|k}} =$$

$$= \frac{n_k \cdot \left(h_{11|k} \cdot h_{22|k} - h_{12|k} \cdot h_{21|k}\right)^2}{h_{1\cdot|k} \cdot h_{\cdot 1|k} \cdot h_{\cdot 2|k} \cdot h_{2\cdot|k}}$$

$$\text{mit } \tilde{h}_{ij|k} = \frac{h_{i\cdot|k} \cdot h_{\cdot j|k}}{n_k}$$

Verteilung unter H_0: approximativ χ^2-verteilt mit $((I-1)(J-1))$ Freiheitsgraden

Ablehnungsbereich: $\chi^2_k > \chi^2_{1-\alpha}((I-1)(J-1))$

4.1.5 Exakter Test nach Fisher

Der *Exakte Test nach Fisher* testet ebenso wie der χ^2-Vierfeldertest zwei dichotome Merkmale auf stochastische Unabhängigkeit. Da der χ^2-Test approximativ ist, kann er deshalb bei kleinen Stichprobenumfängen nicht angewendet werden. Der exakte Test nach Fisher auch bei kleine Stichprobenumfängen verwendet werden, weil durch die hypergeometrische Verteilung, die dem exakten Test nach Fisher zugrunde liegt, genaue Wahrscheinlichkeiten bestimmt werden können.

Der exakte Test bildet Kombinationen von Zellhäufigkeiten, die bei festen Zeilen- und Spaltensummen entstehen könnten, und berechnet die bedingte Wahrscheinlichkeit für die Zellhäufigkeiten, gegeben die Randhäufigkeiten.

Der Test ist folgendermaßen aufgebaut:

Exakte Test nach Fisher

Annahmen: Zwei unabhängige dichotome Merkmale X , $i = 1, 2$ und Y , $j = 1, 2$,
bedingt durch ein drittes Merkmal Z , $k = 1, 2$,
gruppiert in einer $(2 \times 2 \times 2)$-Kontingenztafel

Hypothese: H_0 :
$P(X = 1|Z = k) = P(X = 2|Z = k) = P$,

H_1 :
$P(X = 1|Z = k) \neq P(X = 2|Z = k)$

Teststatistik:

$$\varphi_k(h_{11|k}) \;=\; \varphi(h_{11|k}|Z = k) \;=\; \frac{\binom{h_{1\cdot|k}}{h_{11|k}} \cdot \binom{h_{2\cdot|k}}{h_{21|k}}}{\binom{h_{\cdot\cdot|k}}{h_{\cdot 1|k}}} \;=$$

$$= \frac{\binom{h_{11|k} + h_{12|k}}{h_{11|k}} \cdot \binom{h_{21|k} + h_{22|k}}{h_{21|k}}}{\binom{h_{\cdot\cdot|k}}{h_{11|k} + h_{21|k}}} = \cdots =$$

$$= \frac{(h_{11|k} + h_{21|k})! \cdot (h_{12|k} + h_{22|k})! \cdot (h_{11|k} + h_{12|k})! \cdot (h_{21|k} + h_{22|k})!}{h_{\cdot\cdot|k}! \cdot h_{11|k}! \cdot h_{12|k}! \cdot h_{21|k}! \cdot h_{22|k}!}$$

Verteilung unter H_0: hypergeometrisch verteilt mit $H(h_{\cdot 1|k}\ ,\ h_{1\cdot|k}\ ,\ h_{\cdot\cdot|k})$

Ablehnungsbereich: $\varphi_k(h_{11|k}) \;<\; \alpha/2$

[Toutenburg and Heumann, 2008], [Lübbert, 1999]

4.1.6 Konvexe Optimierung

Hier wird nur einleitend die konvexe Optimierung erläutert. Sie ist eine allgemeiner Form der linearen Optimierung und Teilgebiet der Optimierung in der angewandten Mathematik. Die konvexe Optimierung bildet die Grundlage für die lineare und nichtlineare Optimierung.

Die sogenannte Zielfunktion wird minimiert. Die Optimierung ist durch bestimmte Restriktionen bedingt und in Form von Gleichungen und Ungleichungen gegeben. Sind alle Nebenbedingungen erfüllt, so ist die Lösung des konvexen Optimierungsproblems zulässig.

Falls sowohl die Zielfunktion als auch die Menge der zulässigen Lösungen konvex sind, stellt dies ein konvexes Optimierungsproblem oder ein konvexes Programm dar.

In der Praxis sind viele Probleme konvex. Zum Beispiel wird häufig auf Quadern optimiert, die immer konvex sind, und Zielfunktion sind oft quadratische Funktionen, die unter bestimmten Voraussetzungen ebenfalls konvex sind.

Als Besonderheit der konvexen Optimierung ist anzuführen, dass jeder lokale Optimalwert auch ein globales Optimum ist. Das bedeutet, dass eine Lösung, die mindestens so gut ist wie alle anderen Lösungen in einer bestimmten Umgebung, auch mindestens so gut ist wie alle zulässigen Lösungen. Deshalb kann man einfach nach lokalen Optima suchen.

Im Folgenden sind zwei Beispiele für weiterführende Literatur für die intensivere Lektüre angeführt: Janzing [2004, Kap. 4], Boyd [2004].

4.2 Durchführung der statistischen Auswertung in R

In diesem Kapitel sollen nun die im Theorieteil 4.1 abgehandelten Zusammenhangsmaße und die erwähnten Tests beispielhaft angewendet werden. Zu Beginn von Kapitel 4 war von der konvexen Optimierung die Rede, die hier als „Werkzeug“ fungieren soll. Die konvexe Optimierung ist die Verallgemeinerung der linearen Optimierung und ist imstande nichtlineare Gleichungen zu optimieren. Nichtlineare Bedingungen sind ebenfalls erlaubt. Die vorangegangen Formeln der statistischen Auswertung sind allesamt entweder fraktionell (Odds Ratio), oder polynomial oder beides (alle anderen erläuterten Maßzahlen und Tests).

Für die konvexe Optimierung wird der im R-Paket „stats“ befindliche Algorithmus „constrOptim“ genutzt. Der Algorithmus minimiert eine Funktion, in diesem Fall die Formeln der jeweiligen statistischen Auswertungsmethode, unter linearen Ungleichheitsbedingungen.

4.2.1 Odds Ratio

Zuerst wird die relativ einfache Formel des Odds Ratio implementiert. Dazu die erste R-Programmierung.

Programmierung und Ergebnisse in R

Das R-Paket „stats“ ist bereits in den Standardeinstellungen von R geladen. Die Randhäufigkeiten und die Bedingungsoperatoren für die konvexe Optimierung werden definiert und zugewiesen.

```
> # Odds Ratio
> # Gegebene Randverteilung: 1-dim. Randhäufigkeiten
> h1.. <- 50
> h2.. <- 50
> h.1. <- 69
> h.2. <- 31
> h..1 <- 45
> h..2 <- 55
> h... <- 100
> # Bedingungsvektor b, Bedingungsmatrix A
> b <- c( -h1.. , -h2.. , -h.1. , -h.2. , -h..1 , -h..2)
> A <- matrix( nrow=6 ,ncol=8 , data= c( -1, 0, -1, 0, -1, 0,
+ -1, 0, -1, 0, 0, -1, -1, 0, 0, -1, -1, 0, -1, 0, 0, -1, 0, -1,
+ 0, -1, -1, 0, -1, 0, 0, -1, -1, 0, 0, -1, 0, -1, 0, -1, -1, 0,
+ 0, -1, 0, -1, 0, -1))
```

Hier wird erst für die Obergrenze des Odds Ratio-Intervalls programmiert. Da der Algorithmus minimiert, sind die Operatoren negativ.

Der Algorithmus benötigt **Startwerte** für die einzelnen Zellhäufigkeiten, die vorgegeben werden müssen. Dazu bietet sich für die Zellhäufigkeiten an, die unter stochastischer Unabhängigkeit zu erwartenden Häufigkeiten zu verwenden. Diese werden im folgenden Code-Abschnitt berechnet und beispielhaft dargestellt.

```
> # zu erwartende absolute Häufigkeiten
> h_111 <- (h1../h...)*(h.1./h...)*(h..1/h...)*h...
> h_112 <- (h1../h...)*(h.1./h...)*(h..2/h...)*h...
```

(...)

```
> h_111
[1] 15.525
> h_112
[1] 18.975
```

(...)

Als nächstes werden die zu erwartenden Häufigkeiten dem Vektor `theta1` zugewiesen und anschließend die Formel des Odds Ratio als Zielfunktion implementiert.

```
> # Objekt
> theta1 <- c(h_111, h_112, h_121, h_122, h_211, h_212, h_221, h_222)
> # zu minimierende Funktion
> OR_ij1 <- function(x) {
+     (x[1]/x[5])/(x[3]/x[7])
+ }
```

Dies ist die Funktion des Odds Ratio zu der Kontingenztafel 2.2 für $k = 1$.

Nach der R-Documentation von Gentleman and Ihaka [NA] ist eine zulässige Region definiert nach der Formel `A %*% theta1 - b >= 0`. Der Startwert muss innerhalb der zulässigen Region liegen. Dies wurde im Folgenden mit den vorhandenen Operatoren nachgerechnet.

```
> Atheta1 <- A %*% theta1 - b
> Atheta1
     [,1]
[1,]    0
[2,]    0
[3,]    0
[4,]    0
[5,]    0
[6,]    0
```

Die Gleichung ist dem Ergebnis nach erfüllt.

Nun wird der Algorithmus „constrOptim“ mit den Operatoren und der Zielfunktion versehen und berechnet. Der Anfangswert (initial value) ist nicht zulässig, obwohl im letzten Abschnitt gerade dies überprüft wurde. R meldet den folgenden Fehler:

```
> # Lösungsalgorthmus für Minimum
> OR_h_111_min <- constrOptim(theta1, OR_ij1, NULL, ui = A, ci = b)
Fehler in constrOptim(theta1, OR_ij1, NULL, ui = A, ci = b) :
        initial value not feasible
```

Wenn man sich, wie im folgenden Code-Abschnitt, den errechneten Output ausgeben lässt, ist folgendes zu sehen.

```
> OR_h_111_min
Fehler: objekt "OR_h_111_min" nicht gefunden
```

Das errechnete Minimum bzw. die Untergrenze für das Odds-Ratio-Intervall kann nicht gefunden werden. Dasselbe passiert beim Berechnen des Maximums bzw. der Obergrenze des Intervalls:

```
> # Lösungsalgorthmus für Maximum
> OR_h_111_max <- constrOptim(theta1, OR_ij1, NULL, ui = A, ci = b,
+ control=list(fnscale=-1))
Fehler in constrOptim(theta1, OR_ij1, NULL, ui = A, ci = b,
+ control = list(fnscale = -1)) :
        initial value not feasible
> OR_h_111_max
Fehler: objekt "OR_h_111_max" nicht gefunden
```

Um zu erfahren, ob es eine andere Möglichkeit gibt, wurde nun im nächsten Teil mit relativen Häufigkeiten gerechnet und zuerst die zu erwarteten relativen Häufigkeiten berechnet. Diese sind im Anhang C.1.1.2 auf Seite 88 zu finden.

Im Anschluss daran werden diesmal die zu erwartenden **relativen** Häufigkeiten dem Objekt für die konvexe Optimierung zu gewiesen. Darunter ist wieder die zu minimierende Zielfunktion zu sehen.

```
> # Objekt
> theta1 <- c(f_111, f_112, f_121, f_122, f_211, f_212, f_221, f_222)
> #zu minimierende Funktion
> OR_ij1 <- function(x) {
+     (x[1]/x[5])/(x[3]/x[7])
+ }
> # Test
> Atheta1b <- A %*% theta1 - b
> Atheta1b
      [,1]
[1,] 49.50
[2,] 49.50
[3,] 68.31
[4,] 30.69
[5,] 44.55
[6,] 54.45
```

Der Test ergibt diesmal eindeutig positive Ergebnisse, was bedeutet, dass eine zulässige Region definiert ist.

Nun wird der gleiche Lösungsalgorithmus wie vorhin angewendet. Diesmal erhält man dabei keine Fehlermeldung und folgenden Output.

```
> # Lösungsalgorithmus für Minimum
> OR_h_111_min <- constrOptim(theta1, OR_ij1, NULL, ui = A, ci = b)
> OR_h_111_min
$par
[1]  4.323484e+01  4.454786e-01  1.251559e-31  1.664349e-01  1.090387e-14
[6]  2.976455e-01 -6.942715e-01  1.708770e-01
$value
[1] -2.199537e+46
$counts
function gradient
     503       NA
$convergence
[1] 1
$message
NULL
$outer.iterations
[1] 51
$barrier.value
[1] 0
```

„par“ gibt „die beste Menge der gefundenen Parameter“ an, d.h. die Menge der einzelnen Positionen im Vektor theta1.

„value“ gibt den Wert der zu optimierenden Funktion, also des Odds Ratio, an. Dieser Wert ist hier nahezu $-\infty$.

Der Ausdruck „counts“ gibt die Anzahl der Funktionsaufrufe an.

„convergence“ gibt an, ob die Optimierung konvergiert. „0“ bedeutet „konvergiert“ und „1“ bedeutet, dass das Iterationslimit erreicht ist, und bis dahin keine Konvergenz ermittelt wurde.

Der „barrier.value“ gibt den Wert der Grenzfunktion am Optimum an. An „der besten Menge der gefundenen Parameter“ und dem errechneten Wert des Odds Ratios sieht man ‚dass die meisten Werte sehr kleine und vor allem der Odds Ratio sehr klein ist mit $-2.199537 \cdot 10^{46}$.

Das Ergebnis scheint zwar keine unzulässige Lösung zu sein, aber sie ist kein sinnvoll interpretierbares Ergebnis.

Nun wird noch die Obergrenze des Odds Ratio-Intervalls mit den relativen Häufigkeiten berechnet. Dabei kommt folgender Funktionswert heraus.

```
> # Lösungsalgorithmus für Maximum
> OR_h_111_max <- constrOptim(theta1, OR_ij1, NULL, ui = A, ci = b,
+ control=list(fnscale=-1))
> OR_h_111_max
$par
[1] 1.940832e-01 1.974475e-01 2.733515e-12 9.042863e-02 1.424559e-01 2.032099e-01
[7] 1.063995e-01 8.832360e-02
$value
[1] 53030467124
$counts
function gradient
     501       NA
```

```
$convergence
[1] 1
$message
NULL
$outer.iterations
[1] 1
$barrier.value
[1] 0.1169739
```

„par“, „die beste Menge der gefundenen Parameter“ beinhaltet sehr kleine, aber positive Werte. Der errechnete Wert für den Odds Ratio liegt bei etwa 53 Milliarden, also ein sehr großer Wert. Das ist zwar auch eine zulässige Lösung, allerdings muss man sich fragen, was dies für eine Aussagekraft hat. Konvergenz wurde auch keine ermittelt. Das Ergebnis des Odds Ratios hat in dieser Berechnung ein riesiges Intervall: $\gamma_{111} = [-2.199537 \cdot 10^{46}, \ldots, 53030467124]$. Das ergibt ein nahezu uninterpretierbares Ergebnis.

Aufgrund dieser Ergebnisse ist zu vermuten, dass bei den anderen statistischen Maßzahlen und Tests ebenfalls keine sinnvollen Ergebnisse herauskommen.

Im Gespräch mit meinen Betreuern ist auch vermutet worden, dass die Ganzzahligkeit bei den hier absoluten Häufigkeiten dem Algorithmus Probleme bereitet. Mit Hilfe meiner Betreuer habe ich nach alternativen Algorithmen in R gesucht, jedoch keine besseren gefunden.

In dem mathematischen Programm MatLab hätte man den Sachverhalt ebenfalls ausprobieren können, allerdings hätte dies den Rahmen dieser Arbeit überschritten. Zum anderen gab es keine Garantie, dass MatLab 100%-ig keine Probleme bei der konvexen Optimierung mit absoluten Randhäufigkeiten gehabt hätte.

4.3 Diskussion

In der Anwendung dieses Kapitels wurde versucht, die Formel des auf Z bedingten Odds Ratio als Funktion mit konvexer Optimierung zu maximieren bzw. zu minimieren. Es sollte ein Odds-Ratio-Intervall, beispielhaft bedingt auf $Z = 1$, für die Kontingenztafel berechnet werden. Die Randverteilungen und Eigenschaften der Kontingenztafeln sind aus dem Kapitel 2 übernommen bzw. weitergeführt worden.

Die Berechnung mit absoluten Randhäufigkeiten war nicht möglich, da der Algorithmus „contrOptim“ eine Fehlermeldungen ausgab, nach denen die Anfangswerte nicht zulässig waren. Deshalb konnte der Algorithmus kein Ergebnis finden bzw. berechnen.

Danach wurde, obwohl dies nicht Teil der Arbeit war, ausprobiert, ob der Algorithmus auf relative Randhäufigkeiten anders reagiert. Dies ist tatsächlich der Fall. Es kamen Ergebnisse heraus, die zwar zulässig sind, allerdings keine sinnvolle Interpretation zulassen. Für das beispielhafte Odds-Ratio-Intervall für h_{111} kam ein riesiges Intervall von quasi $-\infty$ bis $+\infty$ heraus.

Aufgrund dieses Ergebnisses ist die Vermutung sehr stark, dass der Algorithmus „constrOptim“ scheinbar nicht mit ganzen Zahlen umgehen kann. Mit relativen Häufigkeiten, also nicht ganzen Zahlen, scheint es auch nicht zu funktionieren, aber der Algorithmus gibt wenigstens eine Lösung. Begründen kann man dies unter anderem dadurch, dass die relativen Häufigkeiten direkt von den absoluten Werten abhängen, das heißt, das Problem der Ganzzahligkeit wird nicht wirklich dadurch beseitigt.

Das wies darauf hin, dass bei den anderen Maßzahlen und Tests ebenfalls keine oder keine befriedigenden Ergebnisse ermittelt werden, denn die statistischen Maßzahlen und Tests, die in diesem Kapitel bearbeitet werden sollten, beruhen ebenfalls auf ganzen Zahlen. Demzufolge würde der Algorithmus bei diesen Berechnungen ebenfalls keine Lösungen finden.

Im Gespräch mit meinen Betreuern ist eben diese Vermutung, dass die Ganzzahligkeit bei den hier absoluten Häufigkeiten dem Algorithmus Probleme bereitet, zu Tage getreten. Mit Hilfe meiner Betreuer habe ich nach alternativen Algorithmen in R gesucht, jedoch keine besseren gefunden. In dem mathematischen Programm MatLab hätte man den Sachverhalt ebenfalls ausprobieren können, allerdings hätte dies den Rahmen dieser Arbeit überschritten. Zum anderen gab es zu diesem Zeitpunkt keine Garantie, dass MatLab nicht dieselben Probleme bei der konvexen Optimierung nur mit Randhäufigkeiten gehabt hätte.

Kapitel 5

Fazit und Ausblick

Abschließend wird in diesem Kapitel ein Abgleich der Ergebnisse dieser Arbeit und der anfänglichen Ziele vorgenommen.

5.1 Fazit

Ziel dieser Bachelor-Thesis war es, Randverteilungen hinsichtlich ihres statistischen Analysepotentials zu untersuchen. Dazu sollte die Methode der linearen Optimierung für die Berechnungen herangezogen werden. Als Grundlage für die Berechnungen standen keine Einzeldaten, sondern nur Randverteilungen zur Verfügung.

Nach einer Einleitung wurden im zweiten Kapitel die zugrunde liegende Theorie erläutert und mit der statistischen Software R an einem medizinischen Szenario angewendet. Am Beispiel einer fiktiven Studie über „Wirkstoff - Placebo"-Wirkung wurden durch lineare Optimierung aus der gegebenen Randverteilung die Einträge innerhalb der $2 \times 2 \times 2$ Kontingenztafel berechnet. Die Berechnung mit 1-dimensionalen Randhäufigkeiten wurden anschließend durch zusätzlich vorgegebene 2-dimensionale Randhäufigkeiten, angelehnt an die Methode der Fréchet Bounds, erweitert und analog durchgeführt. Die Ergebnisse bestätigten insofern die Erwartungen, dass je mehr Bedingungen angesetzt sind, desto kleiner werden die Intervalle der berechneten Einzelhäufigkeiten.

Im dritten Kapitel wurde mit denselben Methoden wie im vorhergehenden Kapitel eine weitere Anwendung durchgeführt. Das betrachtete Szenario befasste sich mit der Kostenberechnung bei Diabetes-Erkrankungen in Kombination weiterer Merkmale. Hierbei wurden die Kostenintervalle der Behandlung der Patienten mit der jeweiligen Merkmalskombination und die insgesamten Kosten derselben berechnet. Durch die Wahl der Randverteilungen trat in den Ergebnissen der Effekt auf, dass die errechneten Intervalle relativ klein wurden. Dies ist ein sinnvolles Ergebnis, allerdings können anders geartete Randverteilungen völlig andere Ergebnisse, also auch breitere Intervalle, hervorrufen.

Als abschließende Aussage dieses Kapitels kann man festhalten, dass derartige Intervallberechnungen zum Beispiel in Krankenhäusern als grobe Anhaltspunkte gewertet werden können,

falls keine Einzeldaten für die genaue Berechnung zur Verfügung stehen.

Im Kapitel 4 wurden die theoretischen Grundlagen für die Berechnung verschiedener statistischer Analysemethoden erläutert und sollten in R durch konvexe Optimierung auf das Anwendungsbeispiel des zweiten Kapitels angewendet werden. Es stellte sich heraus, dass der verwendete Lösungsalgorithmus in R mit der Ganzzahligkeit der gegebenen Randverteilung nicht umgehen kann. Von R wurden keine Ergebnisse berechnet. Nur mit relativen Häufigkeiten wurden Ergebnisse erzielt, die allerdings nicht aussagekräftig bzw. sinnvoll interpretierbar waren. Deshalb wurde die weitere statistische Auswertung eingestellt, da kein anderer Algorithmus in der kürze der Zeit gefunden bzw. programmiert werden konnte.

Insgesamt ist festzustellen, dass Berechnungen mit ganzen Zahlen, die auf linearer Optimierung basieren, sinnvolle Ergebnisse liefern können. Allerdings ist sehr wichtig zu beachten, wieviele Restriktionen einbezogen werden, wie die gegebenen Randverteilungen beschaffen sind und vor allem welche Erwartungen man die Beschaffenheit der resultierenden Intervalle stellt.

5.2 Ausblick

Der Grundgedanke war es, das statistische Analysepotential von Randverteilungen zu untersuchen. Dies geschah vor dem Hintergrund, dass viele öffentlich zugänglich gemachte Daten (z.B. vom Statistischen Bundesamt) sehr aufwendig anonymisiert werden müssen, um den Datenschutz zu gewährleisten. Wenn das Analysepotential von Randverteilungen sehr hoch wäre, könnte man in einigen Bereichen auf aufwendige Anonymisierung verzichten und ausschließlich Randverteilungen zu Analysezwecken öffentlich zugänglich machen.

Im Allgemeinen könnte man die in dieser Arbeit vorgenommenen Analysen in mehrerer Hinsicht erweitern. Zum einen könnte eine Erweiterung auf mehr als 2 Kategorien pro Merkmal darüber Aufschluss geben, wie sich die Einflüsse und Abhängigkeiten der unterschiedlichdimensionalen Randverteilungen auf Berechnungen auswirken. Auch könnte die Anzahl der beteiligten Merkmale weiter gesteigert werden.

Zum anderen wäre eine Erweiterung auf stetige bzw. metrische Merkmale ebenfalls sinnvoll. Dazu müsste vorher eine Klassenbildung durchgeführt werden. In Kapitel 3 war das quasistetiges Merkmal „Alter“ vertreten, allerdings nur in zwei Gruppen aufgeteilt. Hier könnte eine feinere Unterteilung genaueren Aufschluss über Zusammenhänge geben.

Des Weiteren könnte man sich Schätzungen von Erwartungswerten vorstellen, die auf theoretischen Wahrscheinlichkeiten basieren.

Für die Berechnung der Intervalle der Einzelhäufigkeiten wäre außer der linearen Optimierung eventuell eine weitere Vorgehensweise denkbar gewesen. Man kann die Einzelhäufigkeiten durch Subtraktion der Randhäufigkeiten ausdrücken. Nach der verbleibenden Einzelhäufigkeit müsste differenziert werden und so die Extremalpunkte wie Maximium und Minimum berechnet werden. Dieses Vorgehen müsste für jeden Zelleneintrag der Kontingenztafel wiederholt

werden.

Bezüglich des Kapitels der statistischen Auswertungen (Kap. 4) sind weitere Untersuchungen vorstellbar. Einerseits könnte der Sachverhalt aus Kapitel 4 mit der Software MatLab untersucht werden, bzw. getestet werden, ob statistische Auswertungen, wie in Kapitel 4 beschrieben, möglich sind.

Des Weiteren könnte man versuchen, einen Algorithmus in R zu programmieren, der mit Ganzzahligkeit bei konvexer Optimierung umgehen kann.

Darauf aufbauend könnten verschiedene Tests durchgeführt werden, wie in dieser Arbeit geplant war. Tests haben bei diesem Sachverhalt, dass nur Randverteilungen zur Verfügung stehen, einen höheren Interpretationsgehalt, als reine Zusammenhangsmaße, wie z.B. der χ^2-Kontingenzkoeffizient, weil bei den Tests verschiedene Signifikanzniveaus gleichzeitig (also Intervalle von Signifikanzniveaus) angesetzt werden könnten. Durch das errechnete Testgrößen-Intervall könnte man vielleicht erfahren, ab welchem Wert der Testgröße ein bestimmtes Signifikanzniveau erreicht ist. Dann wüsste man die zu diesen Testgrößen (vom Algorithmus) ermittelten Häufigkeitsverteilungen und wüsste letztendlich, welche Verteilungen in einer Kontingenztafel notwendig sind, um ein bestimmtes Signifikanzniveau der Zusammenhänge zwischen den einzelnen Merkmalen zu erhalten.

Anhang A

Theorie und Anwendung in $2 \times 2 \times 2$ Kontingenztafeln

Hier ist ein allgemeiner Hinweis bzgl. der im Anhang aufgeführten R-Codes zu geben: Da die R-Codes der einzelnen Kapitel sehr umfangreich sind, werden die für die Programmierung zwar wichtigen aber für das Verständnis eher unnötigen Passagen teilweise durch Auslassungspunkte „(...)“ ersetzt, um den Anhang nicht zu groß werden zu lassen. Die vollständigen R-Codes sind auf der beiliegenden Daten-CD abgelegt.

Um die entsprechenden Codes schnell finden zu können, ist auf der letzten Seite des Anhangs ein Inhaltsverzeichnis der Daten-CD aufgeführt, in dem die Kapitelnummer, der Pfad und eine kurze Beschreibung angegeben sind.

A.1 Theorie zur Berechnung der Einzelhäufigkeiten (Kap. 2.1)

A.1.1 Kontigenztafeln (Kap. 2.1.1)

A.1.1.1 Partialtabellen-Darstellung (Kap. 2.1.1)

Die Tabellen A.1 und A.2 sind vergleichbar in j und i aufgeteilt.

$j=1$		Merkmal Z		
		1	2	$h_{i1\cdot}$
Merkmal X	1	h_{111}	h_{112}	$h_{11\cdot}$
	2	h_{211}	h_{212}	$h_{21\cdot}$
	$h_{\cdot 1k}$	$h_{\cdot 11}$	$h_{\cdot 12}$	$h_{\cdot 1\cdot}$

$j=2$		Merkmal Z		
		1	2	$h_{i2\cdot}$
Merkmal X	1	h_{121}	h_{122}	$h_{12\cdot}$
	2	h_{221}	h_{222}	$h_{22\cdot}$
	$h_{\cdot 2k}$	$h_{\cdot 21}$	$h_{\cdot 22}$	$h_{\cdot 2\cdot}$

Tabelle A.1: *Partialtabellen*-Darstellung der dreidimensionalen $I \times K$ - Kontingenztafel, J Ebenen

$i=1$		Merkmal Z		
		1	2	$h_{1j\cdot}$
Merkmal Y	1	h_{111}	h_{112}	$h_{11\cdot}$
	2	h_{121}	h_{122}	$h_{12\cdot}$
	$h_{1\cdot k}$	$h_{1\cdot 1}$	$h_{1\cdot 2}$	$h_{1\cdot\cdot}$

$i=2$		Merkmal Z		
		1	2	$h_{2j\cdot}$
Merkmal Y	1	h_{211}	h_{212}	$h_{21\cdot}$
	2	h_{221}	h_{222}	$h_{22\cdot}$
	$h_{2\cdot k}$	$h_{2\cdot 1}$	$h_{2\cdot 2}$	$h_{2\cdot\cdot}$

Tabelle A.2: *Partialtabellen*-Darstellung der dreidimensionalen $J \times K$ - Kontingenztafel, I Ebenen

A.1.1.2 Einfache dreidimensionale Kontingenztafel (Kap. 2.1.1)

		Merkmal Z				
		1		2		
		$Y=1$	$Y=2$	$Y=1$	$Y=2$	$h_{i\cdot\cdot}$
Merkmal X	1	h_{111}	h_{121}	h_{112}	h_{122}	$h_{1\cdot\cdot}$
	2	h_{211}	h_{221}	h_{212}	h_{222}	$h_{2\cdot\cdot}$
	$h_{\cdot\cdot k}$	$h_{\cdot\cdot 1}$		$h_{\cdot\cdot 2}$		$h_{\cdots}$

Tabelle A.3: Dreidimensionale $I \times K$ - Kontingenztafeln

A.2 Berechnung der Einzelhäufigkeiten in R (Kap. 2.2)

A.2.1 Anwendung der Linearen Optimierung (Kap. 2.2.2)

A.2.1.1 Zielfunktionsvektoren

Für h_{121} :

$$\overline{h_{121}}\,c^T = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} \ ,$$

$$\underline{h_{121}}\,c^T = \begin{pmatrix} 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} \ .$$

Für h_{122} :

$$\overline{h_{122}}\,c^T = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} \ ,$$

$$\underline{h_{122}}\,c^T = \begin{pmatrix} 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} \ .$$

Für h_{211} :

$$\overline{h_{211}}\,c^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} \ ,$$

$$\underline{h_{211}}\,c^T = \begin{pmatrix} 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} \ .$$

Für h_{212} :

$$\overline{h_{212}}\,c^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} \ ,$$

$$\underline{h_{212}}\,c^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} \ .$$

Für h_{221} :

$$\overline{h_{221}}\,c^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} \ ,$$

$$\underline{h_{221}}\,c^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} \ .$$

Für h_{222} :

$$\overline{h_{222}}\,c^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \in \mathbb{N}^{1,8} \ ,$$

$$\underline{h_{222}}\,c^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{pmatrix} \in \mathbb{N}^{1,8} \ .$$

A.2.1.2 R-Code zur linearen Optimierung mit 1-dimensionalen Randhäufigkeiten

Dateiname auf Daten-CD: `Kap2_Lin_Opt_CK_20100813`,
`Kap2_Lin_Opt_CodePlusErgebnisse_CK_20100813`

```
> # Lineare Optimierung
>
> # Paket lpSolve laden
> library(lpSolve)
>
> # Gegebene Randverteilung:
```

(...)

```
> # Vorgaben für lpSolve
> f.con <- A
> f.dir <- c("<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=")
> f.rhs <- B
>
> # Max und Min von h111
> # Max von h111
> # Zielfunktionsvektor c^T
> C_h111max <- c( 1, 0, 0, 0, 0, 0, 0, 0)
> C_h111max_t <- t(C_h111max)
> # Vorgaben für lpSolve
> f.obj1 <- C_h111max_t
> # Lin. Opt.-problem
> h_111_max <- lp ("max", f.obj1, f.con, f.dir, f.rhs, int.vec=1:8)
>
> h_111_max$objval
[1] 45
> h_111_max$solution
[1] 45  5  0  0  0 19  0 31
>
> # Min von h111
> # Zielfunktionsvektor c^T
> C_h111min <- c( -1, 0, 0, 0, 0, 0, 0, 0)
> C_h111min_t <- t(C_h111min)
> # Vorgaben für lpSolve
> f.obj2 <- C_h111min_t
> # Lin. Opt.-problem
> h_111_min <- lp ("max", f.obj2, f.con, f.dir, f.rhs, int.vec=1:8)
>
> h_111_min$objval
[1] 0
> h_111_min$solution
[1]  0 37 13  0 32  0  0 18
>
> if ( h_111_min$objval < 0 ) { h_111_min_w <- h_111_min$objval*(-1)
+  } else { h_111_min_w <- h_111_min$objval }
> h_111_min_w
[1] 0
>
```

```
> h_111 <- paste("h_111 = [",h_111_min_w,", ...,", h_111_max$objval,"]")
>
> # Max und Min von h112
> # Max von h112
> # Zielfunktionsvektor c^T
> C_h112max <- c( 0, 1, 0, 0, 0, 0, 0, 0)
> C_h112max_t <- t(C_h112max)
> # Vorgaben für lpSolve
> f.obj3 <- C_h112max_t
> # Lin. Opt.-problem
> h_112_max <- lp ("max", f.obj3, f.con, f.dir, f.rhs, int.vec=1:8)
>
> h_112_max$objval
[1] 50
> h_112_max$solution
[1]  0 50  0  0 14  5 31  0
>
> # Min von h112
> # Zielfunktionsvektor c^T
> C_h112min <- c( 0, -1, 0, 0, 0, 0, 0, 0)
> C_h112min_t <- t(C_h112min)
> # Vorgaben für lpSolve
> f.obj4 <- C_h112min_t
> # Lin. Opt.-problem
> h_112_min <- lp ("max", f.obj4, f.con, f.dir, f.rhs, int.vec=1:8)
>
> h_112_min$objval
[1] 0
> h_112_min$solution
[1] 19  0  0 31 26 24  0  0
>
> if ( h_112_min$objval < 0 ) { h_112_min_w <- h_112_min$objval*(-1)
+  } else { h_112_min_w <- h_112_min$objval }
> h_112_min_w
[1] 0
>
> h_112 <- paste("h_112 = [",h_112_min_w,", ...,", h_112_max$objval,"]")
>
> # Max und Min von h121
> # Max von h121
> # Zielfunktionsvektor c^T
> C_h121max <- c( 0, 0, 1, 0, 0, 0, 0, 0)
> C_h121max_t <- t(C_h121max)
> # Vorgaben für lpSolve
> f.obj5 <- C_h121max_t
> # Lin. Opt.-problem
> h_121_max <- lp ("max", f.obj5, f.con, f.dir, f.rhs, int.vec=1:8)
>
> h_121_max$objval
[1] 31
> h_121_max$solution
[1]  0 19 31  0 14 36  0  0
>
```

```
> # Min von h121
> # Zielfunktionsvektor c^T
> C_h121min <- c( 0, 0, -1, 0, 0, 0, 0, 0)
> C_h121min_t <- t(C_h121min)
> # Vorgaben für lpSolve
> f.obj6 <- C_h121min_t
> # Lin. Opt.-problem
> h_121_min <- lp ("max", f.obj6, f.con, f.dir, f.rhs, int.vec=1:8)
>
> h_121_min$objval
[1] 0
> h_121_min$solution
[1]  0 19  0 31 45  5  0  0
>
> if ( h_121_min$objval < 0 ) { h_121_min_w <- h_121_min$objval*(-1)
+  } else { h_121_min_w <- h_121_min$objval }
> h_121_min_w
[1] 0
>
> h_121 <- paste("h_121 = [",h_121_min_w,", ...,", h_121_max$objval,"]")
```

(...)

```
> # Liste der Intervalle
> list (h_111 , h_112 , h_121 , h_122 , h_211 , h_212 , h_221 , h_222)
[[1]]
[1] "h_111 = [ 0 , ..., 45 ]"
[[2]]
[1] "h_112 = [ 0 , ..., 50 ]"
[[3]]
[1] "h_121 = [ 0 , ..., 31 ]"
[[4]]
[1] "h_122 = [ 0 , ..., 31 ]"
[[5]]
[1] "h_211 = [ 0 , ..., 45 ]"
[[6]]
[1] "h_212 = [ 0 , ..., 50 ]"
[[7]]
[1] "h_221 = [ 0 , ..., 31 ]"
[[8]]
[1] "h_222 = [ 0 , ..., 31 ]"
```

A.2.2 Anwendung der 2-dimensionalen Fréchet Bounds (Kap. 2.2.3)

A.2.2.1 Zusätzliche Rahmenbedingungen

Gesamt-Bedingungsvektor b_G:

$$b_G = \begin{pmatrix} b \\ -b \end{pmatrix} = \begin{pmatrix} 50 \\ 50 \\ 69 \\ 31 \\ 45 \\ 55 \\ 25 \\ 25 \\ 20 \\ 30 \\ 30 \\ 39 \\ 15 \\ 16 \\ 34 \\ 16 \\ 35 \\ 15 \\ -50 \\ -50 \\ -69 \\ -31 \\ -45 \\ -55 \\ -25 \\ -25 \\ -20 \\ -30 \\ -30 \\ -39 \\ -15 \\ -16 \\ -34 \\ -16 \\ -35 \\ -15 \end{pmatrix} \in \mathbb{Z}^{36,1}.$$

Gesamt-Bedingungsmatrix A_G:

$$A_G = \begin{pmatrix} A \\ -A \end{pmatrix} = \begin{pmatrix}
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
-1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 \\
-1 & -1 & 0 & 0 & -1 & -1 & 0 & 0 \\
0 & 0 & -1 & -1 & 0 & 0 & -1 & -1 \\
-1 & 0 & -1 & 0 & -1 & 0 & -1 & 0 \\
0 & -1 & 0 & -1 & 0 & -1 & 0 & -1 \\
-1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 \\
-1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 \\
-1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & -1
\end{pmatrix} \in \mathbb{N}^{12,8}.$$

A.2.2.2 R-Code zur linearen Optimierung mit 1- und 2-dimensionalen Randhäufigkeiten

Dateiname auf Daten-CD: `Kap2_Lin_Opt_mehrBedingungen_CK_20100813`, `Kap2_Lin_Opt_mehrBedingungen_CodePlusErgebnisse_CK_20100813`

```
> # Lineare Optimierung mit mehr Bedingungen: 2-dim. Randhäufigkeiten
>
> # Paket lpSolve laden
> library(lpSolve)
>
> # Gegebene Randverteilung:
```

(...)

```
> # Vorgaben für lpSolve
> f.con2 <- A2
> f.dir2 <- c( "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=",
+ "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=",
+ "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=", "<=" )
> f.rhs2 <- B2
>
> # Max und Min von h111
> # Max von h111
> # Zielfunktionsvektor c^T
> C_h111max <- c( 1, 0, 0, 0, 0, 0, 0, 0)
> C_h111max_t <- t(C_h111max)
> # Vorgaben für lpSolve
> f.obj1 <- C_h111max_t
> # Lin. Opt.-problem
> h_111_max <- lp ("max", f.obj1, f.con2, f.dir2, f.rhs2, int.vec=1:8)
>
> h_111_max$objval
[1] 25
> h_111_max$solution
[1] 25  9  0 16  5 30 15  0
>
> # Min von h111
> # Zielfunktionsvektor c^T
> C_h111min <- c( -1, 0, 0, 0, 0, 0, 0, 0)
> C_h111min_t <- t(C_h111min)
> # Vorgaben für lpSolve
> f.obj2 <- C_h111min_t
> # Lin. Opt.-problem
> h_111_min <- lp ("max", f.obj2, f.con2, f.dir2, f.rhs2, int.vec=1:8)
>
> h_111_min$objval
[1] -10
> h_111_min$solution
[1] 10 24 15  1 20 15  0 15
>
```

```
> if ( h_111_min$objval < 0 ) { h_111_min_w <- h_111_min$objval*(-1) } else {
+ h_111_min_w <- h_111_min$objval }
> h_111_min_w
[1] 10
>
> h_111 <- paste("h_111 = [",h_111_min_w,", ...,", h_111_max$objval,"]")
>
> # Max und Min in einem
> # Max von h112
> # Zielfunktionsvektor c^T
> C_h112max <- c( 0, 1, 0, 0, 0, 0, 0, 0)
> C_h112max_t <- t(C_h112max)
> # Vorgaben für lpSolve
> f.obj3 <- C_h112max_t
> # Lin. Opt.-problem
> h_112_max <- lp ("max", f.obj3, f.con2, f.dir2, f.rhs2, int.vec=1:8)
>
> h_112_max$objval
[1] 24
> h_112_max$solution
[1] 10 24 15  1 20 15  0 15
>
> # Min von h112
> # Zielfunktionsvektor c^T
> C_h112min <- c( 0, -1, 0, 0, 0, 0, 0, 0)
> C_h112min_t <- t(C_h112min)
> # Vorgaben für lpSolve
> f.obj4 <- C_h112min_t
> # Lin. Opt.-problem
> h_112_min <- lp ("max", f.obj4, f.con2, f.dir2, f.rhs2, int.vec=1:8)
>
> h_112_min$objval
[1] -9
> h_112_min$solution
[1] 25  9  0 16  5 30 15  0
>
> if ( h_112_min$objval < 0 ) { h_112_min_w <- h_112_min$objval*(-1) } else {
+ h_112_min_w <- h_112_min$objval }
> h_112_min_w
[1] 9
>
> h_112 <- paste("h_112 = [",h_112_min_w,", ...,", h_112_max$objval,"]")
>
> # Max und Min in einem
> # Max von h121
> # Zielfunktionsvektor c^T
> C_h121max <- c( 0, 0, 1, 0, 0, 0, 0, 0)
> C_h121max_t <- t(C_h121max)
> # Vorgaben für lpSolve
> f.obj5 <- C_h121max_t
> # Lin. Opt.-problem
> h_121_max <- lp ("max", f.obj5, f.con2, f.dir2, f.rhs2, int.vec=1:8)
>
```

```
> h_121_max$objval
[1] 15
> h_121_max$solution
[1] 10 24 15  1 20 15  0 15
>
> # Min von h121
> # Zielfunktionsvektor c^T
> C_h121min <- c( 0, 0, -1, 0, 0, 0, 0, 0)
> C_h121min_t <- t(C_h121min)
> # Vorgaben für lpSolve
> f.obj6 <- C_h121min_t
> # Lin. Opt.-problem
> h_121_min <- lp ("max", f.obj6, f.con2, f.dir2, f.rhs2, int.vec=1:8)
>
> h_121_min$objval
[1] 0
> h_121_min$solution
[1] 25  9  0 16  5 30 15  0
>
> if ( h_121_min$objval < 0 ) { h_121_min_w <- h_121_min$objval*(-1) } else {
+ h_121_min_w <- h_121_min$objval }
> h_121_min_w
[1] 0
>
> h_121 <- paste("h_121 = [",h_121_min_w,", ...,", h_121_max$objval,"]")
```

(...)

```
> list (h_111 , h_112 , h_121 , h_122 , h_211 , h_212 , h_221 , h_222)
[[1]]
[1] "h_111 = [ 10 , ..., 25 ]"
[[2]]
[1] "h_112 = [ 9 , ..., 24 ]"
[[3]]
[1] "h_121 = [ 0 , ..., 15 ]"
[[4]]
[1] "h_122 = [ 1 , ..., 16 ]"
[[5]]
[1] "h_211 = [ 5 , ..., 20 ]"
[[6]]
[1] "h_212 = [ 15 , ..., 30 ]"
[[7]]
[1] "h_221 = [ 0 , ..., 15 ]"
[[8]]
[1] "h_222 = [ 0 , ..., 15 ]"
```

Anhang B

Kostenanwendung in der Medizin anhand einer $2 \times 2 \times 2$ Kontingenztafel

Hier ist ein allgemeiner Hinweis bzgl. der im Anhang aufgeführten R-Codes zu geben: Da die R-Codes der einzelnen Kapitel sehr umfangreich sind, werden die für die Programmierung zwar wichtigen aber für das Verständnis eher unnötigen Passagen teilweise durch Auslassungspunkte „(...)“ ersetzt, um den Anhang nicht zu groß werden zu lassen. Die vollständigen R-Codes sind auf der beiliegenden Daten-CD abgelegt.

Um die entsprechenden Codes schnell finden zu können, ist auf der letzten Seite des Anhangs ein Inhaltsverzeichnis der Daten-CD aufgeführt, in dem die Kapitelnummer, der Pfad und eine kurze Beschreibung angegeben sind.

B.1 Kosten bei Diabetes-Erkrankungen (Kap. 3.1)

B.1.1 Rahmenbedingungen für Kostenanwendung (Kap. 3.1.1)

B.1.1.1 Zielfunktionsvektoren für Kostenanwendung

Für h_{111} :
$$\underline{h_{111}}\, c^T = \begin{pmatrix} -1200 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} \ .$$

Für h_{112} :
$$\overline{h_{112}}\, c^T = \begin{pmatrix} 0 & 3100 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} \ ,$$
$$\underline{h_{112}}\, c^T = \begin{pmatrix} 0 & -3100 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} \ .$$

Für h_{121} :
$$\overline{h_{121}}\, c^T = \begin{pmatrix} 0 & 0 & 1550 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} \ ,$$
$$\underline{h_{121}}\, c^T = \begin{pmatrix} 0 & 0 & -1550 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} \ .$$

Für h_{122} :
$$\overline{h_{122}}\, c^T = \begin{pmatrix} 0 & 0 & 0 & 3340 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} \ ,$$

$$\underline{h_{122}}\, c^T = \begin{pmatrix} 0 & 0 & 0 & -3340 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} .$$

Für h_{211} :

$$\overline{h_{211}}\, c^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 1500 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} ,$$

$$\underline{h_{211}}\, c^T = \begin{pmatrix} 0 & 0 & 0 & 0 & -1500 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} .$$

Für h_{212} :

$$\overline{h_{212}}\, c^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 3460 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} ,$$

$$\underline{h_{212}}\, c^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & -3460 & 0 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} .$$

Für h_{221} :

$$\overline{h_{221}}\, c^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1870 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} ,$$

$$\underline{h_{221}}\, c^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & -1870 & 0 \end{pmatrix} \in \mathbb{N}^{1,8} .$$

Für h_{222} :

$$\overline{h_{222}}\, c^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3645 \end{pmatrix} \in \mathbb{N}^{1,8} ,$$

$$\underline{h_{222}}\, c^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & -3645 \end{pmatrix} \in \mathbb{N}^{1,8} .$$

B.1.2 Kostenberechnung mit 1-dimensionalen Randhäufigkeiten (Kap. 3.1.2)

B.1.2.1 R-Code zur linearen Optimierung mit 1-dimensionalen Randhäufigkeiten

Dateiname auf Daten-CD: `Kap3_Kosten_Lin_Opt_CK_20100813`,
`Kap3_Kosten_Lin_Opt_CodePlusErgebnisse_CK_20100813`

```
> # Kostenanwendung durch Lineare Optimierung
>
> # Paket lpSolve laden
> library(lpSolve)
>
> # Gegebene Randverteilung: 1-dim. Randhäufigkeiten
```

(...)

```
> # Max und Min cost_h111
> # Max von cost_h111
> C_cost_h111max <- c( 1200, 0, 0, 0, 0, 0, 0, 0)
> C_cost_h111max_t <- t(C_cost_h111max)       # Zielfunktionsvektor c^T
> f.obj1 <- C_cost_h111max_t
> # Lin. Opt.-problem
> cost_h_111_max <- lp ("max", f.obj1, f.con, f.dir, f.rhs, int.vec=1:8)
>
> cost_h_111_max$objval
[1] 30000
> cost_h_111_max$solution
[1] 25  0 15  8  0  0  0 52
>
> # Min von cost_h111
> C_cost_h111min <- c( -1200, 0, 0, 0, 0, 0, 0, 0)
> C_cost_h111min_t <- t(C_cost_h111min)
> f.obj2 <- C_cost_h111min_t
> cost_h_111_min <- lp ("max", f.obj2, f.con, f.dir, f.rhs, int.vec=1:8)
>
> cost_h_111_min$objval
[1] 0
> cost_h_111_min$solution
[1]  0 17 31  0  8  0  1 43
>
> if ( cost_h_111_min$objval < 0 ) { cost_h_111_min_w <- cost_h_111_min$objval*(-1)
+  } else { cost_h_111_min_w <- cost_h_111_min$objval }
> cost_h_111_min_w
[1] 0
>
> cost_h_111 <- paste("Kosten_h_111 = [",cost_h_111_min_w,", ...,",
+ cost_h_111_max$objval,"]")
```

(...)

```
> # Max und Min cost_h222
```

```
> # Max von cost_h222
> C_cost_h222max <- c( 0, 0, 0, 0, 0, 0, 0, 3645)
> C_cost_h222max_t <- t(C_cost_h222max)
> f.obj15 <- C_cost_h222max_t
> cost_h_222_max <- lp ("max", f.obj15, f.con, f.dir, f.rhs, int.vec=1:8)
>
> cost_h_222_max$objval
[1] 189540
> cost_h_222_max$solution
[1] 25  0 15  8  0  0  0 52
>
> # Min von cost_h222
> C_cost_h222min <- c( 0, 0, 0, 0, 0, 0, 0, -3645)
> C_cost_h222min_t <- t(C_cost_h222min)
> f.obj16 <- C_cost_h222min_t
>
> cost_h_222_min <- lp ("max", f.obj16, f.con, f.dir, f.rhs, int.vec=1:8)
>
> cost_h_222_min$objval
[1] 0
> cost_h_222_min$solution
[1]  0  0  0 48 13 12 27  0
>
> if ( cost_h_222_min$objval < 0 ) { cost_h_222_min_w <- cost_h_222_min$objval*(-1)
+  } else { cost_h_222_min_w <- cost_h_222_min$objval }
> cost_h_222_min_w
[1] 0
>
> cost_h_222 <- paste("Kosten_h_222 = [",cost_h_222_min_w,", ...,",
+ cost_h_222_max$objval,"]")
>
> # Liste der Intervalle
> list (cost_h_111 , cost_h_112 , cost_h_121 , cost_h_122 , cost_h_211 ,
+ cost_h_212 , cost_h_221 , cost_h_222)
[[1]]
[1] "Kosten_h_111 = [ 0 , ..., 30000 ]"
[[2]]
[1] "Kosten_h_112 = [ 0 , ..., 77500 ]"
[[3]]
[1] "Kosten_h_121 = [ 0 , ..., 62000 ]"
[[4]]
[1] "Kosten_h_122 = [ 0 , ..., 160320 ]"
[[5]]
[1] "Kosten_h_211 = [ 0 , ..., 37500 ]"
[[6]]
[1] "Kosten_h_212 = [ 0 , ..., 86500 ]"
[[7]]
[1] "Kosten_h_221 = [ 0 , ..., 74800 ]"
[[8]]
[1] "Kosten_h_222 = [ 0 , ..., 189540 ]"
```

B.1.3 Kostenberechnung mit 1- und 2-dimensionalen Randhäufigkeiten (Kap. 3.1.3)

B.1.3.1 Zusätzliche Rahmenbedingungen

Gesamt-Bedingungsvektor b_G:

$$b_G = \begin{pmatrix} b \\ -b \end{pmatrix} = \begin{pmatrix} 48 \\ 52 \\ 25 \\ 75 \\ 40 \\ 60 \\ 19 \\ 29 \\ 21 \\ 31 \\ 15 \\ 10 \\ 25 \\ 50 \\ 14 \\ 34 \\ 11 \\ 41 \\ -48 \\ -52 \\ -25 \\ -75 \\ -40 \\ -60 \\ -19 \\ -29 \\ -21 \\ -31 \\ -15 \\ -10 \\ -25 \\ -50 \\ -14 \\ -34 \\ -11 \\ -41 \end{pmatrix} \in \mathbb{Z}^{36,1}.$$

B.1.3.2 R-Code zur linearen Optimierung mit 1- und 2-dimensionalen Randhäufigkeiten

Dateiname auf Daten-CD: Kap3_Kosten_Lin_Opt_mehrBedingungen_CK_20100813, Kap3_Kosten_Lin_Opt_mehrBedingungen_CodePlusErgebnisse_CK_20100813

```
> # Kostenanwendung durch Lineare Optimierung
>
> # Paket lpSolve laden
> library(lpSolve)
>
> # Gegebene Randverteilung: 1-dim. Randhäufigkeiten, 2-dim. Randhäufigkeiten
```

(...)

```
> # Max und Min cost_h221
> # Max von cost_h221
> C_cost_h221max <- c( 0, 0, 0, 0, 0, 0, 1870, 0)
> C_cost_h221max_t <- t(C_cost_h221max)
> f.obj13 <- C_cost_h221max_t
>
> cost_h_221_max <- lp ("max", f.obj13, f.con2, f.dir2, f.rhs2, int.vec=1:8)
>
> cost_h_221_max$objval
[1] 37400
> cost_h_221_max$solution
[1] 14  0  5 29  1 10 20 21
>
> # Min von cost_h221
> C_cost_h221min <- c( 0, 0, 0, 0, 0, 0, -1870, 0)
> C_cost_h221min_t <- t(C_cost_h221min)
> f.obj14 <- C_cost_h221min_t
>
> cost_h_221_min <- lp ("max", f.obj14, f.con2, f.dir2, f.rhs2, int.vec=1:8)
>
> cost_h_221_min$objval
[1] -18700
> cost_h_221_min$solution
[1]  4 10 15 19 11  0 10 31
>
> if ( cost_h_221_min$objval < 0 ) { cost_h_221_min_w <- cost_h_221_min$objval*(-1)
+ } else { cost_h_221_min_w <- cost_h_221_min$objval }
> cost_h_221_min_w
[1] 18700
>
> cost_h_221 <- paste("Kosten_h_221 = [",cost_h_221_min_w,", ...,",
+ cost_h_221_max$objval,"]")
```

(...)

```
> # Liste der Intervalle
```

```
> list (cost_h_111 , cost_h_112 , cost_h_121 , cost_h_122 , cost_h_211 ,
+ cost_h_212 , cost_h_221 , cost_h_222)
[[1]]
[1] "Kosten_h_111 = [ 4800 , ..., 16800 ]"
[[2]]
[1] "Kosten_h_112 = [ 0 , ..., 31000 ]"
[[3]]
[1] "Kosten_h_121 = [ 7750 , ..., 23250 ]"
[[4]]
[1] "Kosten_h_122 = [ 63460 , ..., 96860 ]"
[[5]]
[1] "Kosten_h_211 = [ 1500 , ..., 16500 ]"
[[6]]
[1] "Kosten_h_212 = [ 0 , ..., 86500 ]"
[[7]]
[1] "Kosten_h_221 = [ 18700 , ..., 37400 ]"
[[8]]
[1] "Kosten_h_222 = [ 76545 , ..., 112995 ]"
```

B.2 Gesamtkosten bei Diabetes-Erkrankungen (Kap. 3.2)

B.2.1 Gesamtkostenberechnung mit 1-dimensionalen Randhäufigkeiten (Kap. 3.2.1)

B.2.1.1 R-Code zur linearen Optimierung mit 1-dimensionalen Randhäufigkeiten

Dateiname auf Daten-CD: Kap3_GesamtKosten_Lin_Opt_CK_20100813, Kap3_GesamtKosten_Lin_Opt_CodePlusErgebnisse_CK_20100813

```
> # Kostenanwendung durch Lineare Optimierung
>
> # Paket lpSolve laden
> library(lpSolve)
> # Gegebene Randverteilung: 1-dim. Randhäufigkeiten
```

(...)

```
> # Max und Min cost
> # Max cost
> C_cost_max <- c( 1200, 3100, 1550, 3340, 1500, 3460, 1870, 3645)
> C_cost_max_t <- t(C_cost_max)       # Zielfunktionsvektor c^T
> f.objG1 <- C_cost_max_t
> # Lin. Opt.-problem
> cost_max <- lp ("max", f.objG1, f.con, f.dir, f.rhs, int.vec=1:8)
> cost_max$objval
[1] 274040
> cost_max$solution
[1]  0  0 13 35  0 25 27  0
>
> # Min cost
> C_cost_min <- c( -1200, -3100, -1550, -3340, -1500, -3460, -1870, -3645)
> C_cost_min_t <- t(C_cost_min)
> f.objG2 <- C_cost_min_t
>
> cost_min <- lp ("max", f.objG2, f.con, f.dir, f.rhs, int.vec=1:8)
> cost_min$objval
[1] -269385
> cost_min$solution
[1]  0  0 15 33 25  0  0 27
>
> if ( cost_min$objval < 0 ) {cost_min_w <- cost_min$objval*(-1)
+  } else { cost_min_w <- cost_min$objval }
> cost_min_w
[1] 269385
>
> cost <- paste("Kosten = [",cost_min_w,", ...,", cost_max$objval,"]")
> print ( cost )
[1] "Kosten = [ 269385 , ..., 274040 ]"
```

B.2.2 Gesamtkostenberechnung mit 1- und 2-dimensionalen Randhäufigkeiten (Kap. 3.2.2)

B.2.2.1 R-Code zur linearen Optimierung mit 1- und 2-dimensionalen Randhäufigkeiten

Dateiname auf Daten-CD: Kap3_GesamtKosten_Lin_Opt_mehrBedingungen_CK_20100813, Kap3_GesamtKosten_Lin_Opt_mehrBedingungen_CodePlusErgebnisse_CK_20100813

```
> # Kostenanwendung durch Lineare Optimierung
>
> # Paket lpSolve laden
> library(lpSolve)
> # Gegebene Randverteilung: 1-dim. Randhäufigkeiten, 2-dim. Randhäufigkeiten
```

(...)

```
> # Max und Min cost
> # Max cost
> C_cost_max <- c( 1200, 3100, 1550, 3340, 1500, 3460, 1870, 3645)
> C_cost_max_t <- t(C_cost_max)       # Zielfunktionsvektor c^T
> f.objG1 <- C_cost_max_t
> # Lin. Opt.-problem
> cost_max <- lp ("max", f.objG1, f.con2, f.dir2, f.rhs2, int.vec=1:8)
> cost_max$objval
[1] 271455
> cost_max$solution
[1] 14  0  5 29  1 10 20 21
>
> # Min cost
> C_cost_min <- c( -1200, -3100, -1550, -3340, -1500, -3460, -1870, -3645)
> C_cost_min_t <- t(C_cost_min)
> f.objG2 <- C_cost_min_t
>
> cost_min <- lp ("max", f.objG2, f.con2, f.dir2, f.rhs2, int.vec=1:8)
> cost_min$objval
[1] -270705
> cost_min$solution
[1]  4 10 15 19 11  0 10 31
>
> if ( cost_min$objval < 0 ) {cost_min_w <- cost_min$objval*(-1)
+  } else { cost_min_w <- cost_min$objval }
> cost_min_w
[1] 270705
>
> cost <- paste("Kosten = [",cost_min_w,", ...,", cost_max$objval,"]")
> print ( cost )
[1] "Kosten = [ 270705 , ..., 271455 ]"
```

Anhang C

Theorie und Anwendung der statistischen Auswertung in $2 \times 2 \times 2$ Kontingenztafeln

C.1 Durchführung der statistischen Auswertung in R (Kap. 4.2)

C.1.1 Odds Ratio (Kap. 4.2.1)

C.1.1.1 R-Code zur konvexen Optimierung mit absoluten Häufigkeiten

Dateiname auf Daten-CD: `Kap4_OddsRatio_Kon_Opt_CK_201008004`,
`Kap4_OddsRatio_Kon_Opt_CodePlusErgebnisse_CK_201008004`

```
> # Statistische Auswertung durch konvexe Optimierung
> # Odds Ratio
>
> # Gegebene Randverteilung: 1-dim. Randhäufigkeiten
```

(...)

```
> # Bedingungsvektor b, Bedingungsmatrix A
```

(...)

```
> # zu erwartende absolute Häufigkeiten
> h_111 <- (h1../h...)*(h.1./h...)*(h..1/h...)*h...
> h_112 <- (h1../h...)*(h.1./h...)*(h..2/h...)*h...
> h_121 <- (h1../h...)*(h.2./h...)*(h..1/h...)*h...
> h_122 <- (h1../h...)*(h.2./h...)*(h..2/h...)*h...
> h_211 <- (h2../h...)*(h.1./h...)*(h..1/h...)*h...
> h_212 <- (h2../h...)*(h.1./h...)*(h..2/h...)*h...
> h_221 <- (h2../h...)*(h.2./h...)*(h..1/h...)*h...
> h_222 <- (h2../h...)*(h.2./h...)*(h..2/h...)*h...
```

(...)

```
> # Objekt
> theta1 <- c(h_111, h_112, h_121, h_122, h_211, h_212, h_221, h_222)
> #zu minimierende Funktion
> OR_ij1 <- function(x) {
+         (x[1]/x[5])/(x[3]/x[7])
+ }
```

(...)

```
> # Lösungsalgorithmus für Minimum
> OR_h_111_min <- constrOptim(theta1, OR_ij1, NULL, ui = A, ci = b)
Fehler in constrOptim(theta1, OR_ij1, NULL, ui = A, ci = b) :
        initial value not feasible
> OR_h_111_min
Fehler: objekt "OR_h_111_min" nicht gefunden
> # Lösungsalgorthmus für Maximum
> OR_h_111_max <- constrOptim(theta1, OR_ij1, NULL, ui = A, ci = b,
+ control=list(fnscale=-1))
Fehler in constrOptim(theta1, OR_ij1, NULL, ui = A, ci = b,
+ control = list(fnscale = -1)) :
        initial value not feasible
> OR_h_111_max
Fehler: objekt "OR_h_111_max" nicht gefunden
> # funktioniert nicht
```

C.1.1.2 R-Code zur konvexen Optimierung mit relativen Häufigkeiten

Dateiname auf Daten-CD: Kap4_OddsRatio_Kon_Opt_mitRelHäuf_CK_20100804, Kap4_OddsRatio_Kon_Opt_mitRelHäuf_CodePlusErgebnisse_CK_20100804

```
> # Statistische Auswertung durch konvexe Optimierung
> # Odds Ratio
```

(...)

```
> # Relative Randhäufigkeiten ...
> # Bedingungsvektor b, Bedingungsmatrix A
```

(...)

```
> # zu erwartende relative Häufigkeiten
> f_111 <- f1..*f.1.*f..1
> f_112 <- f1..*f.1.*f..2
> f_121 <- f1..*f.2.*f..1
> f_122 <- f1..*f.2.*f..2
> f_211 <- f2..*f.1.*f..1
> f_212 <- f2..*f.1.*f..2
> f_221 <- f2..*f.2.*f..1
> f_222 <- f2..*f.2.*f..2
```

(...)

```
> # Objekt
> theta1 <- c(f_111, f_112, f_121, f_122, f_211, f_212, f_221, f_222)
> #zu minimierende Funktion
> OR_ij1 <- function(x) {
+     (x[1]/x[5])/(x[3]/x[7])
+ }
```

(...)

```
> # Lösungsalgorithmus für Minimum
> OR_h_111_min <- constrOptim(theta1, OR_ij1, NULL, ui = A, ci = b)
> OR_h_111_min
$par
[1]  4.323484e+01  4.454786e-01  1.251559e-31  1.664349e-01  1.090387e-14
[6]  2.976455e-01 -6.942715e-01  1.708770e-01
$value
[1] -2.199537e+46
$counts
function gradient
     503       NA
$convergence
[1] 1
$message
NULL
$outer.iterations
[1] 51
$barrier.value
[1] 0
> # Lösungsalgorithmus für Maximum
> OR_h_111_max <- constrOptim(theta1, OR_ij1, NULL, ui = A, ci = b,
+ control=list(fnscale=-1))
> OR_h_111_max
$par
[1] 1.940832e-01 1.974475e-01 2.733515e-12 9.042863e-02 1.424559e-01 2.032099e-01
[7] 1.063995e-01 8.832360e-02
$value
[1] 53030467124
$counts
function gradient
     501       NA
$convergence
[1] 1
$message
NULL
$outer.iterations
[1] 1
$barrier.value
[1] 0.1169739
> # mit relativen Häufigkeiten gibt er Ergebnisse aus, aber die sind für
> # Odds Ratios gegen "-infinity"
> # unbrauchbare Ergebnisse, rel.H. sind sehr klein ...
```

Anhang D

Inhaltsverzeichnis der beiliegenden Daten-CD

Kapitel-nummer	**Pfad**	**Beschreibung**
R-Codes		
2.2.2	<CD/DVD-Laufwerk>\R-Codes\Kap2_Lin_Opt_CK_20100813	Code der Berechnung der Einzelhäufigkeiten mit 1-dim. Randverteilung
	<CD/DVD-Laufwerk>\R-Codes\Kap2_Lin_Opt_CodePlusErgebnisse_CK_20100813	Code der Berechnung der Einzelhäufigkeiten mit 1-dim. Randverteilung, **mit Ergebnissen**
2.2.3	<CD/DVD-Laufwerk>\R-Codes\Kap2_Lin_Opt_mehrBedingungen_CK_20100813	Code der Berechnung der Einzelhäufigkeiten mit 1- und 2-dim. Randverteilung
	<CD/DVD-Laufwerk>\R-Codes\Kap2_Lin_Opt_mehrBedingungen_CodePlusErgebnisse_CK_20100813	Code der Berechnung der Einzelhäufigkeiten mit 1- und 2-dim. Randverteilung, **mit Ergebnissen**
3.1.2	<CD/DVD-Laufwerk>\R-Codes\Kap3_Kosten_Lin_Opt_CK_20100813	Code der Kostenberechnung für die Einzelhäufigkeiten mit 1-dim. Randverteilung
	<CD/DVD-Laufwerk>\R-Codes\Kap3_Kosten_Lin_Opt_CodePlusErgebnisse_CK_20100813	Code der Kostenberechnung für die Einzelhäufigkeiten mit 1-dim. Randverteilung, **mit Ergebnissen**

Tabelle D.1: Inhaltsverzeichnis der Daten-CD

Kapitel-nummer	Pfad	Beschreibung
3.1.3	<CD/DVD-Laufwerk>\R-Codes\Kap3_Kosten_Lin_Opt_mehrBedingungen_CK_20100813	Code der Kostenberechnung für die Einzelhäufigkeiten mit 1- und 2-dim. Randverteilung
	<CD/DVD-Laufwerk>\R-Codes\Kap3_Kosten_Lin_Opt_mehrBedingungen_CodePlusErgebnisse_CK_20100813	Code der Kostenberechnung für die Einzelhäufigkeiten mit 1- und 2-dim. Randverteilung, **mit Ergebnissen**
3.2.1	<CD/DVD-Laufwerk>\R-Codes\Kap3_GesamtKosten_Lin_Opt_CK_20100813	Code der Gesamtkostenberechnung mit 1-dim. Randverteilung
	<CD/DVD-Laufwerk>\R-Codes\Kap3_GesamtKosten_Lin_Opt_CodePlusErgebnisse_CK_20100813	Code der Gesamtkostenberechnung mit 1-dim. Randverteilung, **mit Ergebnissen**
3.2.2	<CD/DVD-Laufwerk>\R-Codes\Kap3_GesamtKosten_Lin_Opt_mehrBedingungen_CK_20100813	Code der Gesamtkostenberechnung mit 1- und 2-dim. Randverteilung
	<CD/DVD-Laufwerk>\R-Codes\Kap3_GesamtKosten_Lin_Opt_mehrBedingungen_CodePlusErgebnisse_CK_20100813	Code der Gesamtkostenberechnung mit 1- und 2-dim. Randverteilung, **mit Ergebnissen**
4.2.1	<CD/DVD-Laufwerk>\R-Codes\Kap4_OddsRatio_Kon_Opt_CK_201008004	Code der Berechnung des Odds Ratios mit absoluten Häufigkeiten
	<CD/DVD-Laufwerk>\R-Codes\Kap4_OddsRatio_Kon_Opt_CodePlusErgebnisse_CK_201008004	Code der Berechnung des Odds Ratios mit absoluten Häufigkeiten, **mit Ergebnissen**
	<CD/DVD-Laufwerk>\R-Codes\Kap4_OddsRatio_Kon_Opt_mitRelHäuf_CK_20100804	Code der Berechnung des Odds Ratios mit relativen Häufigkeiten
	<CD/DVD-Laufwerk>\R-Codes\Kap4_OddsRatio_Kon_Opt_mitRelHäuf_CodePlusErgebnisse_CK_20100804	Code der Berechnung des Odds Ratios mit relativen Häufigkeiten, **mit Ergebnissen**

Tabelle D.2: Fortsetzung Inhaltsverzeichnis der Daten-CD

Glossar

1-dimensionale Randverteilung	Randhäufigkeiten, bei denen eine Dimension bestimmt ist und über die anderen Dimensionen summiert ist; zum Beispiel: $h_{i..}, i = 1, \ldots, I$.	10, 11, 17, 19–21, 26, 29, 33, 34, 46, 49, 62
2-dimensionale Randverteilung	Randhäufigkeiten, bei denen zwei Dimensionen bestimmt sind und über die andere Dimension summiert ist; zum Beispiel: $h_{i \cdot k}, i = 1, \ldots, I,\ k = 1, \ldots, K$.	10, 11, 17, 19, 26, 28, 29, 37, 39, 41, 42, 46, 62
Aggregation	Zusammenfassung von Daten. (von lateinisch: *aggregatio* = 'Anhäufung', 'Vereinigung')	1, 2, 17
Anonymisierung	Anonymisierung ist die Veränderung von personenbezogenen oder unternehmensbezogenen Einzeldaten.	1–4, 17, 63
Assoziationsmaß	Ein Assoziationsmaß ist ein Maß, das den Zusammenhang zwischen zwei oder mehreren Variablen/Merkmalen ausdrückt, wird auch Kontingenzkoeffizient genannt.	48, 49
Deanonymisierung	Makrodaten wieder in Mikrodaten umwandeln.	2, 4
Dichotomie	Aufteilung in zwei Mengen, die disjunkt (d.h. nicht miteinander vereinbar bzw. einander genau entgegengesetzt) sind.	51, 52, 54
Fréchet Bounds	Fréchet Bounds sind ein mathematisches Werkzeug, um Intervallgrenzen zu berechnen.	6, 17, 19, 25, 26, 29, 62
Grundgesamtheit	Die Grundgesamtheit, auch *Population*, bezeichnet die Menge aller statistischen Einheiten (auch Merkmalsträger, Erhebungseinheit) mit übereinstimmenden Identifikationskriterien (sachlich, zeitlich und örtlich).	4, 52

Identifikationsmerkmal	Identifikationsmerkmale sind Merkmale,die eine zweifelsfreie Identifizierung von Personen, Unternehmen oder Gruppen möglich machen.	1
Konsistenz(-bedingungen)	Konsistenz (lateinisch: *con* = 'zusammen' u. *sistere* = 'halten') bedeutet Bestand, Zusammenhalt, Geschlossenheit und In-sich-Ruhen, Gegenbegriff ist *Inkonsistenz*. In der Logik bedeutet Konsistenz die Widerspruchsfreiheit eines axiomatischen Systems.	18, 26, 29
Mikrodaten	Synonym für Einzeldaten; Sie enthalten für einen einzelnen Merkmalsträger die an ihm beobachteten oder gemessenen Merkmalsausprägungen der untersuchten Merkmale.	1–5
Pseudonymisierung	Pseudonymisierung ist die Veränderung nur der direkten Identifikationsmerkmale von personenbezogenen oder unternehmensbezogenen Einzeldaten.	1
Reidentifizierung	Vorgang, der zum eindeutigen (Wieder-)Erkennen einer Person oder eines Objektes dient; Identifizierung.	1, 2, 17

Literaturverzeichnis

Augustin, T. (2009, Sommersemster). Entscheidungstheorie. Vorlesung.

Beck-Bornholdt, H.-P. (2005). *Mit an Wahrscheinlichkeit grenzender Sicherheit: logisches Denken und Zufall*. Reinbek bei Hamburg: Rowohlt-Taschenbuch-Verlag.

Boyd, S. (2004). *Convex Optimization*. New York: Cambridge University Press. http://www.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf, 19.08.2010.

Bundesministerium der Justiz (2009). Bundesdatenschutzgesetz (BDSG). http://www.gesetze-im-internet.de/bundesrecht/bdsg_1990/gesamt.pdf, 25.07.2010.

Cox, L. H. (2002). *Bounds on Entries in 3-Dimensional Contingency Tables Subject to Given Marginal Totals*, Volume 2316/2002, pp. 21–33. Berlin/Heidelberg: Springer-Verlag. http://www.springerlink.com/content/4dnwh47fc1ln5va8/, 25.07.2010.

Fahrmeir, L., R. Künstler, I. Pigeot, and G. Tutz (2004). *Der Weg zur Datenanalyse* (5 ed.). Berlin/Heidelberg: Springer-Verlag.

FDZ (2007). Anonymität von Mikrodaten. http://www.forschungsdatenzentrum.de/anonymisierung.asp, 25.07.2010.

FDZ (2009). *Kapitel III. Anonymisierung von Mikrodaten*. Statistisches Bundesamt. http://www.empiwifo.uni-freiburg.de/lehre-teaching-1/winter-term-08-09/wirtschaftsstatistik, 25.07.2010.

Fienberg, S. E. (N/A). Fréchet and Bonferroni Bounds for Multi-way Tables of Counts With Applications to Disclosure Limitation. Technical report, Department of Statistics, Carnegie Mellon University Pittsburgh. USA. http://www.stat.cmu.edu/tr/tr691/tr691.html, 05.08.2010.

Gentleman, R. and R. Ihaka (N/A). R documentation. http://www.iiap.res.in/astrostat/School07/R/html/stats/html/constrOptim.html, 16.08.2010.

Grötschel, M. (2003/2004, Wintersemster). Lineare Optimierung (Algorithmische Diskrete Mathematik II), Vorlesung. http://www.zib.de/groetschel/teaching/skriptADMII.pdf, 05.08.2010.

Höhne, J. (2003). *Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten*, Volume 42, pp. 69–94. Wiesbaden: Statistisches Bundesamt.

Janzing, D. (2003/2004, Wintersemester). Algorithmentechnik. http://iaks-www.ira.uka.de/home/janzing/AlgotechSkriptDominik.pdf, 19.08.2010.

Kühne, U. (2009). Von Zahlen geblendet - Simpson-Paradox: Scheinbar klare Verhältnisse werden in ihr Gegenteil verkehrt. Eine Warnung vor dem naiven Vertrauen in Statistiken. *Der Freitag N/A*(42), 18. http://www.freitag.de/wissen/0942-wissen-simpson-paradox-philosophie-mathematik, 12.08.2010.

Köster, I., L. von Ferber, and H. Hauner (2005). Die Kosten des Diabetes mellitus - Ergebnisse der KoDiM-Studie. Technical report, PMV forschungsgruppe, Köln. http://www.pmvforschungsgruppe.de/pdf/02_forschung/c_ergebnis_kodim.pdf, 09.08.2010.

Lübbert, D. (1999). Testverfahren aus Statistik A, B und C, (Skript). http://www.luebbert.net//download/statb.pdf, 07.08.2010.

Lilly Deutschland GmbH (2009, November). Diabetes-Arten. http://www.lilly-diabetes.de/patienten/diabetes-verstehen/diabetes-arten.html, 09.08.2010.

Mortensen, U. (2010). Log-lineare Modelle. Technical report, Institut für Psychologie der Johannes Gutenberg-Universität Mainz. http://www.uwe-mortensen.de/LoglineareModelle.pdf, 27.07.2010.

Ronning, G., R. Sturm, J. Höhne, R. Lenz, M. Rosemann, M. Scheffler, and D. Vorgrimler (2005). *Statistik und Wissenschaft: Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten*, Volume 4, pp. 51–101. Wiesbaden: Statistisches Bundesamt.

Rosemann, M. (2006). *Auswirkungen datenverändernder Anonymisierungsverfahren auf die Analyse von Mikrodaten*. Tübingen: Institut für Angewandte Wirtschaftsforschung e.V.

Schoffer, O. (2008). *SAS im Forschungsdatenzentrum der Statistischen Landesämter*, pp. 123–126. Kamenz/Dresden: Statistisches Landesamt des Freistaates Sachsen. http://www.statistik.sachsen.de/22/2_08schoffer.pdf, 25.07.2010.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society Series B*(2), 238–241.

Statistisches Bundesamt (2008). Bundesstatistikgesetz (BStatG). http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/SharedContent/Oeffentlich/AbtA/A2/Rechtsgrundlagen/Statistikbereiche/AllgemeineBestimmungen/010_BStatG,property=file.pdf, 25.07.2010.

Toutenburg, H. and C. Heumann (2008). *Prüfen statistischer Hypothesen* (4 ed.)., pp. 127–164. Berlin/Heidelberg: Springer-Verlag. http://www.springerlink.com/content/r7231163t4332362/fulltext.pdf, 07.08.2010.

Tutz, G. (2009). The Analysis of Contingency Tables: Loglinear and Graphical Models.

Wagner, C. H. (1982). Simpson's paradox in real life. *The American Statistician 36*(1), 46–48.

Wikipedia. Die freie Online-Enzyklopädie (2010, Juli). Anonymisierung und Pseudonymisierung. http://de.wikipedia.org/wiki/Anonymisierung_und_Pseudonymisierung, 01.08.2010.

Wikipedia Die freie Online-Enzyklopädie (2010, April). Konvexe Optimierung. http://de.wikipedia.org/wiki/Konvexe_Optimierung, 19.08.2010.

Erklärung zur Urheberschaft

Hiermit erkläre ich, die vorliegende Bachelor-Thesis selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt zu haben. Die aus fremden Quellen direkt oder indirekt übernommen Gedanken wurden als solche kenntlich gemacht und mit allen notwendigen bibliographischen Angaben ins Literaturverzeichnis aufgenommen.

Die Arbeit wurde weder in dieser noch in ähnlicher Form als Prüfungsleistung für eine andere Prüfung eingereicht.

München, den 23.08.2010

Unterschrift
(Christian Kluge)