

Abschlussarbeiten bei der Arbeitsgruppe Metho(dolo)gische Grundlagen der Statistik und ihre Anwendungen

Prof. Dr. Thomas Augustin, Dr. Marco Cattaneo,
Julia Kopf, Gero Walter, Andrea Wiencierz

7. Februar 2011

1 Vorbemerkungen

1.1 Denkbare idealtypische Schwerpunkte von Abschlussarbeiten

- Verallgemeinerung eines klassischen Verfahrens durch Adaption von verallgemeinerten Methoden (Intervallwahrscheinlichkeit, Messfehler)
- Implementierung und Untersuchung einer Methode aus der Literatur oder Vergleich zweier Methoden
- konzeptuelle oder explorative Weiterentwicklung von bisher nicht-etablierten Ansätzen
- Analyse eines komplexen Datenbeispiels mithilfe einer geeigneten Kombination etablierter Methoden *und* kritischer Diskussion der auftretenden methodischen Probleme
- ausführlicher Literaturüberblick über ein Thema mit strukturierter Aufbereitung unter methodischen Gesichtspunkten

1.2 Bachelor's Thesis, Master's Thesis bzw. Diplomarbeit

Aus den meisten Themenbereichen lassen sich sowohl Themen für Bachelorarbeiten als auch für Master- und Diplomarbeiten gewinnen. Selbstverständlich wird bei der genauen Themenstellung auf die unterschiedlichen Bearbeitungszeiten und Vorkenntnisse Rücksicht genommen.

1.3 Kolloquium

Es wird ein regelmäßiges Kolloquium angeboten, um den Austausch zwischen den einzelnen Studierenden, die eine Abschlussarbeit schreiben, zu fördern und allgemeine Aspekte gemeinsam diskutieren zu können. Jede(r) Studierende soll dabei die Möglichkeit erhalten, die Arbeit etwa zum Zeitpunkt der Anmeldung und nach etwa zwei Drittel der Bearbeitungszeit kurz vorzustellen, um breites Feedback zu den Plänen bzw. Ergebnissen zu erhalten. In das Kolloquium

können bei Bedarf auch allgemeine Tutorien (z.B. zur Literaturrecherche mit Datenbanken oder zum Aufbau von Simulationsstudien) integriert werden.

2 Themenbereiche

2.1 Fehler-in-den-Variablen und fehlende Daten

Fehler-in-den-Variablen-Modelle nehmen die Tatsache ernst, dass in den unterschiedlichsten Anwendungsgebieten häufig die eigentlich interessierenden Variablen nicht direkt und exakt beobachtbar sind; stattdessen stehen nur stark fehlerbehaftete Größen (Proxyvariablen) zur Verfügung. Typische Beispiele sind

- fehlerbehaftete technische Messungen i.e.S.,
- stark gerundete Daten durch eingeschränkte Messpräzision,
- durch Erinnerungseffekte verzerrte Daten,
- zur Anonymisierung bewusst kontaminierte Daten,
- Skalenindizes und andere Operationalisierungen komplexer Konstrukte (in den Wirtschafts- und Sozialwissenschaften, aber zum Beispiel auch in der Arbeitsepidemiologie).

Die meisten statistischen Verfahren setzen allerdings eigentlich vollständige und fehlerfrei beobachtete Daten voraus, so dass bei direkter, naiver Verwendung der Daten deutliche Verzerrungen auftreten, die zu gravierenden Fehlschlüssen bei der inhaltlichen Interpretation führen können. Daher versucht man, den Fehlermechanismus selbst zu modellieren und damit geeignet in die statistische Analyse einzubinden bzw. den Einfluss der Fehler auf das Ergebnis der Analyse zu korrigieren.

Korrekturverfahren für Berkson-Fehler (z.B. im Cox-Modell)

Fehler-in-den-Variablen-Modelle und Intervalldaten

Anwendung bzw. Erweiterung von Fehler-in-den-Variablen-Verfahren auf intervallwertige Daten.

Inverse probability weighted correction (IPWC)

Diese Methoden korrigieren Schätzungen aus fehlerbehafteten Datensätzen mithilfe einer Umgewichtung von Daten (ähnliche Idee wie Horvitz-Thompson-Schätzer in der Stichprobentheorie).

- Literaturüberblick über verschiedene IPWC-Ansätze für zensierte Daten
- Anwendung der sog. Methode der doppelt-robusten Korrektur
- IPWC anwenden auf *unabhängige* Variable in Regression
- Nichtparametrische Accelerated Failure Time Modelle unter Messfehler

Variablenpräselektion und Messfehler

Inwieweit produziert man eigentlich eine Messfehlersituation mit den üblichen Verzerrungen, wenn man zunächst ein Modell auswählt und Ergebnisse dieses Modells in ein anderes Modell einsetzt?

Dichteschätzung unter Messfehlern durch Deconvolution

Lockern der restriktiven Annahmen der üblichen Messfehlermodelle

- Simulationsstudien zur Auswirkung der Annahmen (fehlerbehaftete Daten simulieren, Messfehlerkorrektur unter Annahmen, die nicht dem Fehlerprozess entsprechen)
- Korrekturverfahren für abhängige Messfehler
- Welche Möglichkeiten gibt es, Korrekturverfahren auf nichtnormale Fehlerverteilungen auszuweiten?
Eine Arbeit könnte sich etwa mit Robustheitsüberlegungen bei „nicht ganz“ normalverteilten Fehlern auseinandersetzen, eine andere mit diskreten Fehlerverteilungen wie sie etwa in ökonomischen Experimenten auftreten?
- Systematische Sensitivitätsanalyse von Fehlerkorrekturmodellen

Verbessern der Kleinstichprobeneigenschaften von Fehlerkorrekturverfahren

Die meisten Korrekturverfahren werden „nur“ asymptotisch gerechtfertigt, d.h. sie liefern konsistente Schätzer. Es gibt verschiedene Ansätze durch Entwicklungen höherer Ordnungen die Eigenschaften bei mittleren und kleinen Stichproben zu verbessern.

Vergleich robuste parametrische Verfahren, Messfehlerkorrekturverfahren und nicht parametrische Verfahren

Multiple Imputation und fehlerbehaftete Daten

Inwieweit kann man Techniken der multiplen Imputation, wie sie zur Behandlung fehlender Daten angewandt wird, auf Messfehlermodelle übertragen?

Messfehlerkorrektur durch numerische Integration/ Diskretisierung/ Approximation

In einer Reihe von Situationen lassen sich die messfehlerkorrigierten Schätzgleichungen nur mehr numerisch/approximativ berechnen. Hier sollen verschiedenen Vorgehensweisen miteinander verglichen werden.

Anonymisierungsverfahren

Immer häufiger stellen amtliche Einrichtungen (wie das Statistische Bundesamt oder die Bundesanstalt für Arbeit) Daten zur wissenschaftlichen Analyse zur Verfügung, wobei zur Anonymisierung die Daten absichtlich mit einem bekannten Fehlerprozess kontaminiert werden.

Wichtig ist es hier, zu erkennen, wie empfindlich die Ergebnisse der gängigen statistischen Verfahren gegenüber den gängigen Anonymisierungsverfahren sind, und dann auch gegebenenfalls geeignete Korrekturverfahren zu entwickeln.

Modellierung nichtzufällig fehlender Daten

- Empfindlichkeit von MAR-Modellen bei nicht zufälligem Fehlen
- Modellierung von Dropout-Prozessen bei Paneldaten
- Unterschiedliche Arten eines Selektionsbias (Autogenerierte Daten, quasi-experimententelle Daten,...)

2.2 Intervallwahrscheinlichkeit und verwandte Ansätze

Viele Anwender bezweifeln die Relevanz statistischer Aussagen mit dem Argument, dass die entsprechenden Analysen „überpräzise“ und damit der Komplexität des Gegenstands nicht angemessen sind. Dies gilt insbesondere, wenn eine große Zahl von letztendlich nur mathematisch begründeten Zusatzannahmen zu treffen sind, um mit einem komplizierten statistischen Verfahren bei komplexen Daten überhaupt zu einem eindeutigen Ergebnis zu kommen.

Seit kurzem gewinnt eine prinzipiell andere Vorgehensweise mehr und mehr Anhänger: Man löst sich von dem idealisierten Präzisionsanspruch klassischer statistischer Verfahren und versucht, dafür glaubwürdigere und zuverlässigere Aussagen zu gewinnen, indem man die Information durch eine (nicht notwendig einelementige) *Menge* von passenden Modellen beschreibt. Dabei reflektiert die „Größe“ der Menge die Ungenauigkeit der Information. Entsprechende Mengen von Wahrscheinlichkeiten (oft als *Credal Sets* bezeichnet) führen auf sog. *Intervallwahrscheinlichkeiten* (engl. imprecise probabilities, mit der sog. Dempster-Shafer-Theorie als Spezialfall); bei Mengen von Parameterschätzungen spricht man von *partiell identifizierten* Schätzungen. Besonders populär sind die entsprechenden Methoden als formaler Überbau über verschiedene Bereiche der robusten Statistik und in der Entscheidungstheorie, wo sie helfen, bekannte Paradoxien bei der Modellierung ökonomischer Entscheidungen und unsicheren Expertenwissens zu überwinden.

Dempster-Shafer-Theorie bei gerundeten Daten / Heaping

Dempster-Shafer-Theorie bei Intervalldaten

Intervallwahrscheinlichkeit und Intervalldaten

Modellierung von Intervallbeobachtungen auf Basis verallgemeinerter Wahrscheinlichkeitsmodelle aus dem Bereich der Intervallwahrscheinlichkeit.

Ein Imprecise-Probability-Ansatz für den Berkson-Fehler

Partielle Identifikation

- Literaturüberblick über verschiedene Anwendungsbereiche
- Vergleich von Manskis Modell der Partiellen Identifikation (Ökonometrie) mit dem verwandten Ansatz der systematischen Sensitivitätsanalyse (Mollenberghs u.a.) (Biostatistik)
- Partielle Identifikation bei Fehlklassifikation
- Partielle Identifikation bei Messfehlern
- Partielle Identifikation in Schätzgleichungen, insbesondere bei Rundung / Heaping

Implementierung von Algorithmen für die Entscheidungstheorie mit Intervallwahrscheinlichkeit

Die üblichen Entscheidungskriterien in der Intervallwahrscheinlichkeit sind mit linearer Optimierung darstellbar; eine Implementierung in R und ein umfassender Vergleich der Methoden steht noch aus.

Dynamische (In)kohärenz bei sequentiellen Entscheidungen

Arbeitet man mit einem allgemeineren Wahrscheinlichkeitsbegriff in Entscheidungsproblemen, so gelten unter Umständen einige „Selbstverständlichkeiten“ der sequentiellen Informationsverarbeitung (Hauptsatz der Bayes-Entscheidungstheorie, Bellman-Prinzip der dynamischen Optimierung) nicht mehr, wodurch man sich in einem gewissen Sinn dynamisch inkohärent (widersprüchlich) verhält.

Ein Thema hierzu könnte darin bestehen, die diesbezügliche Diskussion in der Ökonomie und im Operations Research systematisch zusammenzufassen. Eher theoretische Arbeiten sollten diese Problematik für verschiedene verallgemeinerte Entscheidungskriterien untersuchen und die Konsequenzen für die statistische Datenanalyse, die ja auch als (sequentielles) Entscheidungsproblem formalisiert werden kann, untersuchen.

Hables Minimum-Distanz-Schätzer

Hable hat in seiner Dissertation (2009) an unserem Institut ein Minimum-Distanz-Verfahren entwickelt, mit dem parametrisierte Intervallwahrscheinlichkeitsmodelle geschätzt werden können. Die Verfahren sollen weiter untersucht werden und dann auf weitere Modelle ausgedehnt werden.

Ansätze zur Regression mit Intervallwahrscheinlichkeit

Es gibt verschiedene Vorgehensweisen, Regressionsmodelle auf Intervallwahrscheinlichkeiten zu verallgemeinern. Diese sollen implementiert, angewendet, verglichen und weiterentwickelt werden.

Credal Maximum Likelihood

Dieser Ansatz verspricht eine zuverlässigere Modellierung unbeobachteter Heterogenität. Die entsprechenden Methoden sollen implementiert, evaluiert und weiterentwickelt werden, um sie dann auch mit klassischen gemischten Modellen vergleichen zu können.

Untere / obere Schätzgleichungen

Die Theorie der unverzerrten Schätzgleichungen (Scorefunktionen) stellt einen formalen Rahmen zur Gewinnung konsistenter Schätzer bereit. Es soll eine konsequente Verallgemeinerung auf Intervallwahrscheinlichkeit untersucht und weiterentwickelt werden.

Intervallwahrscheinlichkeitsmodelle für kategoriale Daten

Das wohl bekannteste (und einfachste) Intervallwahrscheinlichkeitsmodell ist das sog. Imprecise-Dirichlet-Modell (IDM, Walley, 1996, JRSSB). Interessant wäre es, im Rahmen einer Literaturliste die mittlerweile sehr zahlreichen Anwendungen in verschiedensten Bereichen zu sichten und zu systematisieren.

Als weitere Themen bieten sich Vergleiche des IDMs mit einem alternativen Ansatz (Coolen & Augustin, 2009, IJAR) in verschiedenen Situationen an.

Klassische und verallgemeinerte Markov-Ketten

Markov-Ketten bilden das einfachste dynamische Modell. Es sollen verschiedene Anwendungen, etwa im Marketing, der Soziologie oder den Ingenieurwissenschaften, gesichtet werden und untersucht werden, inwiefern sich die Analyse durch verallgemeinerte Modelle verbessern lassen.

Lineare partielle Information

Isoliert von den Hauptsträngen der Theorie der Intervallwahrscheinlichkeit hat sich in der ökonomischen Entscheidungstheorie der Ansatz der linearen partiellen Information entwickelt. Eine Arbeit soll diese Entwicklung systematisch aufbereiten und die zugehörigen Anwendungen sichten.

Robuste Bayes-Verfahren

Das Generalized iLUCK-model (Walter & Augustin, 2009, JSTP) ist ein Modell für die bayesianische Datenanalyse, bei der anstatt mit einer einzelnen Prioriverteilung mit Mengen von Prioriverteilungen gearbeitet wird. Es zielt auf ein verbessertes Modellverhalten in dem Fall ab, dass das in der Priori ausgedrückte Vorwissen mit den Daten nicht übereinstimmt („prior-data conflict“). Während Posterioriverteilungen aus konjugierten Modellen diesen Konflikt nur ungenügend widerspiegeln, wird prior-data conflict bei Generalized iLUCK-models in der Menge der Posterioriverteilungen sehr deutlich wiedergegeben.

- einfache Datenbeispiele mit dem Generalized iLUCK-model, evtl. Programmieren (Erweiterung des luck-Package) **geeignet für BA**

- Andere Ansätze zur Diagnose von prior-data conflict und Vergleich mit dem Ansatz im Generalized iLUCK-model
- Vergleich des Generalized iLUCK-model mit Ansätzen von Whitcomb, Pericchi & Walley, insbesondere bezüglich der Reaktion auf prior-data conflict
- Behandlung von mehrdimensionalen Parametern im Generalized iLUCK-model, insbesondere die Rolle von prior-data conflict im Mehrdimensionalen
- Das Generalized iLUCK-model kann auch für die Schätzung der Parameter einer linearen Regression verwendet werden. Themen in diesem Zusammenhang sind:
 - Datenbeispiele **geeignet für BA**
 - Verallgemeinerung auf nicht-normalverteilte Zielgrößen über latente-Hilfsgröße-Methode, z.B. auf binäre Zielvariablen
- MCMC für robuste Bayes-Verfahren

Robuste frequentistische Verfahren als Spezialfall von Intervallwahrscheinlichkeit

- Huber-Strassen-Theorie und robuste Test
- Robuste Sequentialanalyse

Rekursive Partitionierungsverfahren und Intervallwahrscheinlichkeit

Vereinfacht gesprochen versuchen rekursive Partitionierungsverfahren durch sukzessives Aufspalten der Daten homogene Subgruppen zu identifizieren. Es gibt verschiedene Ideen, Klassifikations- und Regressionsbäume und darauf aufbauende Methoden durch Betrachtung von Intervallwahrscheinlichkeitsmodellen zu stabilisieren und zu robustifizieren. Diese sollen implementiert, evaluiert, verglichen und weiterentwickelt werden.

Adaptive Verfahren als Alternative zu Intervallwahrscheinlichkeit?

Adaptive Verfahren gehen auch zunächst von einer Menge von unterschiedlichen Modellklassen aus, versuchen aber datenbasiert eine Modellklasse auszuwählen, die dann der weiteren statistischen Analyse zugrunde liegen.

2.3 Likelihood-Methoden

Viele der bekanntesten statistischen Verfahren basieren auf der Likelihood-Funktion (z.B. Maximum-Likelihood-Schätzer oder Likelihood-Ratio-Test) und die Bedeutung von Likelihood-Methoden in den Anwendungen nimmt ständig zu.

Graphische Modelle

Abhängigkeiten zwischen Zufallsvariablen können graphisch repräsentiert werden. Bei der Verwendung von Likelihood-Methoden für graphische Modelle kann deren besondere Struktur ausgenutzt werden.

- **Lernen von graphischen Modellen**
Vergleich von verschiedenen Ansätzen
- **Algorithmen für graphische Modelle**
Entwicklung von Algorithmen für Likelihood-basierte Inferenz
- **Likelihood-basierte Klassifikation durch graphische Modelle**
Praktischer Vergleich von verschiedenen Ansätzen

Robuste Likelihood-Methoden

Die Likelihood-Methoden sind i.A. nicht robust gegenüber kleinen Änderungen in den Verteilungsannahmen. Abhilfe können allgemeinere Modelle oder modifizierte Likelihood-Funktionen schaffen.

- **Robustifizierung der Likelihood-Funktion**
Vergleich von verschiedenen Ansätzen
- **Lokationsschätzung mit Penalisierung von Ausreißern**
Vergleich von verschiedenen robusten Schätzern
- **Regression mit Penalisierung von Ausreißern**
Entwicklung von Approximationsalgorithmen für ML-Regression

Fuzzy Daten und Intervalldaten

Unschärfe Daten können mithilfe von Likelihood-Funktionen repräsentiert werden. Somit können Likelihood-Methoden für deren Analyse verwendet werden.

- **Verwandtschaft zu Fehler-in-den-Variablen-Methoden**
Variable selbst ist präzise und nur unscharf beobachtet
- **Likelihood-Methoden mit Fuzzy Daten**
Likelihood-Funktion und ML-Schätzer mit Fuzzy Daten
- **Nichtparametrische Likelihood-Methoden mit Intervalldaten**
Nichtparametrische Likelihood-Funktion und ML-Schätzer mit Intervalldaten
- **Regression mit Intervalldaten**
Schätzung eines Regressionsmodells mit intervallwertigen Daten bzw. Fuzzy Daten
- **Fuzzifizierung als Robustifizierung**
Lohnt es sich manchmal Daten zu „verunschärfen“ ?
- **Fuzzy-Modellierung intervallwertiger Beobachtungen**
Definition und Analysemöglichkeiten von Random Fuzzy Sets

- **Vergleich von Fuzzy Mengen und Wahrscheinlichkeitsverteilungen**
Vergleich der zwei Unsicherheitsbeschreibungen

Likelihood-basierte Entscheidungstheorie

Entscheidungen können direkt auf Basis der Likelihood-Funktion getroffen werden. Obwohl die Likelihood-basierte Entscheidungstheorie bei den speziellen Entscheidungsproblemen der Statistik (Schätzungen und Tests) sehr erfolgreich ist, wurde die allgemeine Likelihood-basierte Entscheidungstheorie bisher nicht viel erforscht.

- **Entscheidung unter komplexer Unsicherheit**
Vergleich von verschiedenen Ansätzen
- **Diskrete Entscheidungsprobleme**
Entwicklung von Algorithmen für Likelihood-basierte Entscheidungen
- **Stetige Schätzprobleme**
Entwicklung von Algorithmen für verallgemeinerte ML-Schätzer

2.4 Wirtschafts- und Sozialstatistik

Konzentrations- und Armutsmessung

- Überblick über die Anwendung von Konzentrationsmaßen in bestimmten Bereichen (z.B. Beispiel internationale Vergleiche zur Land- oder Einkommensverteilung)
- Anwendung von verschiedenen Konzentrationsmaßen auf diverse Datensätze (z.B. anonymisierte Steuerdaten, beispielsweise zur Untersuchung der Verteilung von Einkommen, Vermögen, ...)
- Armutsmessung (in) für Entwicklungsländer(n), Systematisierung und Analyse von Daten internationaler Organisationen
- Vergleich der Fraktionalisierung von Parteiensystemen in verschiedenen Ländern

Zensus

- Internationale Unterschiede in der Methodik von Volkszählungen
- Simulationsuntersuchungen zum Stichprobenschema des Zensus 2001

Regressionsmodelle in der Konzentrationsmessung

Multivariate Konzentrationsmessung

Es sollen explorativ Ideen gesammelt und evaluiert werden, wie man simultan mehrere Größen in der Konzentrationsmessung betrachten kann.

Praktische Analyse von komplexen Datensätzen

Es stehen eine Reihe von sehr komplexen Datensätzen zur Verfügung, z.B.

- in Kooperation mit der Arbeitsgruppe von Prof. Norman Braun, Institut für Soziologie, stark defizitäre Daten über legalen und illegalen Drogenkonsum.
- In Kooperation mit PD Dr. Ulrich Pötter verschiedene komplexe Studien des Deutschen Jugendinstituts
- Denkbar ist auch eine Reanalyse diverser Arbeiten basierend auf dem Sozioökonomischen Panel (SOEP) unter Verwendung weiterführender Methoden.

Statistische Ergebnisse zu Response-Effekten bei Befragungen

Assoziations-, Korrelations- und Regressionsanalysen bei Randomized Response

Randomized-Response-Techniken (Verfahren der zufallsverschlüsselten Antworten) werden immer häufiger eingesetzt, um auch bei sehr sensiblen Fragen ehrliche Antworten in persönlichen Interviews zu erhalten. Auch wenn eine individuelle Zuordnung der Antworten dabei nicht mehr möglich ist, so kann doch die Häufigkeit der interessierenden Eigenschaft immer noch unverzerrt geschätzt werden. Es sollen verschiedene Vorschläge aus der Literatur diskutiert, evaluiert und weiterentwickelt werden, wie man in dieser Situation auch Zusammenhangsanalysen durchführen kann.

Statistik im Sport

Analyse von Sportdaten mit fortgeschrittenen Methoden, z.B. Zeitreihenanalyse, dynamische Rankings mit BTL-Modellen, generalisierte Regressionsmodelle

Sozialindikatoren

Es gibt eine Vielzahl verschiedenster Sozialindikatoren; diese sollten aus statistischer Sicht aufbereitet, systematisiert und kritisch untersucht werden.

Bibliometrie

Die Bibliometrie versucht, den „wissenschaftlichen Output“ von Forschern zu messen, insbesondere durch Analyse von Daten aus Publikations- und Zitationsdatenbanken (z.B. ISI Web of Knowledge).

- Systematische Reanalyse bibliometrischer Daten, z.B. nach Geschlecht, Wanderungsbewegungen, Netzwerken (Zitierkartelle)
- kritische Evaluation der Maßzahlen zur Messung des wissenschaftlichen Outputs (Output-Indikatoren, Impact-Maße für wissenschaftliche Zeitschriften, ...)

2.5 Dogmengeschichte der Statistik

Es soll über eine rein äußerliche, historische Faktenbeschreibung hinaus fundiertes statistisches Wissen eingesetzt werden, um die Entwicklung methodologischer Positionen und Paradigmen über die Zeit nachzeichnen zu können.

Die nachhaltige Wirkung einer ausgewählten Persönlichkeiten auf die Entwicklung der Statistik

Wie haben jeweils z.B. R.A. Fisher, Bruno de Finetti, John F. Tukey, Abraham Wald die Statistik geprägt?

Die computationale Wende in der Statistik

Die Entwicklung der robusten Statistik

Das Ellsberg-Paradox und seine Wirkung auf die Entscheidungstheorie

Die Geschichte des Instituts für Statistik in München