
Methodische Statistik und Amtliche Statistik – Zwei unterschiedliche Sichten auf Daten?

Wilfried Grossmann
Institute for Scientific Computing,
University Vienna

Inhalt

- Einleitung
- Statistischer Zugang zu Daten
- Statistische Modellierung
- Integration der beiden Ansätze

Einleitung – Definition Statistik

- Definition Statistics

 - Wikipedia (English Version)

Statistics is a mathematical science pertaining to the collection, analysis, interpretation, and presentation of data. It is applicable to a wide variety of academic disciplines,.....; it is also used for making informed decisions in all areas of business and government.

 - *Official Statistics nicht als Teildisziplin aufgezählt*

Einleitung – Definition Statistik

- Definition Statistik

 - Wikipedia (Deutsche Version)

 - Statistik ist eine Zusammenfassung von bestimmten Methoden um empirische Daten zu analysieren.

Einleitung – Definition Statistik

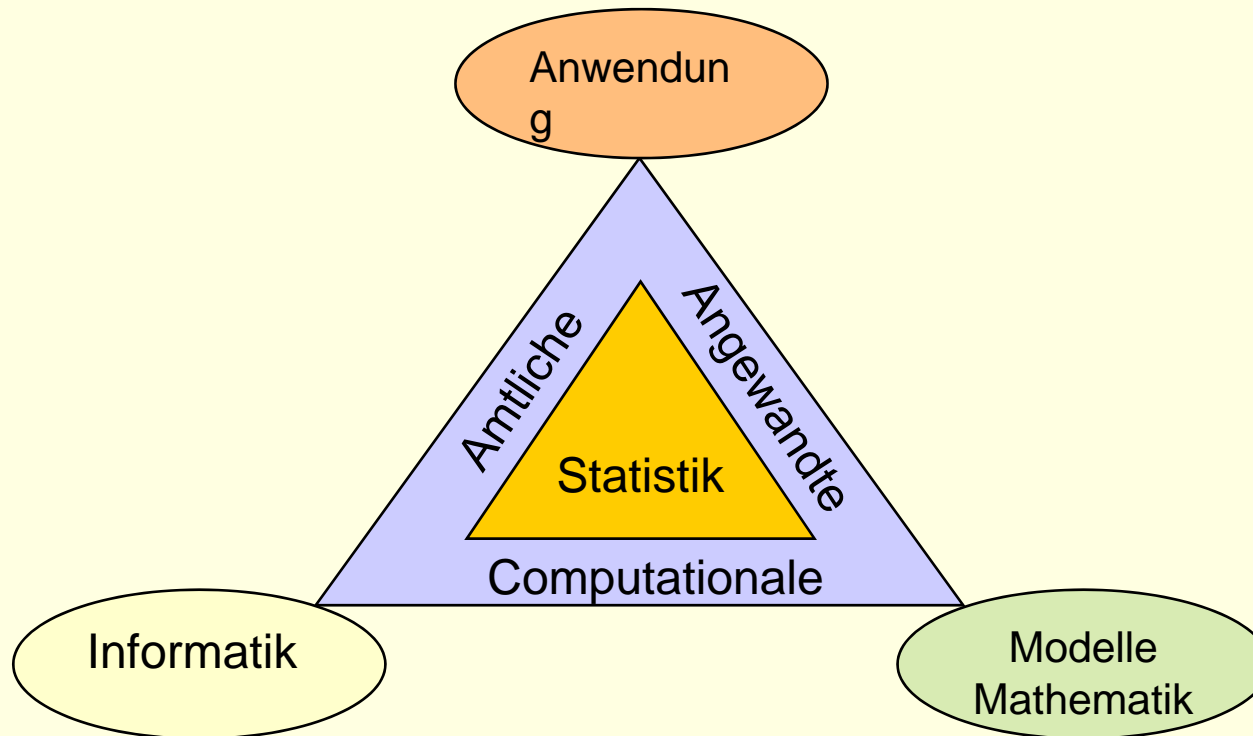
- Beide Definitionen beziehen sich auf die zwei historischen Wurzeln der Statistik
 - Modellierung von zufälligen Mustern in Daten
 - Daten zur Erfassung von ökonomischen, demographischen und sozialen Phänomenen
- Schwerpunkt etwas unterschiedlich
 - English: ***Distinct mathematical Science***, breite Anwendung, eher technisch-mathematisch orientiert
 - Deutsch: ***Empirie***, mehr auf allgemeine Eigenschaften von Statistiken orientiert: „objektiv, reliabel, valide, signifikant, bedeutend“

Einleitung – Statistik und andere Disziplinen

- Andere Disziplinen, die sich ebenfalls mit der Modellierung und Analyse von Daten beschäftigen:
 - Informatik
 - Datenbankenmodelle zur Verwaltung von Daten
 - Algorithmen zur Datenanalyse (Data Mining)
 - Mathematik
 - Numerische Mathematik (Algorithmen)
 - Modelle
 - Buchhaltung, Administration

Einleitung – Statistik und andere Disziplinen

- Statistik im Spannungsfeld Anwendung – Mathematische Modelle – Informatik



Einleitung – Eine Statistik?

- Ziel der meisten Statistische Gesellschaften ist es eine Plattform für alle Richtungen zu sein (**A**kademische Statistik, **A**ngewandte Statistik, **A**mtliche Statistik)
- Integration der verschiedenen Richtungen der Statistik erscheint oft schwierig

Einleitung – Eine Statistik?

- EUROSTAT begann in den 1980er Jahren mit einem Forschungsprogramm für Statistik
 - Brücke zwischen den Zugängen
 - Ressourcen für Statistical Computing in allen Bereichen
 - Verbesserung des Europäischen Statistischen Systems

Einleitung – Eine Statistik?

- NTTS-Konferenzen seit 1992 geben ein Bild der interessierenden Forschungsfragen in diesem Programm
 - Tendenziell ist ein Wechsel der Betrachtungsweise zu erkennen
 - Amtliche Statistik mit informatischer Orientierung in den ersten Jahren
 - Stärkere Orientierung auf Methodische Statistik in der letzten Konferenz

Einleitung – Ziel des Vortrages

- Was ist das Spezifische der Statistik?
 - Unterschiede Statistik – andere Zugängen zur Analyse von empirischen Daten?
- Welche Gemeinsamkeiten bestehen zwischen den beiden Zugängen?
 - Gemeinsamkeiten in der Vergangenheit
 - Neuere Beiträge der beiden Bereiche, die eine stärkere Gemeinsamkeit fördern könnten
- Betrachtung hat einen Akademischen Bias

Statistischer Zugang zu Daten

- Statistik hat einen speziellen Zugang zu Daten, der in beiden Fällen durch einige zentrale Begriffe charakterisiert werden kann
 - Population, Einheiten
 - Variable
 - Messung
 - Sampling, Stichprobe
- Unterschied liegt in der Betrachtung der Begriffe

Statistischer Zugang zu Daten – Population, Einheiten

- Methodische Statistik
 - In vielen Fällen nur ein mentales Konstrukt, potentiell unendlich, Ereignisstruktur (Ω, A)
- Amtliche Statistik:
 - Eine wohldefinierte endliche Grundgesamtheit, in Raum und Zeit genau spezifiziert
- Realisierung der Population (Register) ist oft nicht vollständig möglich und wird mehr in der Informatik betrachtet

Statistischer Zugang zu Daten – Population, Einheiten

- Generische Zugang der Methodischen Statistik ist offen für praktisch alle Anwendungen
 - Fragen der konkreten Realisierung werden im Zusammenhang mit den verschiedenen Anwendungsdisziplinen behandelt
 - Medizin und Biologie
 - Psychologie
 - Technik und Naturwissenschaften
 - Markt- und Meinungsforschung

Statistischer Zugang zu Daten – Population, Einheiten

- Amtliche Statistik ist traditionell auf Wirtschaft und Sozialwissenschaften sowie Verwaltung konzentriert
- Heute besteht Bedarf für viele andere Anwendungen im Rahmen der Amtlichen Statistik:
 - Psychologie (Bildung), Ökologie (Umweltstatistik), Technik (Verkehrsstatistik)
- Amtliche Statistik sollte als **eine umfassende Anwendungsdisziplin** verstanden werden

Statistischer Zugang zu Daten – Variable

- Methodische Statistik:
 - Generische Beschreibung einer Variablen X , die erst in der Anwendung spezifiziert wird
 - Typisierung nach Verwendung in Modellen, z.B. Faktor, abhängige Variable, Confounder, Proxy-Variable
 - Statistik ist nur in beschränktem Maße an der genauen Bestimmung der Definitionen beteiligt

Statistischer Zugang zu Daten – Variable

- Amtliche Statistik:
 - Eigenschaften von wohlunterscheidbaren Einheiten
 - Typisierung nach Bedeutung zur Charakterisierung der Einheiten
 - Oft komplexe Definitionen, z.B. Beschäftigung
 - Oft ein Merkmal, das einfacher bestimmbar
 - Amtliche Statistik ist substantiell an der Definition der Variablen beteiligt

Statistischer Zugang zu Daten – Messung

- Methodische Statistik:
 - Eher generische Einheiten und Labels
 - Festlegung eines Skalenniveaus, oft nicht normativ
 - Messtheorie wird von anderen Disziplinen entwickelt, z.B. Entwicklung von Items in der Psychologie

Statistischer Zugang zu Daten – Messung

- Amtliche Statistik:
 - Komplexe Terminologie,
 - Hierarchische Systeme (Klassifikationen), die Aggregation erlauben, z.B. NACE
 - Eigenständige Entwicklung der Mess-Systeme

Statistischer Zugang zu Daten – Messung

- Informatik leistet einen wertvollen Beitrag in der Strukturierung der Wertebereiche
 - Terminologieserver
 - Klassifikationsdatenbanken
- Derartiges Wissen wird von der Amtlichen Statistik effizient genutzt
- Methodische Statistik sollte sich hier verstärkt an den Erfahrungen der Amtlichen Statistik orientieren

Statistischer Zugang zu Daten – Sampling, Stichprobe

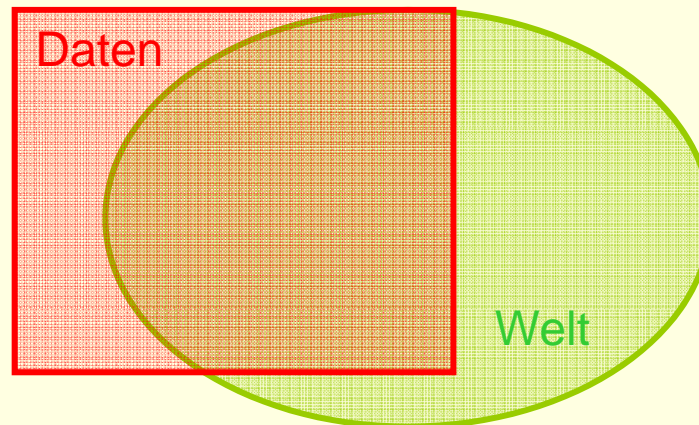
- Statistische Modellierung:
 - Generische Schemata (i.i.d., Zeitreihe, Markovprozess,...) steht im Vordergrund
 - Versuchsplanung, Stichprobentheorie
- Amtliche Statistik:
 - Stichprobentheorie steht im Vordergrund
 - Datenintegration insbesondere von administrativen Daten wird in letzter Zeit wichtiger

Statistischer Zugang zu Daten – Sampling, Stichprobe

- Informatik liefert Werkzeuge für
 - Technischen Lösungen für Surveys
 - Integration von Daten (Record Matching)
- Nutzung mehr in der Amtlichen Statistik

Statistischer Zugang zu Daten – Sampling, Stichprobe

- Beachte den Unterschied Statistik - Informatik
 - Keine Annahme einer geschlossenen Welt, wie sie in der Datenbanktheorie meist verwendet wird



Statistische Modellierung

- Grundkonzept jeder statistischen Modellierung ist das Konzept der Verteilung eines Merkmals (Unsicherheit, Variabilität)
 - Unterscheide das vom Konzept des Fehlers in der Mathematik
- Operationalisierung beruht auf
 - Wahrscheinlichkeit
 - Modellierung
 - Statistical Computing
 - Darstellung der Ergebnisse

Statistische Modellierung – Wahrscheinlichkeit

- Beide Zugänge sind interessiert an
 - Verteilung von Merkmalen in einer Population
 - Im einfachsten Fall Angabe einer Verteilung, z.B. Normalverteilung, Binomialverteilung, Extremwertverteilung,...
 - Meist in Verbindung mit einem deterministischen Modell
 - Fluktuation eines Prozesses in der Zeit
 - Verteilung von Statistiken, z.B. Mittelwert, Anteile,.....

Statistische Modellierung – Wahrscheinlichkeit

- Unterschiede primär in der Begründung des Wahrscheinlichkeitsmodells
 - Methodische Statistik:
 - Zufall aus einer Vielzahl von möglichen Prozessen, die Daten generieren
 - Verteilungen einer potentiell unendlichen Grundgesamtheit, nicht exakt berechenbar

Statistische Modellierung – Wahrscheinlichkeit

- Amtliche Statistik:
 - Zentraler Prozess ist das Modell der Stichprobenziehung
 - Verteilung einer endlichen Grundgesamtheit, die prinzipiell mit marginalen Messfehlern exakt bestimmbar ist
 - Öffentlichkeit missversteht daher Amtliche Statistik oft als Buchhaltung

Statistische Modellierung – Modelle

- Beide Zugänge betonen funktionale Modelle für die Variablen
- Wesentlicher Unterschied sind traditionell
 - Spektrum der verwendeten Modelle
 - Verwendung der Modelle
- Beachte den Unterschied zur Informatik, die sich mehr mit logischen Modellen zwischen Merkmalen beschäftigt (Relationen)

Statistische Modellierung – Modelle

- Methodische Statistik
 - Eine Vielzahl von generischen Modellen zur Strukturbeschreibung
 - Strukturmodelle, z.B. Regression
 - Modelle zur Datenkompression, z.B. Cluster
 - Modelle zur Klassifikation
 - Zentrales Element ist oft die Vorstellung
 - Beobachtung = Erklärungsterm + Residuen
 - Modelle entstehen in einer Anwendung und werden auf andere Anwendungen übertragen

Statistische Modellierung – Modelle

- Amtliche Statistik

- Zentrale Modelle

- Hochrechnung,
 - Formal einfache Rechenmodelle z.B. Indizes

- Der Ansatz

Beobachtung = Erklärung + Residuum
spielt eine geringere Rolle

- Ausnahmen: Regression, Zeitreihenanalyse (Ökonometrie)

- Weniger breite Anwendungsfälle

Statistische Modellierung – Modelle

- In den letzten 20 Jahren verstärkt komplexe Modelle auch in der Amtlichen Statistik
 - Multiple Imputation for Nonresponse in Surveys
 - Rubin, 1987
 - Model assisted Survey Sampling
 - Särndal et al. 1992
 - Finite Population Sampling and Inference
 - Valliant et al., 2000
 - Small Area Estimation
 - Rao 2003

Statistische Modellierung – Modelle

- Synthetische Daten
- Statistical Matching
 - Anwendung der Imputation bei Datenintegration
- Komplexe Indikatoren, z.B. Pisa
 - OECD-Manual 2008 für Composite Indicators, eine Übersicht über klassische multivariate Statistik

Statistische Modellierung – Computing

- Traditionelles Statistisches Computing der methodischen Statistik konzentrierte sich auf die Umsetzung von Standardformeln
- Für die realen Probleme der Amtlichen Statistik vielfach nicht umsetzbar (komplexe Varianzschätzungen)
- Computerintensive Verfahren für Amtliche Statistik nützlich
 - Bootsstrap
 - Bayesian Methods
 - Simulation
 - Visualisierung

Statistische Modellierung – Computing

- Fortschritte der Amtlichen Statistik im Statistical Computing, die auch für die Methodische Statistik von Interesse sind
 - Datenrepräsentation und Metadaten
 - Datenerfassung
 - Disclosure control
- Techniken der Informatik wie Prozess Management oder Workflow Management sollten in beiden Bereichen mehr eingesetzt werden
 - Data Mining Software

Statistische Modellierung – Darstellung der Ergebnisse

- Statistik ist der *Versuch* die Welt so zu beschreiben wie sie ist (F. Ferschl?)
- Unterschiede zwischen Methodischer Statistik und Angewandter Statistik
 - Kunden
 - Anspruch

Statistische Modellierung – Darstellung der Ergebnisse

- Methodische Statistik
 - Experten, Wissenschaftler
 - Ergebnisse eines Modells sind eine Möglichkeit die Realität zu sehen
 - Höhere Akzeptanz der Zufälligkeit
 - Nicht nur „Objektivität“ sondern auch Nützlichkeit eines Modells wird bewertet
 - Signifikanz als wichtige Kenngröße eines Modells

Statistische Modellierung – Darstellung der Ergebnisse

- Amtliche Statistik
 - Breite Öffentlichkeit
 - Darstellung muss oft traditionellen Richtlinien folgen
 - Akzeptanz der Zufälligkeit
 - Konfidenzintervalle nicht vermittelbar?
 - Index als eine Schätzung
 - „Objektivität“ steht im Vordergrund (Randomisierung)
 - Ist ein Index oder die VGR objektiver als ein komplexeres Modell?

Integration Methodische Statistik – Amtliche Statistik

- Methodische Statistik und Amtliche Statistik gehen beide von einer Statistischen Betrachtung der Daten aus
 - Unterschiede in der Betrachtungsweise
 - Statistische Methodik ist mehr auf Prozesse und Modelle orientiert
 - Amtliche Statistik ist mehr auf Dokumentation orientiert
- Als Anwendung hat Amtliche Statistik historisch und auch in ihrer Breite eine Sonderrolle

Integration Methodische Statistik – Amtliche Statistik

- Beide haben entscheidende Beiträge zur Weiterentwicklung der Statistik in den letzten Jahren beigetragen
- Beide Betrachtungen sollten gemeinsam die Entwicklung der Statistik in Zukunft gestalten

Integration Methodische Statistik – Amtliche Statistik

- Wichtige Forschungsbereiche der Statistik in der Zukunft (David Hand, NTTS 2009)
 - massive data sets
 - anomaly detection
 - data quality *
 - new kinds of data *
 - Experiments and social policy *

Integration Methodische Statistik – Amtliche Statistik

- Datenqualität
 - Amtliche Statistik hat eine gute Methodologie zur Beurteilung von Datenqualität entwickelt
 - Weiterentwicklung in Richtung Prozessqualität
 - Unterstützung der Datenvorverarbeitung für alle statistischen Analysen
 - Traditionell wird Genauigkeit (accuracy) oft mit Varianz gleichgesetzt
 - Neue Verfahren sind für komplexe Daten notwendig (Simulation, Bootstrap)

Integration Methodische Statistik – Amtliche Statistik

- New kinds of data
 - Im Umweltbereich gibt es viele Daten und Fragestellungen, die mit herkömmlichen Modellen nicht adäquat analysiert werden können
 - Abfallstatistik: Geeignete Stichprobenmodelle, Analyse von seltenen Ereignissen, Anwendung der Räumlichen Statistik

Integration Methodische Statistik – Amtliche Statistik

- Experiments and social policy
 - Modelle zur Analyse von Daten, z.B. Modelle für qualitative Zeitreihen
 - Simulation
 - Statistische Modelle der Datenintegration
 - Klinische Studien im Zusammenhang mit administrativen Daten zur Gesundheit
- Mikrodaten für die Forschung bieten ein reiches Feld an Möglichkeiten



Vielen Dank für Ihre Aufmerksamkeit