Statistik, empirische Wissenschaften und Wissenschaftstheorie

Prof. Dr. Uwe Saint-Mont

Fachhochschule Nordhausen

München, 4.2.2011

Die Welt der Mathematik

Tukey (1961, Statistical and Quantitative Methodology):

> *[Mathematicians] work with abstract, uncertain entities (like a continuous function defined on* $[0,1]$*) just so long as the assumptions about them which are to be used are unquestioned (like continuity).*
>
> *Unquestioned assumptions lead to secure conclusions.*

Deduktion: Definition, Satz, Beweis

Beispiel:

Definition (Rechtwinkliges Dreieck)

Ein Dreieck mit einem rechten Winkel heißt *rechtwinkliges Dreieck*.

Theorem (Pythagoras)

In einem rechtwinkligen Dreieck ist das Quadrat der Hypothenuse gleich dem Summe der Quadrate der Katheten.

Beweis.

Siehe Euklids „Elemente“. □

Deduktion: Definition, Beweis, Satz

Allgemein:

1. Klar definierte Situation:
   Voraussetzungen - Axiome, Definitionen, spezifische Annahmen
2. Logische Ableitungen, Zusammenhänge
3. Gültige Folgerungen:
   Interessante Lemmata, Sätze, Theoreme und Korollare

Insbesondere:

Starke Voraussetzungen (enge Situation) $\Rightarrow$ Elegante Ergebnisse

Beispiel: Banach-, normierte, metrische und topologische Räume

Gefahren der Deduktion

Tukey (1961). The danger of mathematics to the outside world in general, and to science in particular, is simple:

- *Pure mathematics must take its assumptions most seriously, [...] questioning them not at all.*
- *Pure mathematics must value its results in its own terms, with far less attention to the relation of the assumptions to the real world than to the aesthetic nature of the results.*

Gefahren der Deduktion

Tukey (1961). The danger of mathematics to the outside world in general, and to science in particular, is simple:

- *Yet these are just what science and technology* **must not** *do. Science and technology [. . .] must use the products, and avail itself of the aid of mathematics, yet dare not accept its attitudes.*
- *Every clear-cut problem is artificial, separated from the real world by idealization after idealization.*

Gefahren der Deduktion

Weitere Kommentare zur deduktiven Str(enge):

- Nothing could be more pathetically mistaken than the prefatory claim [. . .] that mathematical rigor 'guarantees the correctness of the results'.
  On the contrary, much experience teaches us that the more one concentrates on the appearance of mathematical rigor, the less attention one pays to the validity of the premises in the real world, and the more likely one is to reach final conclusions that are absurdly wrong in the real world. (Jaynes, 2003)

- It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so. (Mark Twain)

Die Welt der empirischen Wissenschaften

- Primär: Empirische Erfahrung, insbesondere Beobachtungen und Experimente
- Gesucht: Mechanismen, verborgene Strukturen, allgemeine Gesetze
- Substanzielles Wissen steht im Vordergrund, Mathematik als (unentbehrliche) Hilfswissenschaft

Statt vom Allgemeinen zum Speziellen (wie die Deduktion) geht man im Wesentlichen vom Speziellen (bemerkenswerten Beobachtungen, spezifischen Bedingungen, einzelnen Experimenten, konkreten Daten) zum Allgemeinen

Induktion

- Keine logische Strenge oder Sicherheit
- Nur mehr oder minder gute Gründe - „Evidenz“
- Unvollständige Formalisierung, syntaktische und semantische Argumente
- Neues, das über den bisherigen Rahmen hinausgeht, kann berücksichtigt werden
- Fundamentaler, philosophischer Zweifel an der Induktion (Hume)

Humes Induktionsproblem

- Ein „Induktionsprinzip“ (zur Rechtfertigung induktiver Folgerungen) kann weder logisch noch empirisch begründet werden:

- Könnte es logisch-deduktiv begründet werden, wäre es nicht wirklich induktiv
- Würde man es empirisch begründen, argumentierte man zirkulär
- Man hat also nur die Wahl ein solches Prinzip als „transzendent“ (weder logisch noch empirisch gerechtfertigt) zu akzeptieren oder
- Jeder einzelne, konkrete induktive Schritt ist situationsspezifisch zu begründen, wobei aber immer ein prinzipieller Zweifel, eine nicht auszuräumende argumentative Lücke bleibt.

Statistik

Bewegt sich im Spannungsfeld von Deduktion und Induktion:

- Mathematische Statistik versus in den empirischen Wissenschaften verwendete statistische Methoden
- Mathematische Statistik als angewandte Mathematik versus Statistik als Wissenschaft und Kunst der Datenanalyse
- Geplante Datenerhebung (experimentelles Design, CDA) und kreative Auswertung (EDA, IDA, Data Mining)

Statistik und Wissenschaftstheorie: Forschungszirkel

Hypothesen
Setzungen $\nwarrow$ Substanzielle Generalisierung

Formale Ableitung $\swarrow$

Numerisches Resultat

Theorem

Operationalisierung $\searrow$ $\nearrow$ Formale Argumente

Daten

Aktuelle Statistik

Weitgehende Trennung der Aktivitäten:

- Akademisch geben mathematische Statistiker den Ton an
- Empirisch arbeitende Institute beschäftigen mathematisch hinreichend ausgebildete Hilfswissenschaftler zur Datenanalyse und statistischen Qualitätssicherung
- Die meisten Datenanalysen werden von „Anwendern“, also Fachwissenschaftlern oder Praktikern gemacht - ohne einen Statistiker zu Rate zu ziehen

Traditionelle Statistik

Weit größere Überschneidung von Mathematik, empirischer Forschung und Philosophie. Beispiel: R.A. Fisher

- ist berühmt als (mathematischer) Statistiker, v.a. wegen vieler entscheidender methodischer Innovationen (z.B. Likelihood)
- ist zugleich ein bekannter Genetiker
- benutzt Mathematik als Hilfswissenschaft. Am wichtigsten ist immer die gesamte Argumentation, z.B. Randomisierung
- Zentrale Idee: Extraktion von Informationen aus Daten. Beispiel: Parametrischer Verteilungsfamilien
- Die Extraktion funktioniert am zuverlässigsten dann, wenn man bereits im Vorfeld in die Erhebung der Daten „investiert", also ein geeignetes „experimentelles Design" verwendet.

Neyman, Fisher und die Bayesianer

- Bis ins 20. Jahrhundert: Bayesianische Statistik
- Fisher verlässt (engen) Bayesianischen Rahmen, gründet durchgängig mathematisierte Statistik mit eigener Begrifflichkeit (sufficient / ancilliary statistics, efficiency, consistency, reference set, fiducial probability) und induktiver „wissenschaftlicher“ Ausrichtung.
- Neyman gründet Mathematische Statistik (Stochastik steht im Mittelpunkt). Deduktive „mathematische“ Ausrichtung.
- Neo-Bayesianer ab den 1950er Jahren (De Finetti, Savage)
- Heftige Auseinandersetzung um die richtige Ausrichtung der Statistik bis zu Fishers Tod (1962) und darüber hinaus.

Immense Interpretationsunterschiede

Bjørnstad:
[The likelihood principle] is concerned with evaluation of information in the actual *data*, whereas frequentist evaluation is concerned with *method* performance in hypothetical repetitions of the experiment.

Friedman (Meine Hervorhebungen):

- Randomization provides a *known* distribution for the assignment variables; statistical inferences are based on this distribution.
- Modeling *assumes* a distribution for the latent variables; statistical inferences are based on that assumption. Furthermore, model-based inferences are *conditional* on the assignment variables and covariates."

Beispiel: Glaubwürdigkeits- / Konfidenzintervall

On the one hand, the *procedure* of using the sample mean (or some other measure) to estimate μ could be assessed in terms of *how well we expect it to behave;* that is, in the light of different possible sets of data that might be encountered. It will have some *average* characteristics that express the precision we *initially* expect, i.e. before we take our data [. . .] The alternative concept of *final precision* aims to express the precision of an inference *in the specific situation we are studying*. Thus, if we actually take our sample and find $\bar{x} = 29.8$, how are we to answer the question 'how close is 29.8 to μ'? This is a most pertinent question to ask - some might claim that it is the supreme consideration. Within the classical approach we must rest on any transferred properties of the long-term behaviour of the *procedure* itself. (Barnett)

Beispiel: Glaubwürdigkeits- / Konfidenzintervall

$X_1, \ldots, X_n$ iid mit Verteilung $N(\mu, \sigma)$; σ bekannt, μ unbekannt.

Priori-Sicht, Stichprobenraum, long run, hypothetical repetition, average characteristics, procedure, randomization, initial precision: $P(a \leq \bar{X} \leq b) = 1 - \alpha$.

Posteriori-Betrachtung, Parameterraum, konkrete Stichprobe, specific situation, modelling, final precision:
$P(a(x_1, \ldots, x_n; Prior(\mu)) \leq \mu \leq b(x_1, \ldots, x_n; Prior(\mu))) = 1 - \alpha$.

"[...] a confidence interval is a probability statement about the data, given the parameter, rather than one about the parameter, given the data." (Lindley)

Übergang von Priori- zu Posteriori-Perspektive

1. In der frequentistischen Statistik implizit: $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$.
2. In der Bayesschen Statistik explizit: Priori-Verteilung von μ wird mit der beobachteten Stichprobe $x_1, \ldots, x_n$ zur Posteriori-Verteilung von μ verrechnet.
3. Fishers Programm: Übergang von der Prä- zu Post-Betrachtung *ohne* das Bayessche Theorem. (Fiduzialargument)

It should be noted that the aposteriori interpretation of confidence intervals (and thus the implicit fiducial argument and a subconscious switch between aleatory and epistemic probability) was probably centuries old [...] (Hampel)

Theoretiker und Anwender

- 1960/1970er heftige Auseinandersetzungen zwischen Frequentisten und Bayesianern
- Gleichzeitig Auseinanderdriften von „Anwendern“ aller Gebiete und theoretischer Statistik (Tukey: EDA 1977, Leamer 1978)
- 1980er Offener Bruch zwischen Datenanalyse (Data Mining) und mathematischen Theoretikern
- Ab 1980er erste theoretische Ansätze von wissenschaftlicher Seite (Kausalität, Ökonometrie, Informatik)
- Seit 1990er: Ausreifung dieser Ansätze (Greenland, Heckman, Pearl, Rissanen)
- Parallel dazu: Etablierung der Bayesschen Schule (Valencia, Imprecise Probability, Dempster-Shafer-Theorie)

Die Situation heute

„Balkanisierte“ Statistik:

- viele Ansätze und Schulen, die eher nebeneinander existieren als miteinander kooperieren
- Auseinandersetzungen der letzten Jahrzehnte nicht wirklich, also konstruktiv-konzeptionell, überwunden
- Eklektizismus führt dazu, dass jeder den Formalismus wählt, der gerade am besten passt oder sich historisch durchgesetzt hat
- Theoretische Fundierung oder auch nur allgemeiner Konsens, was gute und schlechte Statistik/Datenanalyse ist, existiert kaum
- Mathematische Aspekte dominieren, wissenschaftstheoretische Aspekte spielen fast keine Rolle

Statistik der Zukunft?!

Wiedervereinigte, eigenständige Statistik

- Information als zentraler Begriff,
  Statistik = Informationswissenschaft (Efron)
- Theoretische und praktische Verflechtung mit der Informatik. Beispiele:

  Enger Zusammenhang zwischen Informationstheorie und Wahrscheinlichkeitstheorie,

  Erweiterungen der Wahrscheinlichkeitstheorie (Komplexitätsbegriff, kausale Netze)

  Informationserhebung und -Speicherung (Datenbanken, Informations- und Wissensmanagement, Automatisierung)
- Orientierung an empirischen Problemen, Mathematik als Hilfswissenschaft

Beispiel: Erweiterung der Parametrischen Statistik

- Notation $P_\theta(x)$ zeigt das Wesentliche: Wahrscheinlichkeit P ist nur ein Werkzeug um von der konkreten Beobachtung x zum Parameter θ zu kommen.
- Input: Konkrete Stichprobe $\mathbf{x}$. Output: (Genauere) Aussage über θ als zuvor - insofern Bayesianisch.
- Maximum Likelihood $L_{\mathbf{x}}(\theta)$ ist Standardmethode hierfür - insofern nicht Bayesianisch.
- Problem: Nichtlineare Regression $\theta_0 + \theta_1 z + \theta_2 z^2 + \ldots + \theta_k z^k$ mit den Parametern $\theta_1, \ldots, \theta_k$ und k. Sowohl die Werte θ_i als auch deren Anzahl k ist zu bestimmen.

Minimum Description Length

Aktuelle Lösung:

- Allgemeine Definition der Komplexität $K(\mathbf{x})$ eines Datensatzes $\mathbf{x}$.
- Wahl einer Modellklasse $M_\theta^k = M_{\theta_1,\ldots,\theta_k}$.
- Beschreibung der Daten mithilfe dieser Klasse (=Kodierung), so, dass die Gesamtlänge der Beschreibung minimiert wird.

- Zerlegung der Daten in einen von der Modellklasse erfassten Teil (learnable properties, systematischer Anteil) und einen unsystematischen (algorithmisch zufälligen) Teil.

- Optimal: Auswahl eines Modells, das zu den Daten passt und eher wenige Parameter besitzt.
- Vorteil: Universelle, theoretisch fundierte Methode; Parameter und Daten werden simultan berücksichtigt.

Emanzipation Proclamation (Jaynes)

Every variable x that we introduce is understood to have some set X of possible values. Every function $f(x)$ that we introduce is understood to be sufficiently well-behaved so that what we do with it makes sense. We undertake to make every proof general enough to cover the application we make of it.

It is the assigned homework problem for the reader who is interested in the question to find the most general conditions under which the result would hold.

Strategische Ausrichtung

Also nicht:

- Mathematik als Selbstzweck,
- Statistik als angewandte Stochastik

Sondern:

- Mathematik als Methode / Werkzeug / Sprache um mit logischen strengen Mittel zu analysieren und zu quantifizieren.
- Beispiele: Differential- und Integralrechnung (klassische Physik), Stochastik (Natur- und Sozialwissenschaften), "Causal Calculus" (Pearl), Nonprobabilistic Statistics (Li und Vitányi), ...
- Erweist sich eine Methode als ungenügend, sollte sie erweitert oder durch eine ganz andere Methode ersetzt werden

Entwicklung datenanalytischer Werkzeuge

- Klassisch: Familie $P_\theta(x)$, Priori-Verteilung $P(\theta)$. Bayessches Theorem liefert $P(\theta|x)$.
- Abschwächung durch Fisher zu (Max) Likelihood $L_\theta(x)$ und Fiduzialmethode (Aussagen über θ ohne Verwendung des Bayesschen Theorems).
- Neyman und Pearson, heute orthodoxe Statistik (mathematisch orientiert): Sehr schwache Annahmen, so dass kaum noch Aussagen über θ möglich sind (bspw. Konfidenzintervalle, Tests)
- Konsequente Weiterentwicklungen der Bayesschen Inferenz (z.B. imprecise probability, Nonparametric Bayes)
- In der Praxis entwickelte Methoden mit (noch) geringer theoretischer Fundierung (z.B. Scoring)

Vielen Dank für Ihre Aufmerksamkeit!