Anwendung der Mathematischen Statistik am Rechenzentrum für Forschungsaufgaben der Humboldt-Universität

In diesem Beitrag soll der Versuch unternommen werden, die langjährigen Erfahrungen der Abteilung Angewandte Mathematik des Rechenzentrums auf diesem Gebiet darzustellen und dabei einige allgemeine Prinzipien herauszuarbeiten. Zunächst möchte ich den Begriff Mathematische Statistik und deren Anwendung von dem der beschreibenden (oder deskriptiven) Statistik abgrenzen.

Die **beschreibende Statistik** beschränkt sich im wesentlichen auf das Sammeln und Aufbereiten von Daten sowie die Berechnung einfacher Kennzahlen (Häufigkeiten, Mittelwerte, Standardabweichungen, Extremwerte u.a.) und deren Darstellung als Tabellen oder Graphiken wie z.B. in Statistischen Jahrbüchern. Bei dieser Betrachtungsweise ist man z.B. kaum imstande, einen "kleinen" Unterschied zweier Mittelwerte dahingehend zu bewerten, ob er als "bedeutsam" (da vielleicht in der Tendenz erwartet) oder als "unbedeutend", also "zufällig" anzusehen ist. Eine Regel zur Entscheidung einer solchen eigentlich einfach erscheinenden Frage ist nicht Gegenstand der beschreibenden Statistik.

Die **Mathematische Statistik** ist dagegen eine mathematische Theorie, die auf der Wahrscheinlichkeitsrechnung aufgebaut ist. Das Grundprinzip besteht darin, daß ein Datenmaterial als Stichprobe, d.h. als zufällige Auswahl aus einer sogenannten "Grundgesamtheit" angesehen wird. Diese "Grundgesamtheit" ist im allgemeinen eine rein theoretische Konstruktion und wird durch eine "Wahrscheinlichkeitsdichte" beschrieben, die man als (allgemeines) "Modell für die Daten" ansehen kann. Dazu kommen dann die speziellen Modelle zur Entscheidung der zu prüfenden Hypothesen.

Die **Anwendung der Mathematischen Statistik** ist also die Anwendung einer mathematischen Theorie, um zu fundierten Antworten auf Fragestellungen aus der Forschung der Fachbereiche zu gelangen. Das bedeutet, daß diese Fragestellungen zunächst gemeinsam von dem Wissenschaftler des Fachbereichs und dem Mathematiker in die Gestalt mathematischer Modelle gebracht werden, welche man als geeignet vermutet. Nach den Berechnungen zu diesen Modellen müssen diese mathematischen Ergebnisse interpretiert und wieder in die Sprache des Fachwissenschaftlers übersetzt werden.

Nach diesen äußerst groben Umrissen zur Mathematischen Statistik und ihrer Anwendung möchte ich zwei Projekte mit sehr komplexen Datenbeständen skizzieren.

1. Bei dem Projekt aus dem **Institut für Tierzüchtung und Haustiergenetik** (Rinderzüchtung) geht es u.a. um die Untersuchung von Möglichkeiten der züchterischen Einflußnahme auf die Eutergesundheit. Gleichzeitig wird dieses Datenmaterial noch für andere wissenschaftliche Untersuchungen genutzt, beispielsweise für die Schätzung der Depression (Verminderung der Werte) bei den Milchleistungsmerkmalen infolge Mastitiserkrankung, aber auch zur Analyse des Zusammenhangs von Farbmerkmalen (Färbung des Tieres) und der Milchleistung.
Aufgrund von Untersuchungen in vielen leistungsfähigen Milchrindpopulationen muß angenommen werden, daß eine höhere Milchleistung der Kühe vermehrt Eutererkrankungen zur Folge hat. Neben prophylaktischen (Hygiene) und therapeutischen (Antibiotikabehandlungen) Maßnahmen werden deshalb züchterische Möglichkeiten zur Verbesserung der natürlichen Abwehrmöglichkeiten der Milchkühe in Erwägung gezogen. Die eigenen Untersuchungen hatten das Ziel, geeignete Selektionsmerkmale für die Einbeziehung in ein Zuchtprogramm zur Verbesserung der Resistenz gegen Eutererkrankungen zu finden.
Grundlage für die Einschätzung und Beurteilung der Brauchbarkeit eines Selektionsmerkmales sind mit hoher Zuverlässigkeit geschätzte genetische Parameter, die nur aus großen Stichproben geschätzt werden können.
Eine Resistenzzüchtung verbessert nicht nur die Wirtschaftlichkeit der Milchproduktion, sondern leistet auch einen positiven Beitrag zur Ökologie, indem der Medikamenteneinsatz reduziert werden kann. Damit kommt man der Verbraucherforderung nach Produkten von gesunden und unbehandelten Tieren entgegen.
Zur Beantwortung der o.g. Fragestellung liegt ein umfangreiches Datenmaterial von 15 000 Tieren vor, welches sich aus verschiedenen Dateien zusammensetzt, die in vielfältiger Weise miteinander verbunden werden. Das sind vor allem

- die Datei der Leistungs- und Qualitätsmerkmale der Milch pro Tier (Mütter bzw. Töchter) und pro Laktation (bis zu 8 Laktationen pro Tier), also eine Datei mit Längsschnittcharakter

- die Datei der Krankheitsmerkmale (Längsschnittcharakter)
- die Datei der Väter
- die Datei der Farbmerkmale bei ausgewählten Tieren
- zwei weitere, sehr spezielle Dateien (monatliche Milchleistung bzw. Eiweißwerte).

2. Zu dem Projekt aus dem **Institut für Anthropologie** gehören 12 sehr unterschiedliche Dateien, darunter drei Längsschnittdateien, wobei noch Daten aus den alten Bundesländern hinzukommen werden.
Das gesamte Datenmaterial (66 000 Probanden in Querschnittsstudien, 6 300 in Längsschnittstudien) dient wissenschaftlichen Untersuchungen u.a. zu folgenden Problemen:
- Verlauf der körperlichen Entwicklung im Wachstumsalter
- regressive Zuordnung von speziellen Bekleidungsmaßen zu aktuellen Leitmerkmalen
- Studien über Kopf- und Kieferwachstum, auch hinsichtlich der säkulären Akzeleration
- Ermittlung von Normwerten, Variabilitätsbereichen und Bewertungskriterien hinsichtlich Körperzusammensetzung und Umwelteinflüssen
- Analyse individueller Entwicklungsverläufe aus Längsschnittdateien
- Ermittlung der Körperendhöhe, auch im Zusammenhang mit der säkulären Akzeleration
- Standardisierung von Normwerten und Variabilitäten zum Körperbau.

Mit dieser nur stichpunktartigen Schilderung der Aufgaben aus beiden Instituten, denen natürlich weitere Aufgaben aus anderen Fachbereichen hinzugefügt werden könnten, ist es vielleicht gelungen, die Bedeutung der Dienstleistung des Rechenzentrums darzustellen.

Es sollen nun einige Bemerkungen zu den **Phasen bei der Anwendung der Mathematischen Statistik** folgen.
Zuerst ist stets die **logische Struktur des Datenmaterials** zu beschreiben, insbesondere seine Gruppenstruktur (bzw. Schichtungen), gegebenfalls noch Bedingungen von Zuordnungen (z.B. Schlüsselmerkmale), wenn mehrere Dateien in irgendeiner Weise zusammengeführt werden müssen.
Daneben erfolgt die **Präzisierung** (und Anpassung an die Datenstruktur) **der fachspezifischen Fragestellungen** sowie ein erster Versuch zur Wahl eines allgemeinen Modells.
Nach diesen Festlegungen "als erste Näherung" erfolgen nun oftmals umfangreiche **Datenprüfungen**, die mit den die logische Struktur beschreibenden Merkmalen beginnen müssen. Es folgt die Prüfung der "normalen" Merkmale (nicht zu verwechseln mit normalverteilten Merkmalen). Nachdem hierbei zutage getretene Fehler behandelt worden sind, erfolgt die Prüfung aller aus den ursprünglichen Merkmalen zu berechnenden neuen Merkmale oder Hilfsgrößen, die zum Zwecke der Nachprüfung irgendwelcher Relationen zwischen den Merkmalen gebildet werden. Es würde zu weit führen, die verschiedenen Methoden der Prüfung, der Fehlermitteilung und der Fehlerbehandlung zu beschreiben. Nach der Korrektur der "letzten" Fehler werden stets einige **Übersichtsauswertungen** vorgenommen, die zur Beschreibung der Datei sehr nützlich sind. Zu diesem Thema möchte ich aus dem Buch von Günter Bollinger und anderen: BMDP Statistikprogramme für die Bio-, Human- und Sozialwissenschaften, Fischer-Verlag 1983, Seite 2, zitieren:

> "Statistiklehrbücher vermitteln oft fälschlicherweise den Eindruck, als bestünde die statistische Analyse nur aus der Entscheidung für eine statistische Methode und der Auswahl des entsprechenden Computerprogramms. In der Praxis ist dies nur der letzte Schritt im Rahmen der Datenverarbeitung. Voraus gehen mehr oder weniger umfangreiche Arbeiten, die sich unter den Stichwörtern Datenprüfung und Datenaufbereitung zusammenfassen lassen. Diese Arbeiten beanspruchen normalerweise ein Vielfaches an Zeit gegenüber der eigentlichen Prüfung der Forschungshypothese bzw. Fragestellung. Sie sollten mit aller Sorgfalt durchgeführt werden, da kein noch so hoch entwickeltes statistisches Verfahren Fehler oder Verzerrungen in den Daten später ausgleichen kann."

Nach der erfolgten Bereinigung der Datei steht nun der Vergleich der beobachteten Datenstruktur mit der in dem allgemeinen Modell bzw. in der Grundgesamtheit geforderten Datenstruktur zur Debatte. Zur jetzt anstehenden Auswahl der speziellen Modelle möchte ich aus dem Buch von Dieter Rasch: Biometrie Einführung in die Biostatik, Verlag Harri Deutsch, 1989, Seite 18, zitieren:

> "Zunächst muß entschieden werden, ob auf eine Klasse dieser Modelle zurückgegriffen werden kann oder ob ein davon abweichendes Modell entwickelt werden muß... Viel diffiziler ist dann schon die Wahl des speziellen Modells. Es gibt noch keine ausgearbeitete Theorie der Modellwahl, die praktischen Ansprüchen gerecht wird. Hier muß auf die **Erfahrung** zurückgegriffen werden. Bei der Auswahl des Modells geht es nicht darum, ein sogenanntes "richtiges" Modell zu finden, sondern vielmehr darum, ein für die Aufgabenstellung "pas-

sendes" Modell auszuwählen, d.h. ein Modell, das die Versuchsergebnisse hinreichend genau zu beschreiben gestattet. Gerade in der Auswahl eines solchen Modells besteht die **Kunst der Anwendung statistischer Verfahren**...
Eine ungeeignete Modellwahl ist nicht, wie bei manueller Aufbereitung oft möglich, sofort erkennbar."

Zuletzt noch einige Worte zur Interpretation der Ergebnisse.
Zu einer Begutachtung der Drucklisten durch einen Statistiker ist unbedingt zu raten, ganz besonders bei allen multivariaten Verfahren mit ihren vielen Varianten. Hierzu noch ein Zitat von G. Bollinger:

"Die Auswertung der Drucklisten dürfte dem Benutzer dann Probleme bereiten, wenn er Programme rechnet, die seine statistischen Kenntnisse übersteigen. Da hochkomplexe statistische Techniken durch einfache Kommandofolgen abgerufen werden können, besteht die Gefahr, daß der unerfahrene Benutzer bei der Interpretation der Ausgabe überfordert ist."

Zum Schluß noch eine allgemeine Bemerkung.

Der Umstand, daß jedes Modell eine erhebliche Vereinfachung gegenüber dem "Original" darstellt und außerdem mit einer Reihe von Bedingungen belegt ist, sollte keine Überbewertung von Testentscheidungen erlauben. Lothar Sachs schreibt in seinem Buch: Angewandte Statistik, Springer Verlag, 1992, S. 14 :

"Nicht die statistische Signifikanz, sondern die praktische Relevanz zählt."

Andreas Baudisch