Betrachtungen über Statistik und Computer

1 Die Statistik als Arbeitsgebiet eines Universitätsrechenzentrums

Seit den sechziger Jahren, als Computer ihre Rolle zu spielen begannen, ist die Statistik ein wichtiges Arbeitsfeld der Hochschulrechenzentren. Das begründet sich durch die spezielle Position der Statistik im Gesamtgefüge der Wissenschaften. Stärker als viele andere Fächer hat die Statistik eine integrative Aufgabe zu erfüllen, denn sie wird in den empirischen Wissenschaften vielfältig zur Datenverdichtung, zur graphischen Veranschaulichung von Resultaten, zur Erkennung von Trends und zur Ursachenforschung benutzt. Das Besondere der Statistik im Vergleich zu anderen Hilfsdisziplinen und auch zu anderen Bereichen der Mathematik ist, daß sie bereits auf einem niedrigen Niveau der Modellierung anwendbar ist. Die Statistik basiert auf sehr allgemeinen und oft sehr einfachen Prinzipien wie

- der Genauigkeitserhöhung durch Mittelwerts- oder Medianberechnung (Gesetz der großen Zahlen),
- der Tendenz zur Entstehung normalverteilter Zufallsgrößen bei der Summation vieler Einflüsse (zentraler Grenzwertsatz),
- der Annahme, daß sich Abhängigkeiten zwischen Variablen näherungsweise durch lineare Funktionen beschreiben lassen.

Somit erfordert Statistik für ihre Nutzung nicht die vollständige Kenntnis der Bewegungsgesetze der betrachteten Objekte. Für die statistische Analyse ist es oft ausreichend, wenn neben den auszuwertenden Daten, Informationen über die formal-organisatorische Struktur der durchgeführten Experimente zur Verfügung stehen, wobei entsprechend der Anwendungssituation in jedem Fall noch zusätzliche Angaben über den Typ der Verteilung der Daten, über bestehende stochastische Unabhängigkeiten usw. notwendig sind.

Es kann nicht übersehen werden, daß sich in den letzten Jahren wesentliche Veränderungen in der Tätigkeit der Universitätsrechenzentren auch in der Frage der Statistik vollzogen haben. Die große Masse der statistischen Auswertungen läuft heute nicht mehr wie früher auf zentralen Rechenanlagen, sondern hierfür werden die PCs der Wissenschaftler- und Laborarbeitsplätze sowie die in den Instituten installierten Workstations eingesetzt. Heute bemüht sich das Rechenzentrum stärker um die Verbesserung des gesamt-universitären Rahmens für die Statistik. Durch Aufbau passender Rechnernetze an der Universität wird der Zugriff zur Statistik-Software erleichtert. Ebenso wie bisher werden vom Rechenzentrum Beratungen und Schulungen zu Themen der Statistik angeboten. Und für sehr rechenintensive und komplizierte Optimierungs- und Simulationsaufgaben im Zusammenhang mit der Statistik sind die zentralen Rechner des Universitätsrechenzentrums nach wie vor verfügbar.

In den ersten Jahrzehnten des Bestehens des Rechenzentrums der Humboldt-Universität sind große Anstrengungen unternommen worden, das wissenschaftliche und das technologische Niveau der statistischen Arbeiten zu erhöhen. Wenn Aktivitäten dieses Umfangs auch heute nicht mehr erforderlich sind - es gibt heute eine Reihe von professionell gestalteten Systemen der Statistik- und Graphik-Software -, so bleibt doch die grundsätzliche Aufgabe der gesamten Universität bestehen, die Methoden der Statistik in voller Breite und auf hohem Niveau verfügbar zu halten. Allein schon die Feststellung, daß im internationalen Maßstab ein Wissenschaftszweig "computational statistics" existiert, der seine Impulse aus der tiefen Problematik der Kopplung von statistischen Theorien mit Methoden der Informatik und der praktischen Datenverarbeitung empfängt, sollte für das Universitätsrechenzentrum Anlaß sein, die Frage der Beherrschung der Statistik sehr ernst zu nehmen.

2 Die enge Beziehung von Statistik und Computer

Im folgenden sollen einige Entwicklungen der letzten Jahrzehnte skizziert werden, die die Untrennbarkeit der modernen Statistik und der Computertechnik unterstreichen.

<u>Das Problem der Minimierung bzw. Maximierung einer nichtlinearen Funktion:</u> Viele statistische Fragestellungen führen auf das Problem, auf numerischem Wege eine Maximum-Likelihood-Schätzung zu bestimmen. Beispiele hierzu sind die nichtlineare Regression, die Faktoranalyse und allgemeinere Strukturrelationsanalysen, die logistische Regression und die logistische Diskrimination. Als Lösungsverfahren stehen die Newton-Raphson-Prozedur und die Quasi-Newton-Methode sowie auch der EM-Algorithmus zur Verfügung. Die gleiche Numerik wird beim Minimum-Quadrat-Ansatz der nichtlinearen Regression benutzt.

<u>Die Anwendung von Resampling-Methoden:</u> In der modernen Statistik wird in vielfältiger Weise die

Strategie des Erzeugens von Teilstichproben einer gegebenen Stichprobe (cross-validation, jack-knife, bootstrap) zum Einsatz gebracht. Die Verfahren dienen dazu, die Verzerrung bzw. die Streuung von nichtlinearen Schätzungen zu bestimmen, und sie erlauben, Schätzungen mit reduzierter Verzerrung zu konstruieren. In der Diskriminanz- und der Regressionsanalyse werden Cross-Validation-Methoden für die annähernd erwartungstreue Ermittlung des Entscheidungsfehlers und zur Modellwahl benutzt. Die Resampling-Verfahren erfordern einen hohen Rechenaufwand.

Die Anwendung von verteilungsfreien statistischen Testverfahren: Neben den "klassischen" Testmethoden der parametrischen Statistik, die zum großen Teil auf dem Verteilungsmodell der Normalverteilung basieren, gewinnen die Testverfahren der nichtparametrischen, verteilungsfreien Statistik zunehmende Bedeutung. In diese Verfahren gehen die Beobachtungen meistens nur in diskretisierter Form oder in Form ihrer Rangzahlen ein. Dabei werden spezielle mathematische Methoden und spezielle Techniken zur Generierung und Beherrschung der sämtlichen Permutationen der Stichprobenelemente zum Einsatz gebracht. Die Verfahren benötigen eine hohe Rechenleistung.

Verteilungsfreie statistische Verfahren, die auf Dichteschätzungen basieren: Bei einer Reihe von statistischen Problemen wird in letzter Zeit verstärkt die Dichteschätzung mittels Kernfunktionen angewandt. Dieses Vorgehen bewährt sich u.a. bei Regressions- und Diskriminationsaufgaben, bei denen man nicht von speziellen Verteilungsannahmen Gebrauch machen möchte. Die Verfahren setzen allerdings etwas größere Stichproben voraus.

Anwendung von Simulationen für statistische Problemstellungen: Bei der Lösung theoretischer oder praktischer statistischer Probleme erweist es sich oftmals als nützlich, die jeweilige Entscheidung durch einen Rechenalgorithmus zu simulieren. Insbesondere für die Klärung noch ungelöster theoretischer Fragen kann dieser Weg sehr hilfreich sein; durch Simulationen werden Hypothesen ohne großen wissenschaftlichen Aufwand bekräftigt oder widerlegt, so daß die spätere Untersuchung erleichtert wird. So stellt sich die Methode der Computersimulation als neue Dimension unseres Denkens dar. Bei praktischen Anforderungen, z.B. bei der Bestimmung kritischer Werte einer Teststatistik oder bei der Ermittlung der für ein Problem notwendigen Stichprobenumfänge, können durch Simulationen in einfacher Weise Näherungswerte gewonnen werden.

3 Nachträgliche Anmerkungen zu unserer Arbeit

Die Frage der numerischen Genauigkeit: Die heute üblichen Numerikprozessoren lassen enorme Rechengenauigkeiten zu. Einerseits ist das erfreulich, andererseits sollte man nicht glauben, allein mit Hilfe hoher numerischer Präzision algorithmische und statistische Schwierigkeiten lösen zu können. Eine übertriebene Genauigkeit entspricht nicht dem Charakter der Statistik als einer auf niedrigem Niveau der Modellbildung bei beschränktem Datenumfang betriebenen empirischen Untersuchung. Wenn einige statistische Verfahren dazu neigen, Ungenauigkeiten zu zeigen, so sollte man sich die Frage stellen, ob die Art der Modellierung und der benutzte Lösungsweg dem Problem adäquat sind. Gute, stabile Algorithmen benötigen keine übermäßige Rechengenauigkeit. Es gibt Ausnahmen von dieser Regel, etwa die Situation, daß schon die Eingangsdaten nur mit hoher Stellenzahl darstellbar sind, oder daß riesige Stichprobenumfänge vorliegen, oder daß bei numerischen Hilfsrechnungen Approximationen mit hohen Genauigkeitsanforderungen nicht vermeidbar sind (wie z.B. bei der Berechnung von p-Werten der F- oder der Chiquadrat-Verteilung).

Keine übertriebene Anpassung an die Daten: In den statistischen Verfahren sollte die Anpassung an die Daten dort aufhören, wo sich die Anwendungsrelevanz der Ergebnisse durch "Übermodellierung" verschlechtert. Das gilt insbesondere für die Variablenselektion in den multivariaten Verfahren. Eine zu große Variablenzahl ist sinnlos oder sogar schädlich.

Nicht nur Anwendung der traditionellen multivariaten Verfahren: Die traditionellen multivariaten Verfahren, die auf geometrischen Vorstellungen basieren, besitzen den Nachteil, mit wachsender Dimension immer instabiler zu werden. Die als Ausweg vielfach eingesetzten traditionellen Variablenselektionsverfahren haben ihrerseits den Mangel, die Möglichkeit der gegenseitigen Glättung von korrelierten Variablen zu ignorieren. Daher sollte man den Blick auch auf redundanznutzende, stabilisierende Algorithmen richten.

Die Frage der Bevorzugung verteilungsgebundener oder verteilungsfreier Tests: Wenn die Mittelwerte zweier Populationen zu vergleichen sind, kann die Frage, ob der t-Test oder der Mann-Whitney-Wilcoxon-Test anzuwenden wäre, nicht immer eindeutig beantwortet werden. Beachtet man nur die Güte der Tests, so ist die Anwendung der verteilungsfreien Methode zu empfehlen, denn die Güte

des Mann-Whitney-Wilcoxon-Tests ist nie wesentlich schlechter als die des t-Tests, der Mann-Whitney-Wilcoxon-Test kann aber bei speziellen Verteilungen erheblich besser sein. Bezieht man aber auch andere, mehr wissenschaftsstrategische Gesichtspunkte in die Betrachtung ein, so kann durchaus auch der t-Test bevorzugt werden, denn er steht in einer sehr einfachen Beziehung zu den beschreibenden Größen Mittelwert und Streuung und besitzt darüber hinaus vielfältige, sehr übersichtliche Verallgemeinerungen im Gesamtgebäude der Statistik (Varianzanalyse, multivariate Analyse).

Zur Relativität der Rolle von Theorie und Praxis bzw. von strenger Mathematik und computerbezogener Empirie: Der Computer ist heute aus der Wissenschaft nicht mehr wegzudenken. Das gilt in besonders hohem Maße für die Statistik. Aber es bleibt auch richtig, daß noch so intensives Rechnen nicht die durch Sätze und Theorien gesicherte mathematische Erkenntnis überflüssig machen kann. Jedes praktische Resultat ist eine neue Erfahrung, ist aber dennoch kein endgültig anerkanntes Wissen. Das sollte uns immer wieder ermuntern, neben datenbasierten Forschungen auch prinzipielle theoretische Untersuchungen durchzuführen.

Jürgen Läuter*, Magdeburg

* Der Autor ist seit 1990 Professor an der *Medizinischen Akademie Magdeburg* und leitet dort die *Abteilung für Biomathematik und Medizinische Informatik.* Am Beginn seiner wissenschaftlichen Laufbahn gehörte er zu der kleinen Gruppe von jungen Mathematikern, die im Jahr 1964 das Rechenzentrum an der Humboldt-Universität zu Berlin (damals am II. Mathematischen Institut) gründeten. Er hat durch seine Arbeiten, bei denen er - ganz im Sinne der abschließenden Worte dieses Beitrags - stets bestrebt war, über die Alltagsroutine hinauszuweisen, wesentlich dazu beigetragen, der elektronischen Datenverarbeitung in Lehre und Forschung an der Humboldt-Universität Geltung und Ansehen zu erwerben. Ihm ist es zu verdanken, daß sich die Arbeitsrichtung der rechnergestützten Mathematischen Statistik an unserem Rechenzentrum auf Dauer etabliert hat. Diesem Arbeitsgebiet blieb er treu, als er 1970 das Rechenzentrum verlassen und eine Tätigkeit an der Akademie der Wissenschaften der DDR aufgenommen hatte, wie auch die Verbindung zwischen ihm und Mitarbeitern unseres Rechenzentrums nie zum Erliegen kam. Ein beredtes Zeugnis dafür ist dieser Gastbeitrag, über den wir uns ganz besonders freuen.

PS:
Von Herrn Professor Dr. Jürgen Läuter ist jüngst das Buch

Stabile multivariate Verfahren
Diskriminanzanalyse - Regressionsanalyse - Faktoranalyse
Berlin: Akademie Verlag, 1992, 208 S., DM 68,-

erschienen.

Edmund Suschke