EDV in Medizin und Biologie **14** (2), 33–39, ISSN 0300-8232

EDV-unterstützte Übungen in naturwissenschaftlichen Studiengängen

G. Janke-Grimm

Zusammenfassung

In naturwissenschaftlichen Studiengängen werden in den ersten Semestern mathematische und statistische Grundkenntnisse vermittelt. Ein größerer Lernerfolg wird erzielt, wenn der Vorlesungsstoff durch schriftliche Übungen vertieft wird. Die Durchführung solcher Übungen ist bei großen Studentenzahlen nur mit EDV-Anlagen möglich. In diesem Beitrag wird eine Gruppe von Programmen vorgestellt, die einerseits Übungsaufgaben erstellen und andererseits die Ergebnisse der Studenten auf ihre Richtigkeit hin überprüfen. Es wird auf die Besonderheiten der Übungssimulation und der Ergebnisprüfung eingegangen.

Summary

In a course of studies of the natural sciences the first semesters offer an introduction into the basic knowledge of mathematics and statistics. To enhance the understanding thereof the lectures are accompanied by written exercises. The application of these exercises for a large number of students is only possible with the aid of computer-systems. This paper suggests the use of a group of programs which provides exercises and controls the students' results. In particular, the simulation of exercise and the check-out of results are discussed.

1. Einleitung

Biometrie, genauso wie die grundlegenden Fächer Mathematik und Statistik, kann von Nichtmathematikern am besten durch ständiges Üben an verschiedenen Aufgaben erlernt werden. Aufgrund unterschiedlicher Studiengänge und Prüfungsordnungen ersetzen solche im Semester durchgeführten Übungen nicht eine abschließende Leistungskontrolle in Form einer mündlichen oder schriftlichen Prüfung, sondern sind vielmehr so konzipiert, daß die Studenten den in der Vorlesung behandelten Stoff selbständig üben. Ein größerer Lernerfolg wird erzielt, wenn alle Studenten das gleiche Aufgabenkonzept erhalten, in das unterschiedliche Zahlenwerte eingesetzt sind. Dadurch muß jeder Teilnehmer zwangsläufig selbst die Rechnung durchführen.

Die Erstellung und Überprüfung solcher Übungen ist bei großen Studentenzahlen, wie sie gerade in den unteren Semestern auftreten, nur durch den Einsatz von EDV-Anlagen durchführbar. Bereits seit 1967 werden computererstellte Übungen im Institut für Statistik und Biometrie der Tierärztlichen Hochschule Hannover verwendet (RUNDFELDT und AUKES, 1970). In den letzten Jahren wurde ein den gesamten Lehrplan umfassendes Übungssystem aufgebaut und in den verschiedenen naturwissenschaftlichen Studiengängen eingesetzt.

Das Prinzip der Übungen ist, daß die Studenten den in der vorherigen Stunde gelehrten Stoff anhand einer oder mehrerer Aufgaben selbständig üben. Dabei können sie das Vorlesungsskript und eigene Aufzeichnungen benutzen und die anwesenden Betreuer fragen. Auch die Zusammenarbeit der Studenten untereinander ist zulässig.

2. Darstellung des Themenkreises der Übungen

Vom Institut für Statistik und Biometrie der Tierärztlichen Hochschule Hannover werden verschiedene naturwissenschaftliche Studiengänge betreut, in denen Biomathematik, Mathematik, Statistik oder Biometrie gelehrt wird und in denen das bestehende Übungssystem eingesetzt wird:
- Veterinärmedizin (Biomathematik im 1. Semester)
- Diplombiologe (Mathematik und Statistik im 4. Semester)
- Gartenbau (Mathematik und Statistik im 1. und 2. Semester, Biometrie im 3. und 4. Semester).

Im weiteren soll nur auf die Übungen eingegangen werden, die für die Fachrichtung Gartenbau erstellt wurden. In den anderen Fachrichtungen werden ähnliche Unterrichtsinhalte mit vergleichbaren Lernzielen vermittelt.

Der Vorlesungsstoff und die korrespondierenden Übungen werden in der nachfolgenden Übersicht dargestellt.

Mathematik und Statistik im 1. Semester

Kapitelüberschrift des Vorlesungsstoffes	Übungsthema
1. Einleitung	
2. Zahlen und Verknüpfungsarten	1. Σ- und π-Zeichen 2. Mischungsrechnung 3. Wurzeln und Logarithmen
3. Folgen und Reihen	4. Folgen und Reihen
4. Zinseszins- und Rentenrechnung	5. Zinseszins-Aufgaben 6. Tilgungsrechnung
5. Mengenlehre	7. Verknüpfungs-Aufgaben
6. Kombinatorik und Multinome	8. Grundbegriffe 9. Binome/Multinome 10. Binom bei großem n
7. Wahrscheinlichkeitsrechnung	11. Empirische Verteilung 12. Poisson-Verteilung
8. Statistik	13. Mittelwert, Streuung 14. χ^2-Verfahren 15. t-Test oder F-Test
9. Rechnen mit ungenauen Zahlen	

Mathematik und Statistik im 2. Semester

Kapitelüberschrift des Vorlesungsstoffes	Übungsthema
10. Vektoren und Matrizen	16. Vektorrechnung 17. Matrizenmultiplikation 18. Matrixinversion
11. Analyse von Gleichungssystemen	19. Gleichungssystem (3·3) 20. Überbestimmtes Gleichungssystem 21. Lineare Optimierung
12. Funktionen	22. Pentaden und Gleitmittel 23. Regression oder Korrelation 24. Multiple Regression
13. Interpolationsverfahren	25. Quadratische oder kubische Interpolation

Biometrie im 3. Semester

Hier werden lediglich die Themen der Vorlesungsstunden wiedergegeben, da für jedes Thema eine entsprechende Übung existiert.

Themen der Vorlesungsstunden
1. Häufigkeitsverteilungen
2. Maßzahlen einer Verteilung
3. Normalverteilung: Schiefe, Exzeß und Potenzmomente einer Stichprobe
4. Prüfverteilungen und Freiheitsgrade
5. Das Testen von Hypothesen
6. Die Varianz von Stichprobenfunktionen
7. t-Test
8. Einfaktorielle Varianzanalyse
9. Multiple Vergleiche von Mittelwerten
10. Das varianzanalytische Modell
11. Die zweifaktorielle Varianzanalyse (einfache Besetzung)
12. Die zweifaktorielle Varianzanalyse (mehrfache Besetzung)
13. Modelle der zweifaktoriellen Varianzanalyse mit mehrfacher Besetzung
14. Hierarchische Varianzanalyse

Biometrie im 4. Semester

Themen der Vorlesungsstunden
1. Lateinisches Quadrat
2. Streuungstest
3. Lineare Regressionsanalyse bei zwei Variablen (Modell I)
4. Lineare Regressionsanalyse bei zwei Variablen: Varianzen, Vertrauensbereiche und Tests
5. Regressionsanalyse – Modell II
6. Partielle lineare Regressionsanalyse
7. Nichtlineare Regressionsanalyse
8. Korrelation und partielle Korrelation

Für die Vorlesungen und die begleitenden Übungen im 1. und 2. Semester wird ein ausführliches Skriptum von Prof. Dr. Rundfeldt herausgegeben, da der Stoff dieser beiden Semester in der gewünschten Zusammenstellung in keinem Lehrbuch vorhanden ist.

Die aufbauenden Vorlesungen und Übungen des 3. und 4. Semesters sind stark am deutschsprachigen Lehrbuch »Grundlagen der Statistik« von E. WEBER orientiert. Zusätzlich werden Umdrucke verteilt, in denen Schwerpunkte des Vorlesungsstoffes übersichtlich dargestellt sind.

3. Prinzipieller Aufbau der Übungen

3.1 Allgemeiner Ablauf einer Übung und ihre Vor- und Nachbereitung

Zu Semesterbeginn werden die Namen aller Studierenden erfaßt, mit einer vierstelligen Kenn-Nummer versehen und in einem permanenten Plattenfile gespeichert. Dieser File bleibt während des gesamten Semesters erhalten und dient dazu, Übungsergebnisse und erzielte Zensuren zu speichern.

Das Erstellen der Übungsbogen mit den zu bearbeitenden Aufgaben erfolgt per Computer. Dazu werden die gespeicherten Personenangaben vom Plattenfile abgerufen und in der Kopfzeile des Übungsblattes ausgedruckt. Gleichzeitig werden die Ergebnisse der einzelnen Übungsaufgaben für jeden Teilnehmer errechnet und gespeichert.

Die von den Studenten bearbeiteten Übungsbogen werden per Computer auf ihre Richtigkeit hin überprüft. Dabei werden vom Plattenfile die gespeicherten Personenangaben und die Ergebnisse der Übungsaufgaben abgerufen. Nach dem Überprüfen aller Aufgaben wird eine Zensur vergeben, die wiederum auf Platte gespeichert wird.

Am Ende des Semesters stehen im permanenten File außer den Personenangaben je Teilnehmer die Ergebnisse aller Übungsaufgaben und die erzielten Zensuren. Diese Angaben können dazu verwendet werden, Leistungsnachweise auszustellen.

Die Übungsaufgaben werden jeweils im Anschluß an die Vorlesungen verteilt. Wenig rechenintensive Übungen werden in der anschließenden Übungsstunde durchgeführt. Bei aufwendigen Rechnungen, wie sie zum Teil in der Biometrie vorkommen, können die Aufgaben zu Hause gelöst werden. Die Studenten tragen die Ergebnisse der einzelnen Aufgaben unter Voranstellung der ihnen zugewiesenen Kenn-Nummer in einen Markierungsbogen ein.

Mittels eines Markierungslesers werden die eingestrichenen Markierungen in Lochkarten übertragen und per Computer geprüft. Es gibt Anerkennungskriterien für jede Übung, nach welchen die folgenden Zensuren vergeben werden:

1 = fehlerfrei oder mit geringer Anzahl an Fehlern, so daß die Übung anerkannt wird
2 = so viele Fehler in den Ergebnissen, daß die Übung nicht anerkannt wird
0 = Übung wurde nicht abgegeben.

Sobald mindestens ein Aufgabenergebnis als falsch erkannt wurde, erfolgt ein Korrekturausdruck mit dem jeweils richtigen Computerergebnis und dem Ergebnis des Studenten in Verbindung mit der Angabe, ob die Übung anerkannt wurde oder nicht. Die Markierungsbogen und gegebenenfalls die Korrekturbogen werden den Studenten in der nächsten Übungsstunde zurückgegeben, wobei in der Regel eine Besprechung der Aufgaben stattfindet.

Die vergebenen Zensuren können am Ende des Semesters – wie bereits erwähnt – als Grundlage für eine Erfolgsbescheinigung verwendet werden. Dazu wird festgelegt, wieviele Übungen abgegeben und wie viele davon anerkannt sein müssen, damit die Bescheinigung ausgestellt wird.

Die Verwendung von Markierungsbogen zum Eintragen der Ergebnisse ist zunächst für die Studenten ungewohnt. Das Verfahren wird aber relativ schnell angenommen und erlernt. Es scheint keine echte Alternative für diese Art der Eintragungen zu geben. Ein Ablochen der Ergebnisse von handschriftlichen Eintragungen auf den Übungsblättern ist zeitaufwendig und birgt außerdem die Gefahr in sich, daß ein richtiges Rechenergebnis durch fehlerhaftes Ablochen als »falsch«

gewertet wird. Die Benutzung von Klarschriftbelegen könnte eine Erleichterung bedeuten. Voraussetzung dafür ist jedoch eine deutliche Normschrift, die ebenfalls erlernt werden muß. Erfahrungen unseres Institutes haben gezeigt, daß das exakte Schreiben der Normzahlen schwerer ist und häufiger zu Fehlern führt als die Verwendung von Markierungsbogen. Somit werden zunächst bei uns die Ergebnisse auch weiterhin in Markierungsbogen eingetragen. Dabei sollte aber gerade in der ersten Übungsstunde die Handhabung der Bogen besonders gut erklärt werden.

3.2 Programmablauf zum Erstellen einer Übung

Die folgenden Angaben über programmiertechnische Einzelheiten beziehen sich auf die Rechenanlage MODCOMP 7870 mit dem Betriebssystem MAX IV und die Programmiersprache FORTRAN IV.

Die Programme, welche die Übungsaufgaben erstellen und drucken, folgen alle einem Schema, wobei der eigentliche Rechenteil von Programmteilen umschlossen wird, in denen das Auffinden des Plattenfiles und das Drucken der Übungsbogen sowie der Ergebnisse auf einem Extraausdruck organisiert wird. In der Abbildung 1 ist das allgemeine Flußdiagramm dieser Programme dargestellt.

Die Simulation der Zahlenwerte für die Übungsbogen erfolgt entsprechend dem Typ der einzelnen Aufgabe mittels gleichverteilter oder normalverteilter Daten. Damit Übungsaufgaben mit annähernd gleichem Niveau für alle Studenten erstellt werden, muß die Simulation immer in vorgegebenen Grenzen erfolgen, die in Abhängigkeit von der Art der Aufgabe festzusetzen sind. Zu enge Grenzen bewirken unter Umständen identische Zahlenwerte und Ergebnisse für mehrere Studenten. Hingegen bergen zu weite Grenzen die Gefahr, daß Aufgaben mit unterschiedlichem Schwierigkeitsgrad gestellt werden und der Bereich der möglichen Ergebnisse zu groß wird.

Bei jeder Aufgabe muß überlegt werden, in welcher Weise die Simulation erfolgen soll. Gerade bei Aufgaben aus dem Bereich der Regressionsrechnung oder bei Varianzanalysen müssen die Einzelwerte nach einem genauen Plan simuliert werden, der vielfach eine Umkehrung des exakten Rechenweges ist. Durch diese Art der Simulation wird sichergestellt, daß sowohl die Zahlenwerte als auch die zu berechnenden Ergebnisse innerhalb eines vorgegebenen Bereiches liegen. Als Beispiel soll hier die Simulation für eine lineare Regression zwischen zwei Variablen x und y genannt werden. Aufgabe ist es, den Regressionskoeffizienten b und den Achsenabschnitt a zu berechnen. Die Simulation beginnt damit, eine bestimmte Anzahl von gleichverteilten Werten, den x-Werten, zu erstellen. Danach werden der Regressionskoeffizient b und der Achsenabstand a mittels gleichverteilter Daten in vorgegebenen Grenzen (z. B. b: 2–4 und a: 10–30) simuliert. Die Berechnung der y-Werte erfolgt nach der Regressionsgleichung $y = a + bx$. Erst nach diesen Schritten liegen die Wertepaare x/y vor, die in der Aufgabe ausgedruckt werden. Aus Gründen der Genauigkeit müssen nun aus diesen Wertepaaren die Kenngrößen b und a der Regressionsgleichung mittels der Summe der Abweichungsquadrate, der Summe der Abweichungsprodukte und der Mittelwerte brechnet werden. Dieses ist notwendig, da die zuerst simulierten Größen für b und a von den später berechneten abweichen. Der Unterschied ist in vielen Fällen minimal, kann aber beim Prüfen der Ergebnisse zu einer Fehlentscheidung führen.

Eine weitere Schwierigkeit für das Programmieren solcher Übungen entsteht ebenfalls durch die Simulation. Die erzeugten Zahlen werden mit der rechnerinternen Stellengenauigkeit simuliert und verrechnet. Die meisten Aufgaben verwenden aber ganzzahlige Werte bzw. Zahlen mit einer oder zwei Kommastellen. Somit müssen die Werte vor einer weiteren Verrechnung auf die gewünschte Stellenzahl gerundet werden. Diese Rundungsprozedur wird für alle End- und Zwischenergebnisse durchgeführt, um möglichst präzise zu dem Ergebnis zu gelangen, das der Student mit Hilfe eines Taschenrechners errechnen kann.

Häufig kommt es dazu, daß Zwischenergebnisse nicht mehr zu runden sind, da sie den INTEGER-Zahlenbereich überschreiten. Die benutzte Rundungsprozedur addiert zu den berechneten Zahlen 0,55, wandelt in INTEGER um und dann zurück in REAL und erreicht damit eine Rundung, die den allgemein gültigen Forderungen genügt.

Sollen Ergebnisse beispielsweise auf zwei Stellen gerundet werden, erfolgt eine Modifikation der Rundungsprozedur:

```
ERG  = ERG × 100. + 0.55
IERG = ERG
ERG  = IERG
ERG  = ERG/100.
```

Hier wird deutlich, daß der INTEGER-Bereich sehr schnell überschritten wird und ein Runden von Zwischen- und Endergebnissen nicht mehr möglich ist. Dadurch entstehen Ergebnisse, die erheblich von denen abweichen, die manuell oder mittels Taschenrechner erarbeitet werden.

Außerdem benutzen die Studenten Taschenrechner mit unterschiedlicher Rechengenauigkeit. Die daraus resultierenden Abweichungen vom Computerergebnis müssen ebenfalls berücksichtigt werden.

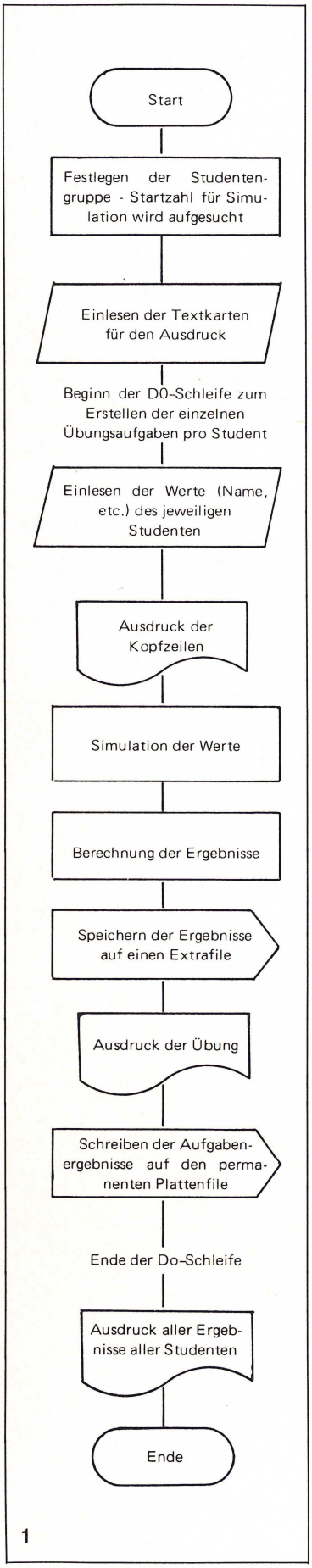
3.3 Programmablauf zum Prüfen der Ergebnisse

Die Programme, die die Ergebnisse der Studenten überprüfen, folgen einem gleichen Schema. Die Ergebnisse und die vierstellige Kenn-Nummer stehen nach der Verarbeitung der Markierungsbogen auf Lochkarten. Vom Plattenfile werden die gespeicherten Ergebnisse des Rechners eingelesen.

Im mittleren Programmteil werden die beiden Ergebniswerte miteinander verglichen, und es wird entschieden, ob das Ergebnis des Studenten richtig ist bzw. als richtig anerkannt werden kann oder ob falsch gerechnet wurde. Alle anderen Programmteile steuern das Lesen der Daten und das Drucken der Korrekturbogen. Abbildung 2 zeigt das allgemeine Flußdiagramm eines solchen Programmes.

Wie bereits aufgezeigt, kommt es aufgrund von unterschiedlichen Genauigkeiten der verwendeten Taschenrechner sowie nicht zu rundenden Zwischen- und Endergebnissen zu Abweichungen zwischen den Computerergebnissen und den Ergebnissen der Studenten, auch wenn »richtig« gerechnet wurde. Diese Differenzen müssen beim Prüfen beachtet werden. Für jedes Ergebnis wird deshalb eine Prüftoleranz vorgegeben, die in der Regel 1 bis 5 % des vom Computer brechneten Wertes beträgt. Sie muß groß genug sein, um alle Rundungsungenauigkeiten aufzufangen, d. h. sie nicht als Fehler zu werten. Andererseits darf sie nicht zu groß sein, da sonst auch echte Rechen- bzw. Verfahrensfehler »nur« als Rundungsungenauigkeit angesehen werden könnten. Die Größe der Prüftoleranz ist abhängig von der Art und Länge des Rechenganges und vom Zahlenwert des Endergebnisses. Führt eine langwierige Rechnung über viele Zwischenwerte zu einem zahlenmäßig kleinen Ergebniswert, muß die Prüftoleranz relativ groß sein.

Fehlentscheidungen bei der Ergebnisprüfung werden von



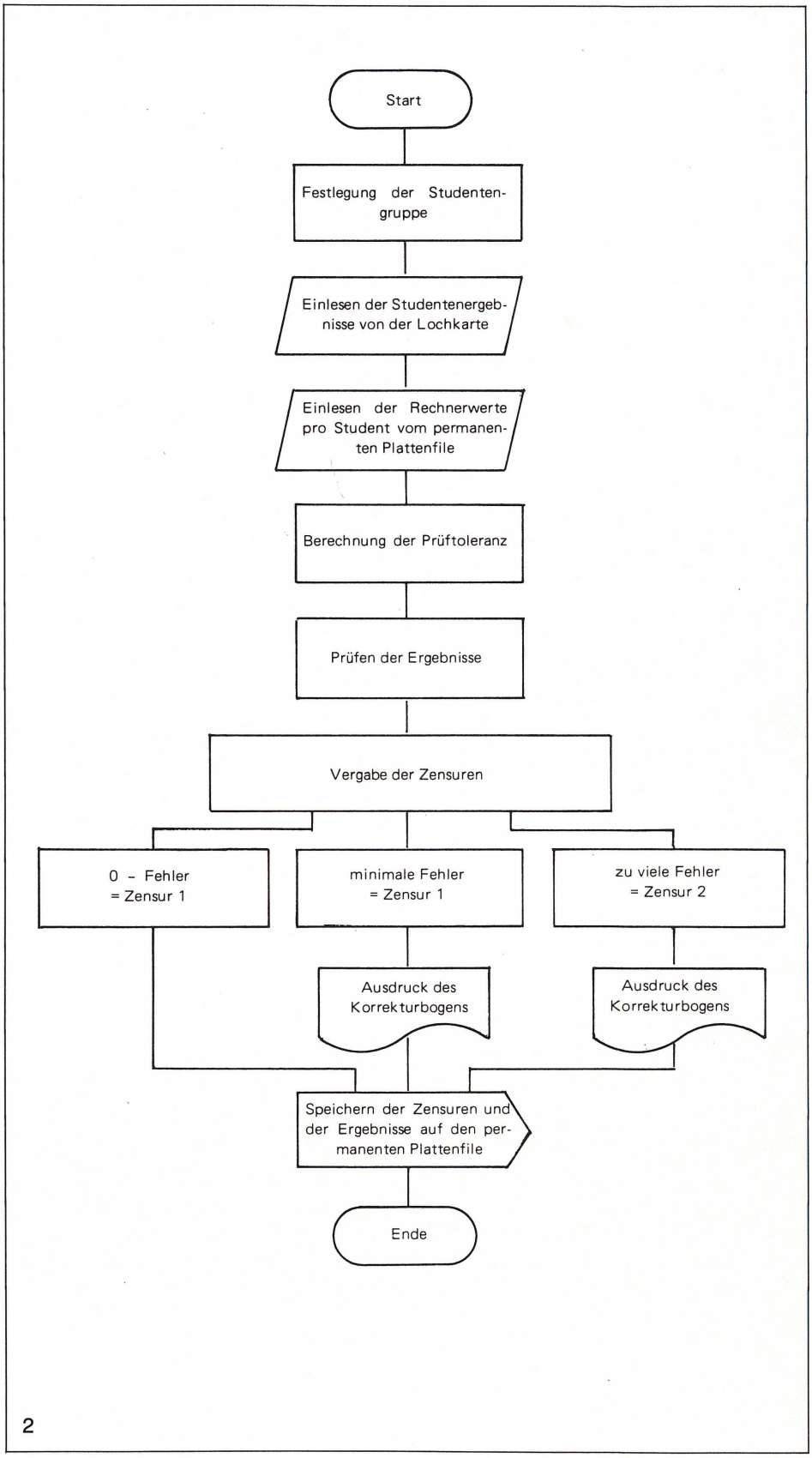


Abb. 1. Allgemeines Flußdiagramm der Programme zum Erstellen einer Übung.

Abb. 2. Allgemeines Flußdiagramm der Programme zum Prüfen einer Übung.

```
     UNIVERSITAET HANNOVER                          SOMMER-SEMESTER 1980
      BIOMETRIE                  UEBUNG NR. 11                    GRUPPE  0

       SEHR GEEHRTER HERR THOMAS BECKER ,                          201-3001

  ES WURDE AN 7 BLAETTERN EINES GUMMIBAUM DIE BLATTBREITE (X)
  UND DIE BLATTLAENGE (Y) IN MM GEMESSEN.
  WIE LAUTET DIE LINEARE BEZIEHUNG ZUR VORHERSAGE DER BLATTLAENGE
  AUFGRUND DER BLATTBREITE UND ZUR VORHERSAGE DER BLATTBREITE
  AUFGRUND DER BLATTLAENGE.

  BLATTBREITE   63.7   71.4   79.1   86.7   94.4  102.1  109.8
  BLATTLAENGE  125.3  132.8  133.4  157.1  154.0  163.9  177.0

  BITTE BERECHNEN SIE FUER DIE REGRESSIONSGERADE VON Y AUF X
  DIE Y-WERTE DER ELLIPSE FUER DIE ERSTEN 3 X-WERTE BEI EINER
  IRRTUMSWAHRSCHEINLICHKEIT VON ALPHA = 5 PROZENT.

  BITTE STREICHEN SIE IHRE KENN-NUMMER 3001 IN WORT 1-4
  UND DIE ERGEBNISSE WIE FOLGT.
  ERGEBNISSE                                          FORM       WORT
  REGRESSION Y AUF X. REGRESSIONSKOEFFIZIENT B         XYYY      5- 8
                      ACHSENABSCHNITT A1               XXXY      9-12
  VORZEICHEN VON A1 (+ = 1 UND - = 2)                            13
  REGRESSION X AUF Y. REGRESSIONSKOEFFIZIENT B         XYYY     14-17
                      ACHSENABSCHNITT A2               XXXY     18-21
  VORZEICHEN VON A2 ( + = 1 UND - = 2 )                          22
  DIFFERENZ VON A1 UND A2 (0 = 1 UND UNGLEICH 0 = 2 )            23
  BITTE STREICHEN SIE BEI DEN FOLGENDEN WERTEN
  IMMER DEN GROESSEREN WERT ZUERST.
  X-WERT 1        Y1 UND Y2                            XXX      24-26, 27-29
  X-WERT 2        Y1 UND Y2                            XXX      30-32, 33-35
3 X-WERT 3        Y1 UND Y2                            XXX      36-38, 39-41
```

Abb. 3. Ausdruck für die Übung zum Thema Regressionsanalyse Modell II.

den Studenten nur selten reklamiert. Im allgemeinen werden sie nur dann erkannt, wenn das Rechenergebnis des Studenten richtig war, aber aufgrund der Prüftoleranz als falsch ausgewiesen wurde. Der umgekehrte Fall, daß ein Ergebnis als richtig anerkannt wurde, obwohl es falsch war, aber innerhalb des Toleranzbereiches lag, wird zumeist nicht bemerkt und schon gar nicht reklamiert. Somit wird beim Prüfen der Ergebnisse mit den festgelegten Prüftoleranzen ein gewisses Risko eingegangen. Da es sich bei dem beschriebenen System um Übungs- und nicht um Prüfungsaufgaben handelt, sollten diese Fehlentscheidungen nicht überbewertet werden.

4. Spezielles Problem

In diesem Abschnitt soll anhand eines Beispiels aus dem Fach Biometrie im 4. Semester das im vorigen Abschnitt Beschriebene erläutert werden.

Thema der Übung ist die Regressionsanalyse Modell II. In den vorangegangenen Vorlesungen und Übungen wurden die Regressionsanalyse Modell I mit einfacher und mehrfacher Besetzung und das Testen der verschiedenen Parameter behandelt. Die nun folgende Aufgabe besteht darin, eine Regressionsanalyse durchzuführen, bei der sowohl die abhängige Größe (x) als auch die unabhängige Größe (y) Zufallsvariablen sind. Unter der Annahme, daß die Meßwerte von x und y einer bivariaten Normalverteilung folgen, entsteht um die beiden möglichen Regressionsgeraden je ein eliptischer Vertrauensbereich. Die Grenze dieser Bereiche kann in Abhängigkeit einer vorgegebenen Irrtumswahrscheinlichkeit und der entsprechenden Meßwerte nach folgender Formel berechnet werden:

$$(1/(1-r^2))\,(x_N^2 - 2\,r x_N \cdot y_N + y_N^2) = \chi_\alpha^2$$

Nach dem Auflösen der Gleichung entsteht für jeden unabhängigen Wert ein Paar von abhängigen Werten, wobei jeweils ein Wert oberhalb und einer unterhalb des vorhandenen unabhängigen Wertes liegt.

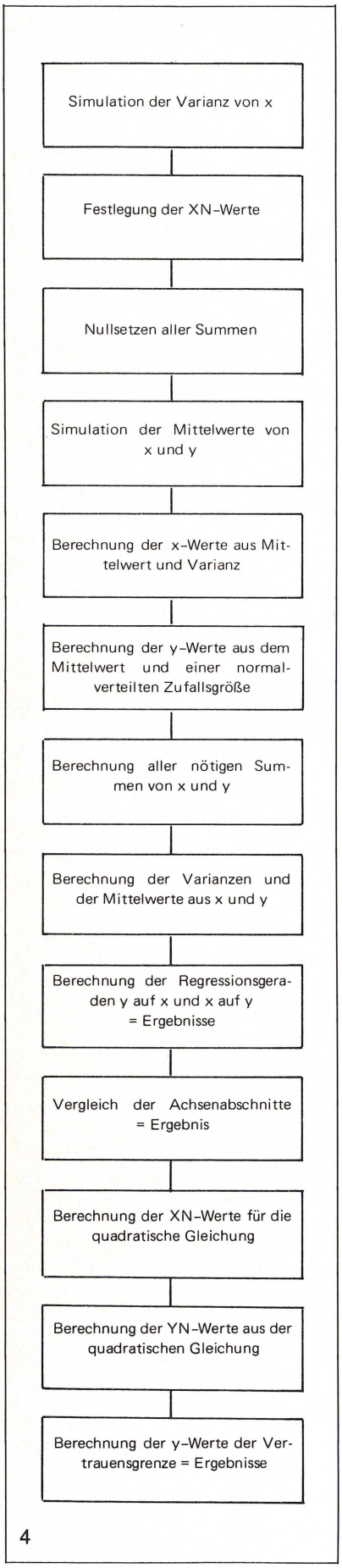


Abb. 4. Flußdiagramm für die Übung zum Thema Regressionsanalyse Modell II. Es ist nur der Programmteil dargestellt, in dem die Zahlenwerte simuliert und die Ergebnisse berechnet werden.

UNIVERSITAET HANNOVER SOMMER-SEMESTER 1980

BIOMETRIE UEBUNG NR. 11 GRUPPE 0

SEHR GEEHRTER HERR THOMAS BECKER , 201-3001

ZWISCHENERGEBNISSE DER BERECHNUNG FUER DIE ELLIPSENPUNKTE

X = X-WERTE

Y = Y-WERTE

XN = (X - X-QUER)/S(X)

YN1 = 1. LOESUNG DER QUADRATISCHEN GLEICHUNG

YN2 = 2. LOESUNG DER QUADRATISCHEN GLEICHUNG

YG = Y-WERT OBERHALB DER GERADEN

YU = Y-WERT UNTERHALB DER GERADEN

X	63.700	71.400	79.100	86.700	94.400	102.100	109.800
Y	125.300	132.800	133.400	157.100	154.000	163.900	177.000
XN	-1.386	-0.922	-0.458	0.000	0.464	0.928	1.392
YN1	-0.819	-0.306	0.177	0.631	1.068	1.481	1.864
YN2	-1.859	-1.475	-1.062	-0.631	-0.171	0.312	0.826
YG	134.000	143.000	153.000	161.000	169.000	177.000	185.000
YU	114.000	121.000	129.000	137.000	146.000	155.000	165.000

5

Abb. 5. Ergebnisausdruck der Übung zum Thema Regressionsanalyse Modell II.

Abbildung 3 zeigt den Übungsausdruck. An den x- und y-Werten ist zu erkennen, daß zwischen ihnen ein linearer Zusammenhang besteht, wenn die Regression von y auf x berechnet wird. (Auf die genauen Anweisungen zum Eintragen der Ergebnisse in Markierungsbogen, die im unteren Teil der Aufgabe stehen, soll hier nicht näher eingegangen werden.)

Das Flußdiagramm dieser Übung ist in Abbildung 4 dargestellt. Dabei ist lediglich der Teil des Programmes gezeigt, daß im Grobdiagramm (Abbildung 1) mit den Begriffen »Simulation der Werte« und »Berechnung der Ergebnisse« belegt ist. Die anderen Programmteile sind, wie bereits erwähnt, rein organisatorischer Art und dienen der Erstellung von Daten und dem Ausdruck der Übungen.

Die Simulation und Durchführung der Rechnung ist ein Beispiel dafür, daß es bei einer solchen Aufgabe nicht möglich ist, x- und y-Werte unabhängig voneinander und unabhängig von den gewünschten Endergebnissen zu simulieren. Die Erstellung der Ausgangsdaten beginnt vielmehr mit der Simulation eines Zwischenergebnisses und der Festlegung der anderen Zwischenergebnisse (x_N-Werte). Die y-Werte der Vertrauensbereiche werden vielfach falsch berechnet, so daß ein Sonderteil in diese Übung eingefügt wurde. Jeder Student erhält einen Ergebnisausdruck, in dem die Ausgangswerte, Zwischen- und Endergebnisse tabellarisch dargestellt sind (siehe Abbildung 5).

5. Diskussion

Die Erfahrungen mit dem dargestellten Übungssystem waren in den letzten Jahren durchweg positiv. Besonders für das

Fach Biometrie läßt sich sagen, daß nach Einführung der Übungen in den anschließenden mündlichen Prüfungen von den Studenten bessere Noten erzielt wurden. Die meisten Prüflinge hatten ein fundiertes Wissen über die verschiedenen statistischen Verfahren erworben, das sicherlich auch auf das eigene Rechnen komplizierter Aufgaben zurückzuführen ist.

In Ergänzung zu den Übungen werden in der Regel 1 bis 2 Multiple-Choice-Tests pro Semester durchgeführt, die ebenfalls dazu beitragen, die Kenntnisse über einzelne Verfahren und über Zusammenhänge zwischen verschiedenen Sachverhalten zu vertiefen. Diese Multiple-Choice-Tests werden gleichfalls mittels EDV erstellt und ausgewertet (SCHWARZ, AUKES und REDEKER, 1976). Aus einem Fragenkatalog wird für jeden Studenten eine Anzahl von Fragen per Computer ausgewählt. Die Fragen werden jeweils mit fünf zufallsmäßig angeordneten Antworten ausgedruckt, von denen aber jeweils nur eine Antwort richtig ist. Der Student trägt als Ergebnis die Nummer derjenigen Antwort auf dem Markierungsbogen ein, die seiner Meinung nach richtig ist.

Die Übungsteilnehmer müssen das Ergebnis ihrer Rechnungen in computerlesbarer Form darstellen. Wie in Abschnitt 3.1 ausgeführt, scheint das Eintragen der Zahlenwerte in die Markierungsbogen die beste Möglichkeit zu sein, um Übertragungsfehler zu vermeiden und außerdem den institutsinternen Aufwand gering zu halten. Dennoch ist das Prüfen der Übungen und die damit verbundene Zensurenvergabe der arbeitsintensivste Teil des Übungssystems. Ist es dem Dozenten nicht möglich, diesen Arbeitsaufwand zu erbringen und kann er auf das Erstellen von Leistungsnachweisen verzichten, so bietet sich die folgende, weniger arbeitsintensive Modifikation des Übungssystems an:

Die Übungsbogen werden wie beschrieben erstellt und an die Teilnehmer der Vorlesung erteilt. Die Kontrolle der Rechenergebnisse wird nicht im Institut vorgenommen, sondern den Studenten selbst überlassen. Als Hilfe dazu kann in einer späteren Vorlesung ein Ergebnisbogen mit den Ausgangsdaten, Zwischen- und Endergebnissen verteilt werden. Dieser Bogen würde dem in Abbildung 5 dargestellten Ausdruck entsprechen und könnte sofort im Anschluß an die Übungsbogen mitgedruckt werden. Der Vorteil dieser Modifikation liegt darin, daß das arbeitsintensive Prüfen der Ergebnisse nach jeder einzelnen Übung entfällt und demzufolge alle Übungen einschließlich der Ergebnisbogen im voraus gedruckt werden können. Der Nachteil besteht darin, daß der Dozent keine Kontrolle darüber hat, ob die Übungen überhaupt gerechnet wurden und wie viele Einzelergebnisse richtig waren bzw. bei welchen Aufgaben vorrangig Rechenfehler auftraten. Außerdem entfällt die Möglichkeit, die Beteiligung an den Übungen als Grundlage für eine Erfolgsbescheinigung am Ende des Semesters zu verwenden (siehe auch Abschnitt 3.1). Die Übungen sind dann lediglich als ein Angebot an die Studierenden zu betrachten, Rechengänge zu trainieren.

Ein Problem, das immer wieder in den einzelnen Jahrgängen auftrat, soll abschließend behandelt werden. Häufig haben die Studenten die Übungen der Vorjahre als Muster und bearbeiten die aktuellen Aufgaben nicht aufgrund eigener Überlegungen, sondern vorwiegend nach dem Schema der Vorjahresübungen. Ein Versuch, dieses Problem einzuschränken, bestand darin, alle Übungen in zwei Versionen bereitzuhalten, die sich geringfügig voneinander unterscheiden und abwechselnd eingesetzt werden. So soll beispielsweise im 1. Semester bei Kapitel 8, Übung 15 in Version 1 entweder der t-Test (Mittelwertvergleich) durchgeführt werden oder für die gleichen Ausgangsdaten in Version 2 der F-Test (Varianzhomogenitätstest) berechnet werden. Dadurch wird die Möglichkeit, »nach Schema« zu rechnen, eingeschränkt, und die Studenten müssen, auch wenn die Vorjahresübungen bekannt sind, ein Mindestmaß an eigener Rechenarbeit in die Übung einbringen. Dennoch ist die derzeitige Situation nicht zufriedenstellend. Aus diesem Grund ist geplant, laufend neue Übungen in das System aufzunehmen und die Anzahl der möglichen Übungsaufgaben zu den genannten Themen zu erhöhen. Die Erweiterungsmöglichkeiten dieses Übungssystems sind bei weitem nicht ausgeschöpft.

Abschließend läßt sich feststellen, daß die Ausarbeitung eines Übungssystems, wie es hier vorgestellt wurde, seine Pflege und Erweiterung lohnend ist. Probleme, die beim Programmieren selbst auftreten, können recht gut gelöst werden, da die Simulation und Berechnung einzelner Zahlenwerte kontrolliert werden kann (s. Abschnitt 3). Den Studenten wird einerseits eine Übungsmöglichkeit geboten, anhand derer sie selbst überprüfen können, ob sie den Vorlesungsstoff verstanden haben. Andererseits erhält der Dozent eine objektive Kontrollmöglichkeit über die Leistungen der Studenten.

Literatur

RUNDFELDT, H., und E. AUKES (1970): Zum Einsatz von Datenverarbeitungsanlagen für die Durchführung und Auswertung von Hochschulprüfungen. EDV in Medizin und Biologie **2,** 43–51.

SCHWARZ, R., E. AUKES und R. REDECKER (1976): Erfahrungen mit der Durchführung studienbegleitender Leistungskontrollen (Multiple-Choice-Fragen) im Unterricht der Gewebe- und mikroskopischen Organlehre. EDV in Medizin und Biologie **4,** 131–139

WEBER, E. (1980): Grundriß der biologischen Statistik. 8. Auflage, Gustav Fischer Verlag, Stuttgart.

Eingegangen am 17. Dezember 1982.

Anschrift der Verfasserin: Dipl.-Ing. agr. Gisela Janke-Grimm, Institut für Statistik und Biometrie der Tierärztlichen Hochschule Hannover, Bischofsholer Damm 15, D-3000 Hannover 1.

EDV in Medizin und Biologie **14** (2), 40–44, ISSN 0300-8232

Computer model of archive and diagnosis of brain tumours based on the WHO classification

J. R. Iglesias, M. J. Sanchez, A. Sendra and A. Mohnhaupt

Summary

1500 Brain tumours were studied histologically with two principal techniques: Haematoxilin-Eosin and Mallory Trichrom stainings. In each tumour 45 histological characteristics were collected. A Personal-Computer Commodore/8032 with a dual drive floppy disk/8052 was used. The programming language used was Commodore Microsoft Basic. The tumours were classified into the 15 statistically significant groups according to the WHO classification. Using the Bayes' theory of decision, the discrimination function ($G_i(X)$) and the probability »a posteriori« of the 15 diagnoses or clusters were calculated. These methods allow a mathematical diagnosis of the tumours with probabilities between 72.77 and 92.70 % and their objective archives based on the existence or absence of histological characteristics.

Zusammenfassung

Es wurden 1500 Hirntumoren histologisch durchuntersucht mit zwei Färbemethoden: Hämatoxilin-Eosin und Mallory-Trichrom. In jedem Tumor wurden 45 histologische Merkmale erfaßt mit Hilfe eines Personal-Computers, Commodore/8032 und einem dual drive floppy disk/8050 in Basic programmiert. Die Tumoren wurden in den 15 statistisch häufigsten Diagnosen gruppiert, wie sie in der WHO-Klassifikation üblich sind. Aufgrund der Bayesschen Theorie wurden die Diskriminierungsfunktionen ($G_i(X)$) und die Wahrscheinlichkeit »a posteriori« der 15 verschiedenen Diagnosen berechnet. Mit Hilfe dieser Methode wird eine mathematische Tumordiagnose mit 72,77 % bis 92,70 % Wahrscheinlichkeit erreicht, dadurch basiert ein objektives Archiv auf der Existenz der histologischen Merkmale.

Introduction

The morphological classification of the Brain Tumours serve multiple purposes:

1. To distinguish between different tumour-processes with different clinical units.
2. To facilitate the morphological and clinical comparison between the observations of different authors.
3. To allow to establish a prognosis of tumour illness in each patient.
4. To evaluate the different techniques for clinical, neuroradiological diagnosis and neurosurgical methods.

In the literature there are various »classical« classifications of the tumours of the Central Nervous System. Sometimes they consider only cytological aspects, in other cases architectural patterns, in other embryological similarities. Elective histological methods as silver impregnations are only used in few laboratories. Any of these »classical« classifications use similar and consistent morphological criteria and they are not applicable to all laboratories of Pathology or Neuropathology. Under these circumstances the subjective interpretation utilizing different techniques after different schools establish in many cases a multiplicity of diagnoses in one tumour. In these cases morphological identification and comparison of tumours and the principal finalities of tumour classification are difficult to standardize.

The World Health Organization [14] felt that a standardized nomenclature was necessary. ZÜLCH [15] emphasized actually the »Babel-like discrepancies« existing in the terminology of the intracranial tumours. The actual classification of the WHO cannot solve the problems of interpretation. Very often the tumours consist of a mixture of various cells and are difficult to classify.

The aim of the present work is, with the help of a personal computer, to find out if the different tumour groups of the WHO classification are identical to the groups or clusters based on the same histological characteristics using same histological and traditional methods.

Material and methods

1500 Brain Tumours were histologically studied with two principal techniques: Haematoxilin-Eosin and Mallory Trichrom (only in a few cases were other Trichromstainings used). In each tumour 1–10 slides were studied and the existence or absence of 45 traditional histological characteristics were collected. These histological characteristics are based on different authors [2, 3, 6–8, 11, 13] and defined specifically.

1. Nodular Architecture: Arrangement of tumorous tissue in large isolated and morphological-similar cell groups.
2. Papillary Architecture: Prominent mesenchymal columns completely surrounded by tumour cells.
3. Cystic Architecture: Presence of cystic spaces, larger than cell-vacuoles with or without content.
4. Diffuse Infiltration: Diffuse arrangement of tumour cells without limitation to normal tissue.
5. Nidus formation of nuclei: The nuclei are tightly arranged in small groups.
6. Rows of nuclei: Arrangement of nuclei in single or multiple rows, independent of nuclear axis.
7. Diffuse distribution of nuclei: The arrangement of nuclei is diffuse, without an obvious organization.

8. Reticular Arrangement: The tumour cells are building a meshwork.
9. Fascicular Arrangement: Arrangement of tumour cells in parallel and fascicular bundles.
10. Concentric Arrangement: Arrangement of tumour cells in concentric structures around a centre.
11. Perivascular arrangement: Arrangement of tumour cells in concentric structures around a vessel.
12. Nuclear quantity: Nuclear concentration in the tumour tissue.
13. Necrosis: Localized destruction of tumour tissue.
14. Haemorrhages: Extravasation of red blood cells in tumour tissue.
15. Calcifications: Deposit of calcified material in tumour tissue.
16. Stroma: Collagenous, reticulin or elastic fibres with or without cells but without relation to vessel walls.
17. Vascularisation quantity: Content of blood vessels in tumour tissue.
18. Anomalous vessels: Any change in the vessel course and/or structure of the vessel wall.
19. Vascular occlusion: Complete or incomplete occlusion of the vessel lumen by embolic or thrombotic material.
20. Tumour invasion of the vessel: Presence of tumour cells in the vascular lumen.
21. Polymorphic cells: Any variation in size or shape of the cell body.
22. Average size of perikaryon: Average size of predominating cell forms. Perikaryon: Part of cytoplasm surrounding the nucleus.
23. Polymorphic nuclei: Any variation in size or shape of nuclei.
24. Average size of nuclei: Average size of predominating form of nuclei.
25. Spherical nuclei: Predominating spherical-shaped nuclei.
26. Oval nuclei: Predominating oval nuclei.
27. Oblong nuclei: Predominating oblong-shaped nuclei.
28. Annular arrangement of nuclei with lumen: Ring-shaped arrangements of nuclei (not ependyma) leaving a central lumen.
29. Annular arrangement of nuclei without a central lumen: Ring-shaped arrangement of nuclei with central cytoplasm.
30. Quantity of chromatin: Nuclear chromatin density.
31. Typical mitosis: Presence of typical mitoses in tumour cells.
32. Atypical mitosis: Presence of atypical mitoses in tumour cells.
33. Glial fibres: Intracellular birefringent fibrillary structures.
34. Degenerative changes without calcification: Excessive accumulation of various material (except calcifications) intra- or extracellular.
35. Lympho-granulo-plasmacellular Infiltration: Peri- or intratumoral accumulation of lymphocytes, leucocytes or plasmacells.
36. Epithelial or tubular arrangement of ependyma: Tubular or epithelial cellular formations identifiable as ependyma.
37. Neuronal quantity: Presence of neurons in tumour tissue.
38. Astrocytic quantity: Presence of differentiated astrocytes in tumour tissue.
39. Astroblastic quantity: Presence of astroblasts in tumour tissue. Usually visible with specific staining methods. Astroblast: Smaller than astrocytes, triangular or oval nuclei with foot-plates.
40. Glioblastic quantity: Presence of glioblasts in tumour tissue. Usually visible with specific staining methods. Glioblast: Mono- or bipolar glial cell with clumsy foot-plates.
41. Honeycomb-like cellular changes: Areas of cells with spherical nuclei and clear perinuclear cytoplasmic halos.
42. Oligodendroglial quantity: Presence of oligodendrocytes in tumour tissue. Usually visible only with specific staining methods.
43. Macrophages and microglial quantity: Presence of macrophages and microglial elements in tumour tissue.
44. Dedifferentiated cells: Presence of unidentifiable cell elements.
45. Not brain-owned cells: Presence of cells without similarity to brain-owned elements.

A personal computer Commodore/8032 with a dual drive floppy disk/8050 was used. The programming language used was Commodore's Microsoft Basic. The tumours were in the 15 more statistically significant groups after WHO classification classified. Table 1 shows the number and distribution of 1500 brain tumours.

Clusters	Diagnosis	Number
1	Meningioma	153
2	Glioblastoma	195
3	Neurinoma	101
4	Neurofibroma	51
5	Astrocytoma	180
6	Oligodendroglioma	166
7	Ependymoma	62
8	Medulloblastoma	89
9	Ca. Metastasis	175
10	Astroblastoma	26
11	Angioma	62
12	Haemangioblastoma	30
13	Pituitary Adenoma	127
14	Unclassified Tumours	25
15	Neuroblastoma	58
		1500

Table 1. Number and distribution of the 1500 tumours of the Central Nervous System in the present study.

Results

Mathematical Fundament

The Bayes' theory of decision is one of the most important statistical instruments for recognition of clusters. The Bayes' theorem says:

$$P(A/B) = \frac{P(B/A) * P(A)}{P(B)} \quad (1)$$

P(A/B) = Probability »a posteriori« is the probability that the elements of B belong to A too.
P(B/A) = Conditional probability or probability that the elements of A belong also to B.
P(A) = Probability that an element belongs to group A.
P(B) = Probability that an element belongs to group B.
In our case tumor X_j will be defined:

$$X_j = (X_{1j}; X_{2j}; \ldots X_{fj} \ldots X_{Fj})$$

X_{fj} = value of each characteristic.

F = 45 characteristics in each tumour.

The name of each class (cluster or tumour diagnosis) will be i.

i = 1, 2, 3, . . . 15.

After (1) then

$$P(W_i/X) = \frac{P(X/W_i) * P(W_i)}{P(X)}$$

$P(W_i/X)$ = Probability »a posteriori« that one tumour with specific values in each characteristic belongs to diagnosis or class i.
$P(X/W_i)$ = Conditional probability that one determinated tumour X_j, belongs to diagnosis i and has in his characteristics the value X.
$P(W_i)$ = Probability »a priori« that a determinated tumor X_j belongs to diagnosis or class i.
P(X) = Probability that a tumor has a characteristic with the value X independent of the diagnosis.
The purpose of our study is to calculate the maximum probability »a posteriori« that a tumour with known values of his

characteristics belongs to one of the 15 clusters (class or diagnosis):

$$P(W_i/X) = \max P(W_k/X)$$

We need to know:

1. The values of the probability »a priori« in each class $P(W_i)$. Probability »a priori« is the probability that the tumour belongs to a specific group. These probabilities are calculated as a mean of probabilities for each group described in the literature 2, 3, 5, 7, 12, 13 (see Table 2).
2. Control of the binary distances of the tumour are a normal distribution.

Each tumour will be defined by the study of 45 microscopic characteristis. If the characteristic is present will be crossed with »yes« or »1«, if absent with »no« or »0«. All characteristics are independent from each other. From these characteristics the binary distances are to be calculated.

We call Ni (control pattern) the number of tumours that belong as the result of a histological study to class or diagnosis i (i = 1, 2, 3, . . . 15). Each diagnosis contains different numbers of tumors because it is very difficult to obtain a sufficient number of infrequent tumours (see table 1). After a primary classification it is necessary to exclude from the Ni (control pattern) the tumours which are not mathematically representative of each diagnosis. This will be made with the study of the »binary distance« because all characteristics are binary.

$$D^2_{gj} = \frac{M_{gj}}{F}$$

D^2_{gj} = »binary distance« between tumour g and tumour j.
M_{gj} = Number of characteristics with identic value in tumour g and tumour j.
F = Number of characteristics.
D^2_{gj} = »binary distance« between tumour g and tumour j.
If the tumour g and tumour j are identical then $D^2_{gj} = 1$. If the tumours are completely divergent then $D^2_{dj} = 0$. To eliminate the »not representative tumours« it is necessary:

a) To obtain the »binary distance« between each tumour and the other tumours in each class or diagnosis.
b) To sum these »binary distances« of each class.
c) To verify that the values of »binary distance« of each class form a normal distribution with similar mean and standard deviation.
d) To calculate the mean and standard deviation of the Ni.
e) To apply the S^2-Test to tumours with D^o than the mean with p = 0.05 and k = 1. Tumours which do not satisfy the last requirement are excluded.

3. The values of conditional probability.

The conditional probability that a tumor (j) of the control pattern belongs to diagnosis i and the characteristic f takes the value 1 is:

$$P(X_{fj}/W_i) = P_f^{(i)X_{fj}}$$

if characteristic f takes the value 0 then

$$P(X_{fj}/W_i) = (1-P_f^{(i)})^{(1-X_{fj})}$$

The conditional probability that a tumour j belongs to diagnosis i is the product of the independent probabilities of his characteristics, which leads to the binomial probability function.

Then

$$\text{Ln } P(X_j/W_i) = \sum_{f=1}^{F} X_{fj} * \text{Ln} \frac{P_f^{(i)}}{1-P_f^{(i)}} + \sum_{f=1}^{F} \text{Ln } (1-P_f^{(i)})$$

$$\text{Ln } P(W_i) + \text{Ln } P(X_j/W_i) = G_i(X)$$

$G_i(X)$ = Discrimination function

$$G_i(X) = \text{Ln } P(W_i) + \sum_{f=1}^{F} X_{fj} * \text{Ln} \frac{P_f^{(i)}}{1-P_f^{(i)}} + \sum_{f=1}^{F} \text{Ln } (1-P_f^{(i)})$$

$$G_i(X) = \text{Ln } P(W_i) + \sum_{f=1}^{F} \text{Ln } (1-P_f^{(i)}) + \sum_{f=1}^{F} X_{fj} * \text{Ln} \frac{P_f^{(i)}}{1-P_f^{(i)}}$$

Then

$$AV_o = \text{Ln } P(W_i) + \sum_{f=1}^{F} \text{Ln } (1-P_f^{(i)})$$

$$AV_f = \text{Ln} \frac{Pf^{(i)}}{1-P_f^{(i)}}$$

AV_o = Influence of class (1, 2, 3, . . . 15)
AV_f = Influence of characteristic (1, 2, 3, . . . 45) (Table 3)

$$\text{Then } G_i(X) = AV_o + \sum_{f=1}^{F} AV_f * X_{fj} \qquad (2)$$

Computer Study

A tumour which belongs to a class with the $G_i(X)$ discrimination function has a maximum value of one of the 15 classes. Their probability »a posteriori« can be calculated with the formula (1). In practice to establish an »a posteriori« probability of a diagnosis, the following steps have to be taken:

1. To calculate the values of the 15 discrimination function in the tumour problem.
2. To order in descending rank.
3. To calculate the 14 differences between the 15 discrimination functions.
4. To calculate the mean of the differences.
5. If the first difference > mean, then the tumour belongs to the first class. If the first difference is < mean, then the tumour belongs to the first and second class, etc.

	Cluster of Diagnosis	Probability »a priori«
1.	Meningioma	0.093
2.	Glioblastoma	0.320
3.	Neurinoma	0.038
4.	Neurofibroma	0.039
5.	Astrocytoma	0.172
6.	Oligodendroglioma	0.031
7.	Ependymoma	0.039
8.	Medulloblastoma	0.058
9.	Ca. Metastasis	0.078
10.	Astroblastoma	0.027
11.	Angioma	0.020
12.	Haemangioblastoma	0.011
13.	Pituitary Adenoma	0.054
14.	Unclassified Tumours	0.008
15.	Neuroblastoma	0.006
		0.994 ≈ 1

Table 2. Cluster or diagnosis of the Brain Tumours with their Probabilities »a priori« calculated as a mean of probabilities according to various Authors [2, 3, 5, 7, 12, 13].

6. Their probability must be calculated after the formula (1):

$$P(W_i/X) = \frac{G_i(X)}{\sum_{i=1}^{c} G_i(X)}$$

The calculation to establish the probability »a priori« and the coefficients of the discrimination functions have been performed also on a personal computer.

After the evaluation of these coefficients of the discrimination function for each class and each characteristic (Table 3), the diagnosis method has been tested on a group of 180 new Tumours of the Central Nervous System. An example of the mathematical classification is summarized in the Table 4. The computer displays and outputs values of diagnosis in descending rank of their probabilities.

In Table 5 the new models are ranked in progressive classes and are separated into two groups. On the left side »Right classified 1« indicates that the first computer diagnosis is identical with the traditional histological diagnosis. On the right side »Right classified 2« indicates that one of the computer diagnosis in each case is identical with the traditional diagnosis. Tumours under »Right classified 1« have a mean diagnosis probability of 72.77 %. Tumours under »Right classified 2« have a probability of 92.70 %.

Discussion

The Bayesian formula which allows the recognition of clusters [1, 9, 10] has been applied to the mathematical diagnosis of Brain Tumours. On the base of the existence of histological characteristics it is possible to identify cluster or in other words diagnostic groups. The cluster have been formed in compliance with the diagnosis of the main groups of the WHO classification. The results of the calculations after Probability Theory of cluster analysis allows to classify mathematically tumours of the Brain with a probability between 72.77 and 92.70 %.

Class 1: Meningioma	
AVO + S AVF (I) * XFJ	
AVO = − 43.0082	(23) = 0.0540675
AVF (1) = − 0.733969	(24) = 3.58351
(2) = − 4.59512	(25) = 1.8563
(3) = − 3.58352	(26) = 4.5951
(4) = − 1.8563	(27) = 0.0540675
(5) = 0.382992	(28) = − 4.59512
(6) = − 2.8622	(29) = − 4.59512
(7) = 0.271934	(30) = 2.8622
(8) = − 1.45529	(31) = − 1.8563
(9) = 1.64223	(32) = − 4.59512
(10) = 3.58351	(33) = − 4.59512
(11) = 0.271934	(34) = − 1.8563
(12) = 2.42775	(35) = − 1.28785
(13) = − 0.382992	(36) = − 4.59512
(14) = 1.45529	(37) = − 4.59512
(15) = 0.613105	(38) = − 4.59512
(16) = 0.860201	(39) = − 4.59512
(17) = 4.5951	(40) = − 4.59512
(18) = 1.64223	(41) = − 4.59512
(19) = − 1.64223	(42) = − 4.59512
(20) = − 1.8563	(43) = − 1.28785
(21) = − 0.0540674	(44) = 0.382992
(22) = − 1.45529	(45) = 0.271934

Table 3. Example of Discrimination function (Df) according to the formula (2). Each tumour group has a Discrimination function with two indices: AVO is the »weight« of each class and AVF (I) is the influence of each characteristic in these classes. XFJ have a value 1 if the characteristic is present.

Tumour 284	Traditional diagnosis:	Meningioma	
	Computer diagnosis: 1.	Meningioma	99.97454 %
Tumour 304	Traditional diagnosis:	Glioblastoma	
	Computer diagnosis: 1.	Glioblastoma	85.88129 %
	2.	Astrocytoma	13.57956 %
	3.	Oligodendroglioma	.26414 %
	4.	Astroblastoma	.19111 %
	5.	Unclassified Tumour	.08368 %
Tumour 352	Traditional diagnosis:	Neurinoma	
	Computer diagnosis: 1.	Meningioma	66.94604 %
	2.	Neurinoma	26.46900 %
	3.	Ependymoma	6.12090 %
	4.	Astrocytoma	.44294 %
Tumour 559	Traditional diagnosis:	Oligodendroglioma	
	Computer diagnosis: 1.	Oligodendroglioma	81.00350 %
	2.	Glioblastoma	15.47739 %
	3.	Ca. Metastasis	3.20030 %
Tumour 724	Traditional diagnosis:	Ca. Metastasis	
	Computer diagnosis: 1.	Ca. Metastasis	89.34260 %
	2.	Haemangio-blastoma	2.27420 %

Table 4. Examples of mathematical classification of Brain Tumours calculated with the computer. The probabilities »a posteriori« were obtained according to the formula (2) and ordered in descending rank.

Class	Number of Tumours	Right classified 1	%	Right classified 2	%
1	13	12	92.30	13	100
2	22	19	86.36	22	100
3	13	6	46.15	13	100
4	8	8	100	8	100
5	16	9	56.25	15	93.75
6	17	16	94.11	16	94.11
7	13	8	61.53	12	92.30
8	15	11	73.33	13	86.66
9	14	13	92.85	14	100
10	3	0	0.00	2	66.66
11	12	12	100	12	100
12	6	4	66.66	6	100
13	14	12	85.71	13	92.85
14	5	0	0.00	1	20.00
15	9	1	11.11	7	77.77
Total	180	131	72.77 %	167	92.77 %

Table 5. Control of the effectivity of the automatic classification of 180 new Brain Tumours. »Right classified 1« indicates that the first computer diagnosis is identical with the traditional pathological diagnosis. »Right classified 2« indicates that one of the computer diagnoses includes the traditional histological diagnosis.

The assessment of the presence or absence of the 45 histological characteristics is very simple and can be performed not only by specialized pathologists but by medical students or paramedical staff as well. All characteristics used are common in standard histological studies. Staining methods of Haematoxilin-Eosin and Thrichroms are generally used in all laboratories of Pathology or Neuropathology.

The histological features are independent of each other. The examiner needs only to state the existence or absence of the respective characteristics on the slides but does not need to quantify the occurrence for example of vessels, mitoses, etc. At present, studies are performed to include morphometrical and quantified histological data into the analysis to intend to obtain better results. Karyometric data analysis have been used to classify Gliomas by Automated Microscopic Picture Analysis [4], but it is a expensive method for every day diagnosis. The present method relies the manual input of the histological data into a computer (30–60 seconds) which then estimates the probability that this tumour with these characteristics belongs to a recognized group, which discrimination function has been evaluated (10–20 seconds).

The present results show that it is possible to distinguish between the main groups of brain tumours from each other by means of objectively obtainable histological characteristics and with a high degree of accurancy. The decision of a particular diagnosis depends on the pathologist, but the computer can show the probability of correct decision. The automatical classification of Brain Tumours can be obtained in a few seconds, requiring only a personal computer.

The frequency of the histological features in a particular tumour group shows only the degree of probability how these characteristics can appear in a diagnosis and can be a relative help to the diagnosis. The discrimination function shows their »importance« in a tumour problem, expressed in mathematical algorithms.

The computer programs in the present study were set up to output only tumours with a probability greater than 0.01. In many cases the computer provides only one diagnosis, but other show different diagnoses in descending rank down to a probability of 0.01. For example: Tumour 724 has a traditional diagnosis of Ca. metastasis (class 9). The computer diagnosis shows the diagnosis of Ca. metastasis with a probability of 0.893426 (9), and Haemangioblastoma with the probability of 0.022742 (12). In our opinion we should decide on Ca. Metastasis (Table 4).

Table 5 shows that the evaluation of correctly classified tumours range between 0 and 100 %. An analysis of the percentages of classification has shown that small percentages are obtained by class 10 (Astroblastom) and class 14 (Unclassified tumours). There are two important causes of incorrectly classified tumours, namely an inappropriate evaluation of the histological characteristics and small quantity of tumours in a class or diagnosis group before the discrimination function can be calculated. In the last case the discrimination function has a small »discriminating effect«. Only 26 Astroblastomas and 25 Unclassified tumours were studied because they are uncommon. With a large number of tumours it will be possible to obtain discrimination function with probabilities > 0.9277 (or 92.77 %). In the present study only 15 main frequent groups of Brain tumours were classified. In the future these methods should be applied to different subgroups of Brain tumours or to other pathological entities.

Acknowledgments

The authors like to thank Prof. S. Obrador-Alcalde (†), Prof. R. Wüllenweber, Prof. E. Kazner, Prof. S. Blümcke and Dr. F. Antoli for their help and Dr. G. Gosztonyi for reviewing the manuscript.

References

[1] ESCUDERO, L. F.: Reconocimiento de Patrones. Centro de Investigacion UAM-IBM, Paraninfo (eds.). Madrid 1973.
[2] HENSCHEN, F.: Tumoren des ZNS und seiner Hüllen. In: LUBARSCH, O., HENKE, F., RÖSSLE, R. (eds.), Handbuch der spez. path. Anatomie u. Histologie, Springer Verlag, Berlin, 1955. Band XIII/3, pp. 413–1040.
[3] JÄNISCH, W., GÜTHERT, H., SCHREIBER, D.: Pathologie der Tumoren des Zentralnervensystems. VEB Gustav Fischer Verlag, Jena, 1976.
[4] MARTIN, H., VOSS, K.: Computerized Classification of Gliomas by Automated Microscope Picture Analysis (AMPA). Acta Neuropathol. (Berl.) **58,** 261–268, 1982.
[5] OBRADOR-ALCALDE, S., SANZ-IBANEZ, J.: Tumores intracraneales. Monografia del Instituto Nacional de Oncologia. Madrid, 1955.
[6] RIO-HORTEGA, PDEL.: Estructura y sistematizacion de los gliomas y paragliomas. Arch. Espan. Oncol. **2,** 411–677, 1932.
[7] RUSSELL, D. S., RUBINSTEIN, L. J.: Pathology of tumours of the nervous system. 4th ed. Edward Arnold Ltd., London, 1977.
[8] SCHERER, H. J.: The form of growth in gliomas and their practical significance. Brain **63,** 1–34, 1940.
[9] SONNENBERG, A.: Sequential iteration of Bayesian Formula by pocket calculator and its use in clinical routine. Comput. Biol. Med. **12,** 357–360, 1982.
[10] STEINHAUSEN, D., LANGER, K.: Clusteranalyse. Einführung in Methoden und Verfahren der automatischen Klassifikation. Verlag Walter de Gruyter, Berlin, 1977.
[11] STOCHDORPH, O.: Die Gewebsbilder der Hirngewächse und ihre Ordnung. Veröff. aus der morph. Pathol. **60,** Gustav Fischer Verlag, Stuttgart, 1955.
[12] ZIMMERMANN, H. M.: Introduction to Tumours of the Central Nervous System. In: MINCKLER, J. (eds.): Pathology of the Nervous System. McGraw-Hill Book Company, New York, 1968, Vol. 2, pp. 1947–1951.
[13] ZÜLCH, K. J.: Biologie und Pathologie der Hirngeschwülste. In: OLIVECRONA, H., TÖNNIS, W. (eds.): Handbuch der Neurochirurgie, Springer Verlag, Berlin, 1956, Bd. 3, pp. 1–702.
[14] ZÜLCH, K. J.: Histological Typing of Tumours of the Central Nervous System. World Health Organization, International Histological Classification of Tumors No. 21, Geneva, 1979.
[15] ZÜLCH, K. J.: Historical Development of the Classification of Brain Tumours and the New Proposal of the World Health Organization (WHO). Neurosurg. Rev. **4,** 123–127, 1981.

Date of receipt: June 23rd, 1983.
The authors' address: José R. Iglesias, M.D., Institut für Neuropathologie, Universitätsklinikum Steglitz, Hindenburgdamm 30, D-1000 Berlin 45.

EDV in Medizin und Biologie 14 (2), 45–49, ISSN 0300-8232

Classification of aquatic bacterial strains: An example of numerical taxonomy in limnology

H.-J. Krambeck and K.-P. Witzel*)

Summary

A Fortran-programsystem, capable of classifying aquatic bacteria strains by means of cluster-analysis is described. In minutes, a VAX 11/750 with a Versatec VM 80, clusters matrices of up to one million Elements and produces graphical output; in addition the hierarchical presentation of the similarities of the distinct clusters in the form of dendrograms is possible.

The predecessors of this present system were useful only in demonstrating the validity of the method because their computing times extended over days; whereas this one, because it is very fast and easy to handle, has the features of a standard method.

Zusammenfassung

Vorgestellt wird ein – in Fortran geschriebenes – Programmsystem, das die Klassifizierung von aquatischen Bakterienstämmen mittels Clusteranalyse erlaubt. Auf einer VAX 11/750 mit einem Versatec VM 80 können im Zeitraum von Minuten Matrizen mit bis zu einer Million Elementen geclustert und graphisch ausgegeben werden; auch ist die hierarchische Darstellung der Ähnlichkeiten der Cluster mit Dendrogrammen möglich.

Die Vorläufer des hier beschriebenen Systems waren wegen der z. T. tagelangen Rechenzeiten nur als Demonstration der Methodik brauchbar, während dieses wegen seiner Schnelligkeit und leichten Bedienbarkeit alle Merkmale einer Standardmethode besitzt.

The problem

The greater number of aquatic aerobic bacterial strains are rather similar in shape; they are mostly more or less elongated, cylindrical or spherical cells. A representative picture of an aquatic bacterial community made by a scanning electron microscope is shown in Fig. 1. Because of their similar shapes, a taxonomic estimation of bacteria is usually based on the results of a series of biochemical investigations. Table 1 shows the series of tests which were applied to the strains presented in this paper.

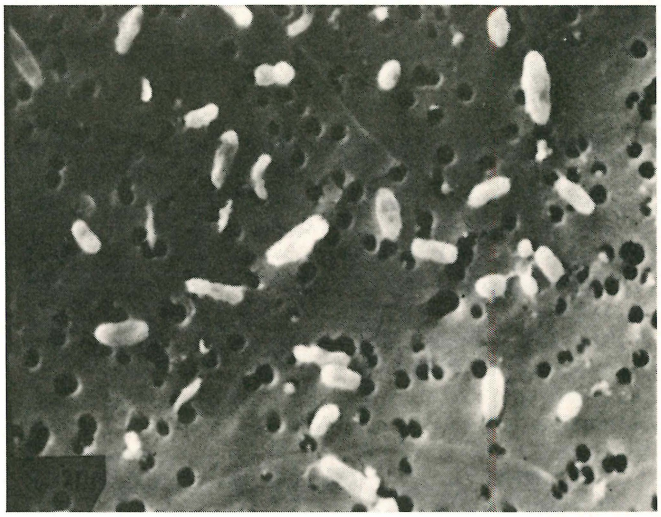


Fig. 1: A photograph of a typical aquatic bacteria population, taken with a scanning electron microscope.

The method

The series of tests of Table 1 produces a vector V_s of properties for each strain s in scope.

$$V_s = \begin{pmatrix} v_{s1} \\ v_{s2} \\ \cdot \\ \cdot \\ \cdot \\ v_{sn} \end{pmatrix}$$ where n = total number of tests and

$$v_{sk} = \begin{cases} 0 & \text{if the result of the test was negative} \\ 1 & \text{if the result of the test was positive} \\ 2 & \text{if the result of the test was not applicable or if the test was not carried out} \end{cases}$$

This last case ($v_{sk} = 2$) allows for the comparison with other series of tests, whose number n was different; in this case the combined set of the valid tests is applied. From the vectors of property V_i and V_j the coefficient of similarity S_{ij} is calculated as follows (SOKAL & MICHENER 1958):

$$S_{ij} = \text{NINT}\left(100.0 \cdot \frac{NS_{ij}}{NS_{ij} + ND_{ij}}\right) \text{ with}$$

*) Herrn Professor Dr. Jürgen Overbeck, unserem verehrten Lehrer, zum 60sten Geburtstag gewidmet.

1.) Kokken
2.) Gerade Stäbchen
3.) Gekrümmte Stäbchen
4.) Fadenförmige Zellen
5.) Zellenden rund
6.) Zellenden verjüngt
7.) Keulenförmige Zellen
8.) Einzelzellen
9.) Zellpaare
10.) Zellketten
11.) Beweglichkeit Mikroskop
12.) Beweglichkeit Agar
13.) Zellänge 0,0–2,6 My
14.) Zellänge 2,6–4,0 My
15.) Zellänge 4,0–UE
16.) Zelldicke 0,0–0,8 My
17.) Zelldicke 0,8–UE
18.) Zelldicke
19.) Zellen unbegeißelt
20.) Zellen polar begeißelt
21.) Sporenbildung
22.) Wachstum in ZOB, Ext. 0,00–0,40
23.) Wachstum in ZOB, Ext. 0,41–0,85
24.) Wachstum in ZOB, Ext. 0,86–1,50
25.) Wachstum in ZOB, Ext. 1,51–UE
26.) Gramfärbung
27.) Hugh-Leifson GLC Aerob
28.) Hugh-Leifson GLC Anaerob
29.) Arginase
30.) Gas aus GLC-Pepton
31.) Säure aus GLC-Pepton
32.) Urease
33.) Katalase
34.) Oxidase
35.) H_2S-Bildung
36.) Indol aus Try
37.) Stärkeabbau
38.) Fak. anaerobes Wachstum
39.) Fluoreszierendes Pigment
40.) Nitratreduktion
41.) Gasbildung im Nitratmedium
42.) Stickstoffbildung
43.) Koloniedurchmesser 0,0–0,4 mm
44.) Koloniedurchmesser 0,4–0,8 mm
45.) Koloniedurchmesser 0,8–1,4 mm
46.) Koloniedurchmesser 1,4-UE
47.) Kolonie rot
48.) Kolonie blau
49.) Kolonie weiß
50.) Kolonie grau
51.) Kolonie gelb
52.) Kolonie durchsichtig
53.) Kolonie durchscheinend
54.) Kolonie ganzrandig
55.) Kolonie welliger Rand
56.) Kolonie gezähnter Rand
57.) Kolonie dünn
58.) Kolonie konvex
59.) Kolonie schleimig
60.) Kolonie ausbreitend
61.) Kolonie lederartig
62.) Kolonie wurzelförmig
63.) Kolonie glatt
64.) Kolonie strukturiert
65.) Kolonie glänzend
66.) Kolonie matt
67.) Kolonie mit konzentr. Ringen
68.) Kolonie mit radialer Struktur
69.) Kolonie mit diff. Pigment
70.) Hugh-Leifson Lac Aerob
71.) Wachstum in Kaz O. C.
72.) Zellen Peritrisch begeißelt
73.) Flocken in Nährlösung
74.) Hugh-Leifson Lac Anaerob
75.) Kolonie rot-orange
76.) Kolonie cream
77.) Kolonie weißgrau
78.) Kolonie gelbcream
79.) Kolonie graucream
80.) Kolonie rosa
81.) Kolonie rund
82.) Gerade + gekr. Stäbchen
83.) Gekr. + fadenf. Stäbchen
84.) Kokken + Stäbchen
85.) 1–2 Zellen
86.) Zellaggregate
87.) 1–2 Zellen + Zellketten
88.) 1–2 Zellen + Ketten + Aggreg.
89.) Zellketten + Aggregate
90.) 1–2 Zellen + Aggregate
91.) NH_3-Bildung aus Pepton
92.) Voges-Proskauer-Reaktion
93.) Methylrotprobe
94.) Citratverwertung
95.) Lackmusmilch, neutral
96.) Lackmusmilch, sauer
97.) Lackmusmilch, basisch
98.) Lackmusmilch, weiß
99.) Lackmusmilch, neutral-koag.
100.) Lackmusmilch, sauer-koag.
101.) Lackmusmilch, basisch-koag.
102.) Lackmusmilch, weiß-koag.
103.) Wachstum auf NB bei 5 C
104.) Wachstum auf NB bei 10 C
105.) Wachstum auf NB bei 20 C
106.) Wachstum auf NB bei 27 C
107.) Wachstum auf NB bei 37 C
108.) Wachstum auf NB bei 45 C
109.) Wachstum auf NB bei pH 4
110.) Wachstum auf NB bei pH 5
111.) Wachstum auf NB bei pH 6
112.) Wachstum auf NB bei pH 7
113.) Wachstum auf NB bei pH 8
114.) Wachstum auf NB bei pH 9
115.) Wachstum auf NB bei 0,0% NaCl
116.) Wachstum auf NB bei 2,5% NaCl
117.) Wachstum auf NB bei 5,0% NaCl
118.) Wachstum auf NB bei 7,5% NaCl
119.) Wachstum auf NB bei 10,0% NaCl
120.) Gelatineverflüssigung
121.) Celluloseabbau
122.) Resistenz gegen Penicill. 0,02%
123.) Resistenz gegen Penicill. 0,2%
124.) Resistenz gegen Polymyxin 0,02%
125.) Resistenz gegen Polymyxin 0,2%
126.) Resistenz gegen Streptom. 0,02%
127.) Resistenz gegen Streptom. 0,2%
128.) Resistenz gegen Tetracyc. 0,02%
129.) Resistenz gegen Tetracyc. 0,2%
130.) Resistenz gegen Chloramph. 0,02%
131.) Resistenz gegen Chloramph. ges.
132.) Resistenz gegen Sulfanils 0,02%
133.) Resistenz gegen Sulfanils ges %

Table 1: The list of the biochemical tests

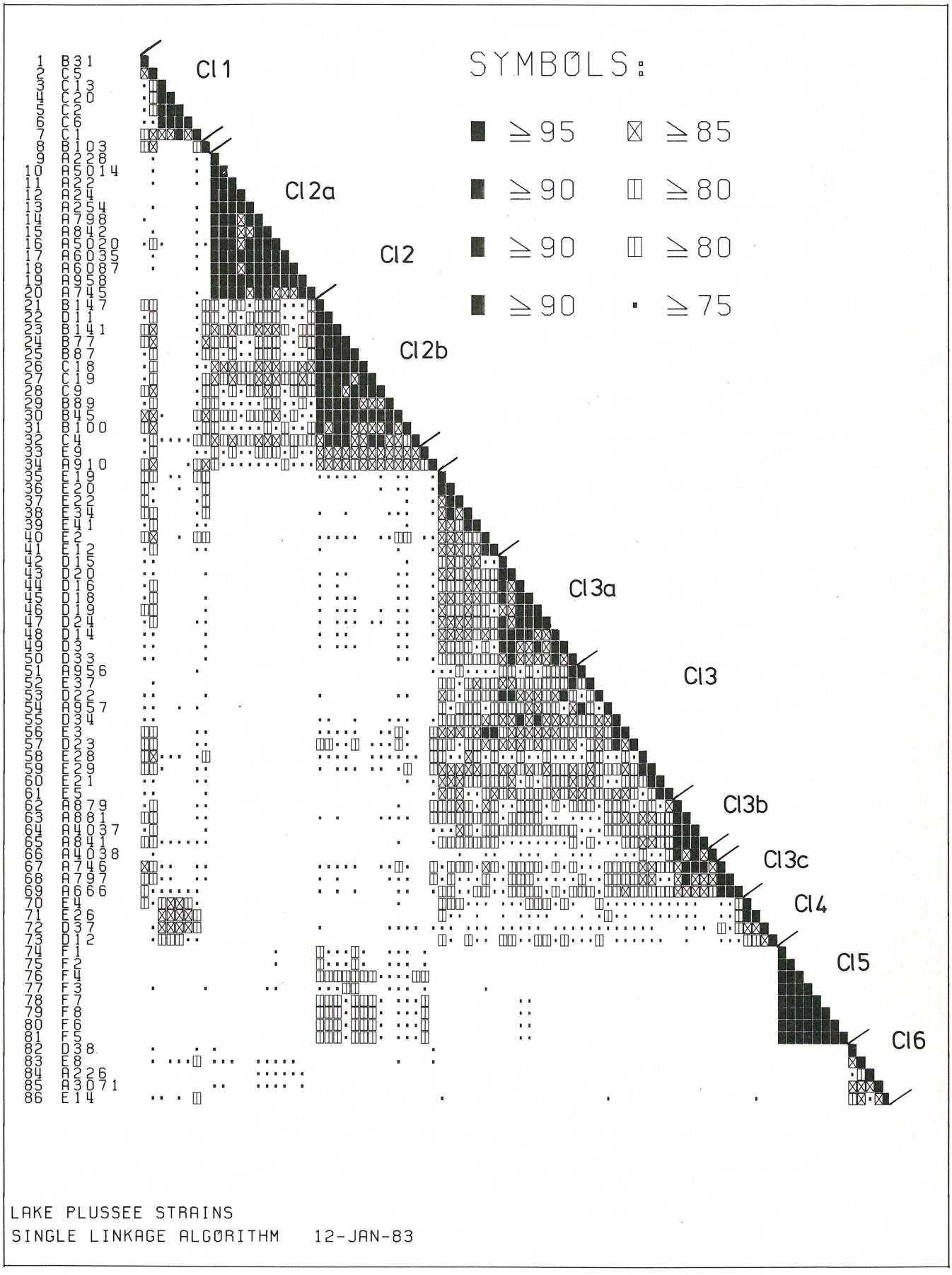


Fig. 2: A cluster matrix from 86 bacterial isolates; the dark triangles below the main diagonal are the groups of similar bacteria.

$$NS_{ij} = \sum_{k=1}^{n} \begin{cases} g_k \text{ if } v_{ik} = v_{jk} \text{ and } v_{ik} \neq 2 \\ O \text{ otherwise} \end{cases}$$

= weighted number of tests, in which strains i and j are in agreement.

$$ND_{ij} = \sum_{k=1}^{n} \begin{cases} g_k \text{ if } v_{ik} \neq v_{jk} \text{ and } v_{ik} \neq 2 \text{ and } v_{jk} \neq 2 \\ O \text{ otherwise} \end{cases}$$

= weighted number of tests, in which strains i and j differ.

NINT = Fortran-Subroutine, which converts real numbers to integer values including rounding.

$$G = \begin{pmatrix} g_1 \\ g_2 \\ \cdot \\ \cdot \\ \cdot \\ g_n \end{pmatrix}$$

is a vector which weights the importance of the tests.

The reason for this vector is the consideration that in bacteriology a property like e.g. »grampositive« is given greater taxonomic significance than e.g. the »thickness of a cell«. Thus, each member of the series of tests can be weighted according to the importance of that test; the significance of S_{ij} can therefore be maximized as a function of those tests. The computer program reads the vector G form the datafile containing the results of the tests; it must be noted however that a comparison of different investigations is only possible, if the same weight vector G is used. Since different opinons exist as how to weight a specific test correctly, in our computer runs the weight factors were set to unity. However the program is flexible enough to easily implement customized weight vectors. The factor 100.0 in the expression NINT, which transforms S_{ij} to the integer range 0–100, has simply technical reasons; because the matrices of similarity (1000 × 1000) can sometimes be quite large, the elements can be declared as bytes instead of reals, which has a significant influence both on computing time and memory requirements.

Calculating the values for every combination of strains results in a triangular matrix of similarity S_{ij} (i = 1, 2, . . ., n; j = 1, 2, . . ., i), which is then processed by a cluster algorithm (Single linkage Clustering, SNEATH & SOKAL 1973), which sorts the matrix according to related groups, i.e. maximum similarity coefficients.

Results

Fig. 2 depicts a clustered matrix with 86 bacterial isolates. The clusters show the different populations (in the dark triangles below the main diagonale) very well; they are annotated as CL 1 to CL 6. In a computer run, in which over 300 bacterial strains from the river Saar and several of its more or less polluted tributaries, as well as strains from a small north German lake were compared, the different isolates could be separated almost completely (K.-P. WITZEL et al. 1981). The program system performed equally well in a comparison of oligocarbophilic and saprophytic strains of bacteria (K.-P. WITZEL et al. 1982b), and in seasonal changes of the bacterial populations of a lake (K.-P. WITZEL et al. 1982a).

In connection with proposed automation of the chemistry of the tests, and the recording of the data by means of marksensecards in the laboratory itself, this previously so laborious classification of aquatic bacteria isolates, has all the features enabling it to become a routine method. Processing the clusterprogram on a 32-bit midicomputer (here: a DEC VAX 11/750) with an electrostatic plotter is very fast, needing about $\frac{1}{100}\ n^2$ (n = number of isolates) CPU-seconds; with 100 strains this takes 2 minutes, with 300 strains about 15 minutes.

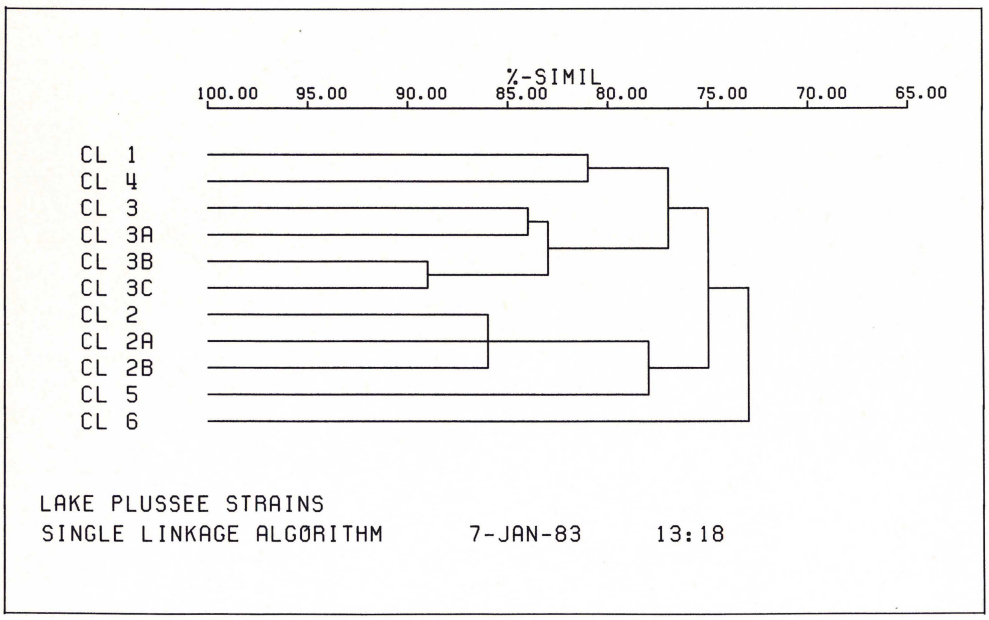


Fig. 3: The dendogram, which depicts the 11 revealed clusters in their mutual relationships and in order of descending similarity coefficients. Remarkable is the data reduction; about 11 300 numbers, are expressed in this method of presentation.

The graphic output comprises not only the clustered matrix, but also the presentation of the hierarchical similarities by means of dendrograms; Fig. 3 depicts the order of the 11 clusters.

Summary and conclusions

The exceedingly large manifold of strains of aquatic bacteria can be depicted very quickly and arranged clearly in the form of cluster-matrices (Fig. 2) and dendrograms (Fig. 3) by the described programme. It's two predecessors (written in Algol 60 on a TR 440 of the University of Saarbrücken and on a PDP 8 of the MPI for Limnology) demonstrated the selectivity and usefulness of the method but due to the lengthy computing times (hours to days), they were too slow for practical use. This programme was translated into Fortran 77 and expanded to include graphic output for cluster matrices and dendrograms; it's typical execution time lasts only minutes and is appropriate as a standard laboratory-method. It can also be used outside the field of bacteriology and is available from the authors.

Acknowledgements

The programming of the dendrogram-graphics was performed by PETER KNOKE, who also translated most of the programme to Fortran. Mrs. SABINE DÜHRKOOP edited the data and typed the manuscript. Dr. CHRISTIANE KRAMBECK provided the photograph of Fig. 1.

References

SNEATH, P. H. A. & R. R. SOKAL, 1973: Numerical taxonomy. The principles and practice of numerical classifications. – Freemann, San Francisco.

SOKAL, R. R. & C. J. MICHENER, 1958: A statistical method for evaluating systematic relationships. – Kansas Univ. Science Bull. **38:** 1409–1438.

WITZEL, K.-P., H.-J. KRAMBECK & H. J. OVERBECK, 1981: On the structure of bacterial communities in lakes and rivers – a comparison with numerical taxonomy of isolates. – Verh. Internat. Verein. Limnol. **21:** 1365–1370.

WITZEL, K.-P., H. J. OVERBECK & K. MOALEDJ, 1982a: Microbial communities in Lake Plußsee – An analysis with numerical taxonomy of isolates. – Arch. Hydrobiol. **94:** 38–52.

WITZEL, K.-P., K. MOALEDJ & H. J. OVERBECK, 1982b: A numerical taxonomic comparison of oligotrophilic and saprophytic bacteria from Lake Plußsee. – Arch. Hydrobiol. **95:** 507–520.

Date of receipt: January 14th, 1983.
The author's address: Dr. H.-J. Krambeck and Dr. K.-P. Witzel, Max-Planck-Institut für Limnologie, Postfach 165, D-2320 Plön/Holstein.

STATISTISCHE VERFAHREN/STATISTICAL METHODS

EDV in Medizin und Biologie **14** (2), 49–53, ISSN 0300-8232

Ansätze zum Suchen nach Strukturen bei Nominaldaten

R. Piepersjohanns

Zusammenfassung

Der Autor schlägt Verfahren vor, die die Suche nach Struktur auf nominalskalierten Variablen erleichtern soll. Eines der vorgestellten Programme arbeitet im Sinne einer Diskriminanzanalyse, die Ergebnisse der beiden anderen werden analog zu einer Clusteranalyse interpretiert. Alle Verfahren basieren auf systematischer Suche.

Summary

The author proposes some methods to facilitate the search for structure in nominal data. Three programs are introduced. One is operating in the sense of discriminant analysis, while the results of the other two are interpreted similar to cluster analysis. The methods are based on systematic search.

Problemstellung

Die heuristische Suche nach Strukturen findet als Ergänzung zum klassischen Hypothesentesten zunehmende Verbreitung. Insbesondere bei wenig fixierten Hypothesen sind spezielle Suchroutinen sicher dem massiven Einsatz statistischer Programmpakete vorzuziehen, wenn dabei auch noch die Signifikanzniveaus interpretiert werden. Es gibt bereits Modelle zur Unterstützung der Suche nach Struktur (z. B. GUHA, aber nicht überall verwendbar, da verschiedene, z. T. nicht immer vorhandene Programmsprachen verwendet werden), aber leider noch nicht in den verbreiteten Programmpaketen wie z. B. SPSS.

Ein Programmsystem (HYPAG, HÄRTNER et al 1980) steht dem Anwender bereits zur schnellen Überprüfung spezieller aussagenlogischer Hypothesen zur Verfügung. Die praktische Erfahrung mit diesem dialogorientierten System zeigte jedoch,

daß viele Benutzer mit der Formulierung sinnvoller Ausgangshypothesen überfordert sind, darüber hinaus sind zum schnellen Arbeiten große Rechenberechtigungen (mittlerer oder großer Rechner, je nach Auslastung) erforderlich. Im folgenden werden daher Programme behandelt, die die Suche nach Strukturen automatisieren, wobei man die Optimalitätskriterien zum Teil variieren kann. Die Programme sind nicht dialoggeeignet, da die systematische Suche sehr zeitaufwendig werden kann.

Es werden drei Programme vorgestellt, welche aus einer nicht zu großen Menge von Variablen eine Teilmenge aussuchen (deren Umfang vorgegeben wird), die in gewissem Sinne die Stichprobe optimal beschreibt. Bei allen Programmen wird aus dieser Teilmenge von Variablen eine Multinomialtafel (Mehrwegetafel, Verallgemeinerung der Vierfeldertafel auf mehr als zwei Gruppierungsvariablen, die auch mehr als zwei Stufen haben dürfen) gebildet, die die Stichprobe in viele disjunkte Klassen teilt.

Das erste Programm (Hypag/S-Disc) arbeitet im Sinne einer Diskriminanzanalyse: Die Klassen (i.e. die Zellen), welche durch die Mehrwegetafel entstanden sind, werden mit einer vorgegebenen Gruppenstruktur verglichen. Für jede Klasse wird nach einem Kriterium entschieden, ob sie die Zuordnung zu einer Gruppe impliziert oder indifferent bleibt. Das Kriterium ist in der Regel das Verhältnis zwischen Treffern und Fehlern: Treffer sind die Angehörigen derjenigen Gruppe, die in dieser Klasse am stärksten vertreten ist, Fehler wären zunächst einmal die Vertreter der anderen Gruppen in dieser Klasse. Wenn das Verhältnis zu ungünstig ist (Vorgabe!), erfolgt keine Zuordnung, ebenso wenn zwei Gruppen in einer Klasse gleich stark vertreten sind.

Das zweite Programm (Hypag/S-Komb) geht davon aus, daß jede Variable der Ausgangsmenge ein „Erfolgskriterium" ist, welches durch ein Unterprogramm des Benutzers erklärt wird. Das Programm sucht Kombinationen von Variablen, auf denen möglichst viele Probanden gleichzeitig »Erfolg« haben, oder wo sie mindestens einmal »Erfolg« haben, d.h. die Variablen können mit »und« und »oder« verknüpft werden.

Das dritte Programm (Hypag/S-Typ) sucht aus allen in Frage kommenden Mehrwegetafeln (es wird wieder die Zahl der beteiligten Variablen vorgeschrieben) aussagekräftige Zellen heraus. Aussagekräftig bedeutet, daß das Produkt aus erreichten Probanden und dessen Komplement zur Gesamtzahl der beteiligten Versuchspersonen maximal wird, dadurch werden triviale Kombinationen in der einen wie der anderen Richtung vermieden.

Alle drei Programme sind in FORTRAN V geschrieben.

Das Diskriminanzprogramm Hypag/S-Disc

In das Diskriminanzprogramm werden Rohdaten eingegeben. Durch die Datenbeschreibung (Variablenzahl v, Format) ist die oben erwähnte Variablenmenge erklärt. Eine Variable ist die Gruppenvariable, analog zu den traditionellen Verfahren zur Diskriminanzanalyse. Ferner wird die Anzahl der Variablen (k) eingegeben, die am „Diskriminanzkonstrukt" beteiligt sein soll sowie die Schichtungskriterien pro Variable; denn für jede Variable werden Auszählungen wie für eine r*k-Feldertafel vorgenommen. Das Programm analysiert nun alle möglichen (k+1)-Wegetafeln, das sind Tafeln, an denen die Gruppenvariable sowie k Variablen aus dem (v−1) großen Pool von Diskriminanzvariablen beteiligt sind. Insgesamt gibt es also (v−1) über k solcher Tafeln.

Für jede dieser Tafeln muß der Datensatz einmal bearbeitet werden, was sehr viel Rechenzeit erfordert. 5000 Mal einen Datensatz von 200 Probanden durchzusehen ist selbst für einen Großrechner schon sehr viel, daraus ergeben sich Beschränkungen für die Größe der Variablenmenge, aus der ausgewählt wird und der Anzahl der Elemente im Diskriminanzkonstrukt. Als Richtschnur zur Einhaltung der Rechenberechtigungen kann die folgende Ungleichung dienen:

$$\binom{v-1}{k} = \frac{(v-1)!}{k! * (v-1-k)!} < 5000$$

Die Zahl rechts ist von der Probandenzahl, den Stufenzahlen der Variablen und den Berechtigungen abhängig.

Nun kann man sich fragen, ob dieser immense Aufwand wirklich erforderlich ist. Dazu ein einfaches Beispiel, eine Tafel aus drei zweistufigen Faktoren A, B und C:

	C=1	
	B=1	B=2
A=1	1	2
A=2	2	1

	C=2	
	B=1	B=2
A=1	2	1
A=2	1	2

Alle zweidimensionalen Randverteilungen haben in jeder Zelle die Zahl drei stehen, man beobachtet also keine Abhängigkeiten. Die offensichtlich vorhandene Beziehung aller drei Faktoren untereinander findet hier in den Tafeln niedrigerer Ordnung keinen Niederschlag. Aus diesem Grunde führen hierarchische Verfahren nicht notwendig zur optimalen Tabelle mit k Faktoren.

Alle untersuchten Mehrwegetafeln werden aufgrund der maximalen Treffermöglichkeit bewertet. Dabei geht man wie folgt vor: in jeder durch eine bestimmte Stufenkombination der gerade betrachteten Variablen erzeugten Zelle gibt es eine weitere Partition nach der Gruppenvariablen (man hat es also eigentlich mit einer (k+1)-Wegetafel zu tun). Je nachdem, in welchem Verhältnis die Gruppen in dieser Zelle zueinander stehen, wird die Entscheidung getroffen, welcher Gruppe die

```
VARIABLEN:
    E4   E6   E7

     1    1    1 ***     2   16   12
     2    1    1 ***     2    7    6
     3    1    1 ***     0    1    4
     1    2    1 ***     1    5   11
     2    2    1 ***     4    6    2
     3    2    1 ***     0    3    1  ++++
     1    3    1 ***     3    5    5
     2    3    1 ***     3    1    1
     3    3    1 ***     1    0    0
     1    1    2 ***     0    0    2
     2    1    2 ***     2    4    4
     3    1    2 ***     7    4    2
     1    2    2 ***     1    8    3
     2    2    2 ***     2    5    0
     3    2    2 ***     9    1    0
     1    3    2 ***     2    3    3
     2    3    2 ***     7    2    3
     3    3    2 ***     4    0    0
     1    1    3 ***     1    0    0
     2    1    3 ***     2    4    0
     3    1    3 ***     9    1    3
     1    2    3 ***     3    4    2
     2    2    3 ***     6    5    0
     3    2    3 ***    14    2    1
     1    3    3 ***     8    3    0
     2    3    3 ***     4    1    0
     3    3    3 ***    14    4    0
```

Abb. 1.

Zelle zugeordnet wird. Zunächst entscheidet man sich für die am stärksten besetzte Gruppe. Alle Probanden in dieser Zelle, die nicht dieser Gruppe angehören, sind damit im Sinne dieser Zuordnungsregel Fehler. Jetzt wird das Verhältnis von Treffern zu Fehlern betrachtet; wenn es den vorgegebenen Wert c nicht übersteigt (Standard: c=1), wird die Entscheidung zurückgenommen, und die Zelle bleibt »unentschieden«.

Abbildung 1 zeigt ein Beispiel für eine solche Multinomialtafel. Die Daten entstammen einer Arbeit von J. ORGASS (Probandenzahl n=271).

Hier wurden k=3 Variablen ausgewählt, deren Namen über der Tafel aufgelistet sind. Links von den Sternchen sind die Stufennummern der unabhängigen Variablen angezeigt, die Reihenfolge korrespondiert zu der über der Tafel. Die angekreuzte Zeile beschreibt also die Zelle mit der dritten Stufe von E4, der zweiten von E6 und der ersten Stufe von E7. Rechts von den Sternchen findet sich die Verteilung der Gruppenvariablen in dieser Zelle: keine Vp in Gruppe eins, drei in Gruppe zwei und eine in Gruppe drei. Diese Zelle wird der Gruppe zwei zugeordnet, wodurch ein Fehler entsteht. Eine besondere Behandlung erfahren Zellen ohne Fehler, wie (3, 3, 1) oder (1, 1, 2): Hier wird bei der Berechnung des Optimalitätskriteriums der gesamten Tafel ein Viertel Fehler angenommen. Dadurch wird verhindert, daß Tafeln mit wenigen richtigen Zuordnungen und insgesamt wenig Fehlern bevorzugt werden; denn das Gesamtkriterium ist wieder das Verhältnis von Treffern und Fehlern – Unentschiedene zählen nicht.

Nach dieser Tafel kann man bereits Regeln für künftiges Verhalten aufstellen. Wenn wie im Beispiel dreistufige Prädiktorvariablen vorliegen, dann gibt es bei drei ausgesuchten Variablen maximal 27 Elemente in der Zuordnungsregel. Hier fallen einige Zellen aus, da nicht entschieden werden kann, welcher Gruppe sie zuzuordnen sind. Insgesamt ergibt sich diese Zuordnungstafel:

ZUORDNUNGSTAFEL, SPALTEN BEDEUTEN WAHRE WERTE

	1	2	3
1	87	24	10
2	11	42	18
3	1	6	17
UNENTSCHIEDEN	12	23	20

In den Spalten stehen die wahren Werte, die linke Spalte enthält also alle Probanden der Gruppe eins. Davon wurden 87 wieder der Gruppe eins zugeordnet, 11 landeten bei Gruppe zwei (Fehler!), und einer wurde fälschlich der Gruppe drei zugeschlagen. Auf 12 Mitglieder der Gruppe eins konnte keine der erstellten Regeln angewandt werden.

Mit drei Variablen aus einem Pool von zehn Variablen ist dies das bestmögliche Ergebnis. In der Praxis kann es aber kaum ausgenutzt werden; denn die daraus abgeleiteten Regeln sind zu kompliziert. Man versucht daher, durch geeignetes Zusammenfassen zu einer einfacheren Struktur zu kommen. Dabei ist allerdings nicht mehr gewährleistet, daß auch dann noch die Auswahl dieser Variablen optimal ist, weswegen vom Programm mehrere Kombinationen von Variablen vorgelegt werden. Der Benutzer kann dann diejenige aussuchen, die eine Struktur beschreibt, welche seinen Hypothesen entspricht.

Zur besseren Übersicht und zur Erleichterung der Weiterverarbeitung wird die Multinomialtafel auch in Form eines (liegenden) Baums ausgedruckt (vgl. Abbildung 2). Dadurch wird eine Ordnung der Faktoren vorgeschlagen.

```
E7      =1       E6       =1       E4
 16/ 44/ 42;102#  4/ 24/ 22; 50#  2/ 16/ 12; 30#
                                   2/  7/  6; 15#
                                    /  1/  4;  5#

                         =2       E4
                  5/ 14/ 14; 33#  1/  5/ 11; 17#
                                  4/  6/  2; 12#
                                   /  3/  1;  4#

                         =3       E4
                  7/  6/  6; 19#  3/  5/  5; 13#
                                  3/  1/  1;  5#
                                  1/  /   ;  1#

        =2       E4       =1       E6
 34/ 27/ 17; 78#  3/ 11/  8; 22#   /   /  2;  2#
                                  1/  8/  3; 12#
                                  2/  3/  3;  8#

                         =2       E6
                 11/ 11/  7; 29#  2/  4/  4; 10#
                                  2/  5/   ;  7#
                                  7/  2/  3; 12#

                         =3       E6
                 20/  5/  2; 27#  7/  4/  2; 13#
                 ==============   9/  1/   ; 10#
                                  4/  /   ;  4#

        =3       E4       =1       E6
 61/ 24/  6; 91# 12/  7/  2; 21#  1/  /   ;  1#
                                  3/  4/  2;  9#
                                  8/  3/   ; 11#

                         =2       E6
                 12/ 10/   ; 22#  2/  4/   ;  6#
                                  6/  5/   ; 11#
                                  4/  1/   ;  5#

                         =3       E6
                 37/  7/  4; 48#  9/  1/  3; 13#
                                 14/  2/  1; 17#
                                 14/  4/   ; 18#
```

Abb. 2.

Die Erstellung der Reihenfolge geschieht durch die folgende Prozedur: zuerst wird die Variable der Kombination (hier: E4, E6, E7) gesucht, welche mit der Gruppenvariable den höchsten Zusammenhang ergibt. Dazu werden die r*k-Feldertafeln der drei unabhängigen Variablen mit dem Kriterium gebildet. Diejenige, welche den größten Chiquadratwert ergibt, wird ausgewählt. Das ist hier E7. Unter der Bedingung E7 = 1 werden nun die verbleibenden möglichen zweidimensionalen Kontingenztafeln untersucht, und es wurde hier E6 ausgewählt. Für E7 = 2 kehrt sich die Reihenfolge um. In jedem Ast des Baumes wird also unabhängig entschieden, nach welchem Kriterium weiter zu verzweigen ist. Auf jeder Stufe sieht man eine Auszählung der Gruppenzugehörigkeit im gerade aktuellen Ast, im unterstrichenen Beispiel ist E7 = 2 und E4 = 3. 27 Probanden sind in diesem Zweig, 20 in Gruppe eins, fünf in Gruppe zwei, und die letzten beiden sind in Gruppe drei. Wie man sofort erkennt, ist der Baum schief, was aber einfacher in der Programmierung ist, andererseits kann man hieraus leicht per Hand eine symmetrische Zeichnung erstellen. Aus diesem Baum kann man z. B. Schlüsse der folgenden Art ziehen:

- wenn E7 gleich drei ist, kommt Gruppe drei praktisch nicht mehr in Frage.
- wenn zusätzlich E4 gleich drei ist, ist mit einigen Fehlern Gruppe eins angezeigt.

Wenn die Trennung der Gruppen nach diesem Muster nicht ausreicht, muß überlegt werden, ob man nicht mehr Variablen aussuchen läßt, so daß die Aussagen differenzierter werden. Eine andere, in diesem Programm nicht realisierte Möglich-

```
DIE 10 BESTEN KOMBINATIONEN VON  3 AUS 10 VARIABLEN:

JOSEF ORGASS: EINSTELLUNGSTEST FUER LEHRLINGE

KOMBINATION  1:   48 TREFFER.
           VARIABLEN:  5  6  8
                 KV1:      36       10        2

KOMBINATION  2:   47 TREFFER.
           VARIABLEN:  2  5  6
                 KV1:      35        9        3

KOMBINATION  3:   45 TREFFER.
           VARIABLEN:  5  6 11
                 KV1:      33        9        3

KOMBINATION  4:   44 TREFFER.
           VARIABLEN:  6 10 11
                 KV1:      28       12        4

  usw.

BISHER ERKLAERT:   0 FAELLE. NEU:  48 FAELLE
DURCH DIE VARIABLEN   5,  6,  8,

BISHER ERKLAERT:  48 FAELLE. NEU:  23 FAELLE
DURCH DIE VARIABLEN   2,  6, 11,

BISHER ERKLAERT:  71 FAELLE. NEU:  18 FAELLE
DURCH DIE VARIABLEN   2,  4,  8,

BISHER ERKLAERT:  89 FAELLE. NEU:   8 FAELLE
DURCH DIE VARIABLEN   6, 10, 11,

usw.
```

Abb. 3.

Als Dichotomisierungskriterium wurde für dieses Beispiel für alle Tests 75 als Grenze gewählt.

Unter »Kombination 1« sind 48 Treffer mit den Variablen 5, 6 und 8 genannt, das bedeutet, daß 48 Probanden auf den drei Tests, die diesen Variablennummern entsprechen, hohe Werte (über 75) erzielten. Neben KV1 ist die Variable »Berufsgruppe« innerhalb der Zelle ausgezählt, von den 48 Testpersonen in dieser Zelle gehören also 36 der Berufsgruppe 1 an usw.

Weiter unten wird als zusätzliche Information ermittelt, wie viele Versuchspersonen mit den besten Zellen zusammen erreicht werden.

Beim Programm Hypag/S-Typ ist die oben erwähnte Vorauswahl nicht möglich, dafür kann dieses Programm mehrstufige Variablen verarbeiten. Es werden allgemein Zellen gesucht, die stark besetzt sind. Allerdings kann eine zu starke Besetzung auch wieder eine triviale Aussage bedeuten, dagegen schützt man sich, indem man die Zellen nach folgendem Kriterium wählt:

$$c = t^*(n - t) \overset{!}{=} \max$$

darin ist t die Trefferzahl der Zelle, die beurteilt wird, und n ist der Stichprobenumfang. Damit ist c einerseits proportional zur Zellenbesetzung, anders ausgedrückt zum Abstand zur leeren Zelle, andererseits ist c auch proportional zum Abstand zur »vollbesetzten Zelle«, welche einem Ergebnis entspräche, bei dem alle Beobachtungseinheiten in einer einzigen Zelle liegen. c erreicht sein Maximum, wenn in einer Zelle genau die Hälfte aller Objekte liegen.

keit ist, aus dem gesamten Variablenpool einen Baum nach diesem Prinzip zu erstellen, wobei man eine geeignete Stopregel definieren muß.

Die Programme Hypag/S-Komb und Hypag/S-Typ

Beide Programme suchen stark besetzte Zellen aus Mehrwegetafeln vorgegebener Dimensionalität, das Programm Hypag/S-Komb ist allerdings stark spezialisiert: es sucht nur nach Erfolgen, wobei Erfolg durch ein spezielles Unterprogramm definiert wird. Alle Variablen werden damit dichotomisiert. Die Suche nach den besten Variablenkombinationen kann auf zwei verschiedene Arten geschehen: entweder man sucht solche, die auf jeder der ausgesuchten Variable »Erfolg« haben, oder man fordert nur, daß auf mindestens einer der ausgesuchten Variablen »Erfolg« ist. Wie in allen hier vorgestellten Programmen wird eine gewisse Grundmenge von Variablen systematisch durchsucht, was dazu führt, daß der gesamte Datensatz v über k mal zu durchsuchen ist (v ist der Umfang des Variablenpools, k die Zahl der ausgesuchten Variablen). Da v über k sehr groß werden kann, muß bei großen Variablensätzen eine Vorauswahl getroffen werden. Dazu besteht hier die Möglichkeit, indem die Variablen nach eindimensionaler Trefferquote geordnet werden.

Der Ausdruck umfaßt die besten Zellen und deren Trefferquoten sowie eine Ordnung der ausgewählten Zellen nach folgendem Muster: an erster Stelle erscheint die Zelle mit der höchsten Trefferquote, danach kommt die Zelle, die von den bisher nicht erfaßten Probanden die meisten erklärt usw.

Zusätzlich können innerhalb der ausgesuchten Zellen maximal vier Variablen ausgezählt werden.

Die folgende Abbildung zeigt ein Beispiel, welches wiederum mit den Daten von J. ORGASS gerechnet wurde.

Abb. 4.

```
DIE 10 BESTEN ZELLEN AUS  3-WEGE-TAFELN

JOSEF ORGASS: EINSTELLUNGSTEST FUER LEHRLINGE

ZELLE  1:
E4       =  1
E5       =  1
E7       =  1
KRITERIUM:  9618, TREFFER:  42, DAVON NEU:  42

ZELLE  2:
E1       =  1
E5       =  1
E7       =  1
KRITERIUM:  9430, TREFFER:  41, DAVON NEU:   9

ZELLE  3:
E3       =  1
E5       =  1
E7       =  1
KRITERIUM:  9048, TREFFER:  39, DAVON NEU:   5

ZELLE  4:
E3       =  1
E4       =  1
E7       =  1
KRITERIUM:  8658, TREFFER:  37, DAVON NEU:   9

usw.

UEBERSCHNEIDUNGEN DER AUSGEWAEHLTEN ZELLEN:
          1    2    3    4    5    6    7    8    9
     2   32
     3   28   26
     4   28   20   28
     5   32   32   20   21
     6   32   32   20   20   32
     7   25   21   18   22   24   21
     8   19   25   19   15   19   21   15
     9   25   21   18   18   21   24   25   16
    10   21   21   15   16   24   24   24   16   24
```

Die Zellen werden geordnet nach dem Maß c ausgegeben, darüber hinaus wird der zusätzliche Beitrag der Zelle zu den vorigen gedruckt. Außerdem wird eine Tabelle mit paarweisen Überschneidungen der Zellen aufgelistet, es erscheint dort jeweils die Zahl der Probanden, die sowohl der einen als auch der anderen Zellen angehören.

Als Beispiel dient wieder der bekannte Datensatz. In der Abbildung 4 sind zunächst die besten Zellen beschrieben, diesmal mit genauen Stufenzahlen, da dieses Programm im Gegensatz zum vorigen mehrstufige Variablen zuläßt.

Darunter ist in der Tabelle für jedes Paar ausgesuchter Zellen die Zahl der Probanden angegeben, die beiden Zellen angehören.

Dieses Verfahren ist dann interessant, wenn in einer Stichprobe mehrere Teilgruppen (Cluster) vermutet werden, die möglicherweise überlappend sind. Dann sollten die Überschneidungen zwischen Zellen, die verschiedene Gruppen beschreiben, sehr gering sein.

Literatur

HÄRTNER, R., K. MATTES, & H. WOTTAWA, Computerunterstützte Hypothesenagglutination zur Erfassung komplexer Zusammenhänge. EDV in Medizin und Biologie, 1980, **11** (2), 53–59.

HAJEK, P. & T. HAVRANEK, The GUHA Method – its Aims and Techniques. International Journal of Man-Machine Studies, 1978, **10**, 3–22.

MATTES, K. Hypothesenagglutionation. Ein heuristisches Verfahren zur Datenanalyse. Version 0. Heidelberg: Unveröffentlichte Programmbeschreibung, 1980.

ORGASS, J., Evaluation eines Diagnosesystems zur Auswahl von Mitarbeitern am Beispiel der Eignungsuntersuchung von Ausbildungsbewerbern. Bochum: Unveröffentliche Diplomarbeit, 1982.

Eingegangen am 22. Oktober 1982.
Anschrift des Verfassers: R. Piepersjohanns, Ruhr-Universität Bochum, Psychologisches Institut, Postfach 10 21 48, D-4630 Bochum 1.

EDV in Medizin und Biologie **14** (2), 53–57, ISSN 0300-8232

Analysis of a limiting dilution assay by using the Single-Hit Poisson Model – an APL-Computer programme

G. Kynast and E. Weber*)

Summary

The mathematical-statistical basis for Limiting Dilution analyses of activated T cells is developed. With the Single-Hit Poisson Model one can determine the frequencies of the effective cells. The relative cell frequency Φ is estimated by the Maximum-Likelihood- (ML-) and the Minimum-Chi-square-(MC-) method. An APL-computer programme is presented for both estimation procedures.

Zusammenfassung

Die mathematisch-statistische Basis für eine Limiting-Dilution-Analyse aktivierter T-Zellen wird dargelegt. Das Single-Hit Poisson Modell liefert eine Möglichkeit, die Häufigkeiten effektiver Zellen zu bestimmen. Diese relative Häufigkeit Φ wird mit Hilfe der Maximum-Likelihood- (ML-) und der Minimum-Chi-Quadrat- (MC-) Methode geschätzt.

Für beide Schätzmethoden wurden APL-Computerprogramme entwickelt.

*) Dedicated to Prof. Dr. Otto Westphal on the occasion of his seventieth birthday.

1. Introduction

»Limiting Dilution Assays in cell biology (and cellular immunology) form a special subclass of all transfer assays in which the particular objective is to measure the frequency of the cells giving rise to a response« (MILLER [1982], p. 219). In the following article we describe a statistical model for analysing data from a Limiting Dilution experiment. The estimation of the frequency of immunocompetent cells in mice (cytotoxic T lymphocyte precursors [CTL-P]) is estimated using the Maximum-Likelihood and the Minimum-Chi-square method, according to TASWELL [1981]. Furthermore an aspect pointed out by MILLER [1982] is taken into account which is important for a suitable experimental design.

2. The Single-Hit Poisson Model

A Limiting Dilution Assay is a quantal response bioassay that analyses an all-or-nothing immune response (FINNEY [1978], OLSCHEWSKI et al. [1979]). The reason for using an indirect assay is the fact that the frequency of CTL-P cannot be measured directly. Therefore the activity of the progeny of

CTL-P is measured on ^{51}Cr-labelled target cells and a response is defined as release of ^{51}Cr. This outcome is a 0-1-variable where 1 stands for positive (immune) and 0 for negative response.

In the sequel, the following assumptions are made. There are x_i, $i = 1,\ldots,I$, responder cells, each on n_i cultures, with $\sum_{i=1}^{I} n_i = N$, such that $x_1 < x_2 < \ldots < x_I$. (Because of the inaccuracy of cell dilutions the dose x_i is a mean value.) After addition of the target cells for every group i, $i=1,\ldots,I$, the number of positively reacting cultures l_i will be observed, i.e. their ^{51}Cr-release is greater than 3 standard deviations above the mean background level. (The 3σ limit is usually taken by this kind of immunological experiments.) In such an assay a positive response is obtained from the progeny of a single CTL-P (single-hit). Now the model considers the number of negative or nonreacting cultures $r_i = n_i - l_i$.

From the assumptions made above it follows that the number of negative responses r_i, $i=1,\ldots,I$, are Binomial distributed, $\mathcal{B}(n_i,P_i)$, with parameters n_i and P_i, where P_i is the probability for a negative response. The probability P_i depends on the number of responder cells in the cultures: $P_i = F(x_i)$, where F is a fixed, but unknown distribution function.

Denoting by k_i, $i=1,\ldots,I$, the number of CTL-P's per dose and per culture which cannot be directly observed and by Φ the probability that a single cell is a CTL-P cell, then k_i is Binomial distributed with parameters x_i and Φ, denoted $L(k_i)=\mathcal{B}(x_i,\Phi)$. Since the number of experimental units x_i of this Binomial distribution is large and cannot be defined and the probability of occurence of one event is small, the Binomial distribution will be approximated by the Poisson distribution $\mathcal{P}$ with the parameter $\lambda_i := \Phi x_i$ (see for example LINDER [1964], p. 352), i.e.

$$\lim_{\substack{x_i \to \infty \\ \Phi \to 0 \\ \Phi x_i = \text{const.}}} \mathcal{B}(x_i,\Phi) = \mathcal{P}(\lambda_i).$$

The probability that a culture reacts negatively is in this case the zero term of the above defined Poisson distribution, i.e.

$$P_i = P(k_i=0) = e^{-\lambda_i} = e^{-\Phi x_i} \quad (i=1,\ldots,I).$$

This expression is the mathematical formulation of the so-called Single-Hit Poisson Model.

A χ^2-test is used for testing the goodness of fit of this model to real data (see for example ARMITAGE [1971], p. 391).

3. Estimation of Φ by linear regression

According to the described model the unknown parameter Φ must be estimated from experimental data, consisting of the number of negatively reacting cultures r_i, $i=1,\ldots,I$. The r_i are $\mathcal{B}(n_i, P_i)$-distributed variables where $P_i = e^{-\Phi x_i}$ depends on the parameter Φ and the value of x_i.

A suitable transformation for achieving a linear relation between dose and effect will be offered as the natural logarithm of the observed data, which leads to the following regression equation:

$$\ln P_i = -\Phi x_i \qquad (i=1,\ldots,I) .$$

Two possibilities to estimate the parameter Φ will be discussed: the Maximum-Likelihood- and the Minimum-Chi-square-method. In addition, a rough graphic method will be mentioned.

Estimation methods

The distribution of the r_i was shown to be

$$L(r_i) = \mathcal{B}(n_i, P_i), \text{ i.e.}$$

$$P(r_i) = \binom{n_i}{r_i} P_i^{r_i} (1-P_i)^{n_i - r_i}$$

with $P_i = e^{-\Phi x_i} \qquad (i=1,\ldots,I)$.

a) ML-estimator $\hat{\Phi}_{ML}$

To obtain an ML-estimator the following expression respectively its logarithm must be maximized as a function of Φ:

$$L = \prod_{i=1}^{I} P(r_i)$$

$$\ln L = \sum_{i=1}^{I} \ln P(r_i)$$

$$= \sum_{i=1}^{I} \left[\ln \binom{n_i}{r_i} + r_i \ln P_i + (n_i - r_i) \ln (1-P_i) \right]$$

$$= \sum_{i=1}^{I} \left[\ln \binom{n_i}{r_i} - r_i \Phi x_i + (n_i - r_i) \ln (1-e^{-\Phi x_i}) \right] .$$

Since the maximum of this function cannot be obtained directly $\hat{\Phi}_{ML}$ will be obtained by the well-known conditions:

$$\left. \frac{d \ln L}{d \Phi} \right|_{\Phi = \hat{\Phi}_{ML}} = 0 \quad \text{and}$$

$$\left. \frac{d^2 \ln L}{d \Phi^2} \right|_{\Phi = \hat{\Phi}_{ML}} < 0 .$$

The first and the second derivative of ln L are described in the appendix.

The estimator $\hat{\Phi}_{ML}$ is then calculated by Newton-Raphson's method of iterative approximation:

$$\hat{\Phi}_{j+1} = \hat{\Phi}_j - \left. \frac{d \ln L / d\Phi}{d^2 \ln L / d\Phi^2} \right|_{\Phi = \hat{\Phi}_j} \qquad j = 0,1,\ldots..$$

As an initial value the least-square estimator, obtained from the linear regression through the origin, is used. The iteration is stopped when the difference between $\hat{\Phi}_{j+1}$ and $\hat{\Phi}_j$ for some j is smaller than some $\varepsilon > 0$, f.e. $\varepsilon = 10^{-16}$ in the APL-programme used here. The numerical stability can also be seen by the value of the first derivative at $\hat{\Phi}_{ML}$ which has to be close to zero.

The 95%-confidence limits for $\hat{\Phi}_{ML}$ are calculated as follows:

$$\hat{\Phi}_{ML} \pm 1{,}96 \sqrt{\widehat{\text{var}}\, \hat{\Phi}_{ML}} ,$$

with $$\widehat{\text{var}}\, \hat{\Phi}_{ML} = \left. \frac{-1}{d^2 \ln L / d\Phi^2} \right|_{\Phi = \hat{\Phi}_{ML}}$$

(c.p. TASWELL [1981], p. 1615)

b) MC-estimator $\hat{\Phi}_{MC}$

From the central limit theorem it can be seen that for fixed i and $n_i \to \infty$

$$L\left(\frac{r_i - n_iP_i}{\sqrt{n_iP_i\,(1-P_i)}}\right) \longrightarrow \mathbb{N}\,(0,1)$$

and furthermore

$$\chi^2 := \sum_{i=1}^{I} \frac{(r_i - n_iP_i)^2}{n_iP_i\,(1-P_i)} \longrightarrow \chi^2_{I-1}\,,$$

where $\mathbb{N}$ denotes the Normal and χ^2 denotes the Chi-square distribution with I-1 degrees of freedom.

Thus it follows that

$$\chi^2 = \sum_{i=1}^{I} \left[\frac{(r_i - n_iP_i)^2}{n_iP_i\,(1-P_i)}\right]$$

$$= \sum_{i=1}^{I} \left[\frac{(r_i - n_ie^{-\Phi x_i})^2}{n_ie^{-\Phi x_i}\,(1-e^{-\Phi x_i})}\right]$$

and the Minimum-Chi-square-method demands to minimize this expression. The estimator $\hat{\Phi}_{MC}$ will be the solution of the equation

$$\frac{d\chi^2}{d\Phi}\Big|_{\Phi = \hat{\Phi}_{MC}} = 0 \qquad \text{under the constraint}$$

$$\frac{d^2\chi^2}{d\Phi^2}\Big|_{\Phi = \hat{\Phi}_{MC}} > 0\,.$$

Again, this expression can be solved by using the Newton-Raphson's method of iterative approximation:

$$\hat{\Phi}_{j+1} = \hat{\Phi}_j - \frac{d\chi^2 / d\Phi}{d^2\chi^2 / d\hat{\Phi}^2}\Big|_{\Phi = \hat{\Phi}_j} \qquad j = 0,1,\ldots$$

The initial value and the stopping criterion are the same as in the ML-method. The calculation of the first and the second derivative of χ^2 is described in the appendix, too.

The 95% confidence limits are calculated using the Normal approximation. This yields:

$$\hat{\Phi}_{MC} \pm 1{,}96\,\sqrt{\widehat{var}\;\hat{\Phi}_{MC}}$$

with

$$\widehat{var}\;\hat{\Phi}_{MC} = \frac{2}{d^2 \ln \chi^2 / d\Phi^2}\Big|_{\Phi = \hat{\Phi}_{MC}}$$

(c.p. TASWELL [1981], p. 1615)

c) Rough graphical method for the estimation of Φ

In Figure 1 the Single-Hit Poisson Model equation $P_i = e^{-\Phi x_i}$ is plotted semilogarithmically. An estimation of the frequency Φ can be directly obtained by the slope of the fit line by eye. Asymptotically it is valid that $L(k_i) = \mathbb{P}(\lambda_i)$; for the simplest case $\lambda_i = 1$ regarded here, i.e. in the mean a single CTL-P will be expected, the following relation holds:

$$\lambda_i = 1 \approx x_i\hat{\Phi} \quad <=> \quad \hat{\Phi} \approx \frac{1}{x_i}\,.$$

Because the probability of a negative response is 0.37 in this case, $P(k_i=0) = 0.37$ when $L(k_i)=\mathbb{P}(1)$ holds, one expects in every 37% of negatively reacting cultures in the mean 1 CTL-P. In practice the experimentator takes this method to get a quick estimate of x_i and therefore an appropriate design of the experiment. Figure 1 shows what happens:

(i) Fit line by eye
(ii) Read off x_i to give $\ln P_i = -1$
(iii) Estimate Φ by $1/x_i$

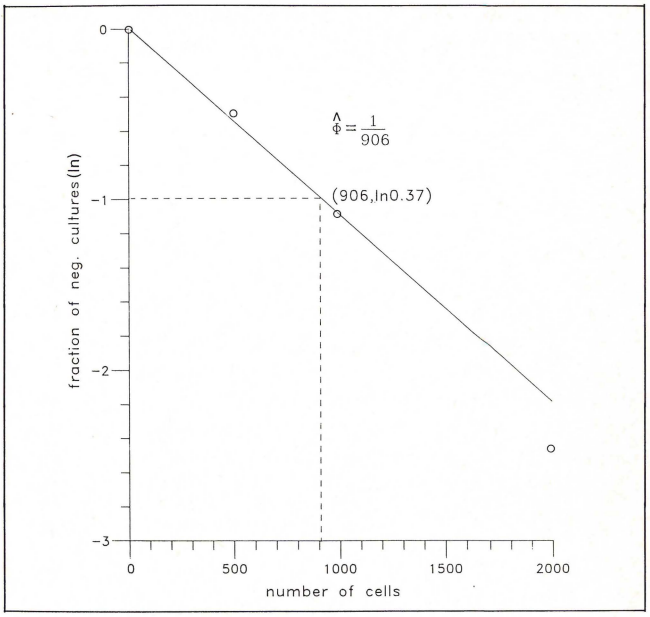


Fig. 1. Rough graphical method for the estimation of Φ.

4. Results

For the experimental performance MILLER ([1982], p. 224) points out the following aspects:

(a) »The culture conditions must be such that a single CTL-P can survive and produce normal progeny after activation.«
(b) »The assay must clearly distinguish cultures containing a single clone of CTL from nonresponding cultures.«
(c) »The presence or absence of CTL-P must be the only factor determining whether a culture will respond.«

The following Table 1 presents a fictious LD experiment. Four cell doses were tested, 24 replicates (=n_i) per cell concentration x_i. The CTL-P-frequency represents fictious cytotoxicity values on ^{51}Cr-labelled target cells.

Table 1 is made by the APL-programme LIDIA, see Appendix.

To the clonal probability in Table 1:

Because k_i, i=1,...,I, the number of CTL-P per dose and per culture, is Poisson distributed with parameter $\lambda_i = \Phi x_i$ the probability of a positive response of a culture – caused by exactly one CTL-P – is $Q_i = P(k_i=1) = \Phi x_i\, e^{-\Phi x_i}$. Remember now the probability of no response – the mathematical formulation of the Single-Hit Poisson Model – : $P_i = P(k_i=0) = e^{-\Phi x_i}$ then $\frac{Q_i}{1-P_i}$ (i=1,...,I) is the conditional probability that a particular responding culture is clonal (=monoclonal). The following graph shows the probability that a particular responding culture is clonal in dependance of the fraction of the observed positive responding cultures $1-P_i$ (see Figure 2).

That means for the Single-Hit Poisson Model: if 1/10 of a group of identical cultures responds positive (i.e. $1-P_i=0.1$) one can calculate

$$P_i = e^{-\Phi x_i} <=> 1 - P_i = 1 - e^{-\Phi x_i},$$

then $\ln P_i = -\Phi x_i <=> \ln (1-P_i) = \Phi x_i$.

For $P_i = 0.9$ $\quad \ln 0.9 = -\Phi x_i$

and $Q_i = \Phi x_i e^{-\Phi x_i} = \Phi x_i P_i = -\ln 0.9\; P_i = -\ln 0.9 \cdot 0.9 \approx 0.095$

finally one gets the conditional probability $\frac{Q_i}{1-P_i} \approx 0.95$.

EXPERIMENT: TEST EXAMPLE, LDA; DATASET: TEST1
MINIMUM CHI SQUARE ESTIMATOR

ESTIMATED PHI: .001099231; STANDARDDEVIATION: .000184753

.95 CONFIDENCE LIMITS: .000737115 AND .001461346

ESTIM. 1/PHI WITH CONF.-LIMITS: 910 684 1357

P-VALUE OF CHI-SQUARE FIT: .93724

VALUE OF THE FIRST DERIVATION: .000000000001346
NUMBER OF ITERATIONS: 5

DOSE	CULTURES	CELLS PER CULTURE	NEG. CULTURES OBSERVED	NEG. CULTURES EXPECTED	FRACTION OF NEGATIVE CULTURES OBSERVED	FRACTION OF NEGATIVE CULTURES EXPECTED	.95 CONF.LIMITS		CLONAL-PROBAB.
1	24	8000	0	.00	.000	.000	.000	.003	.001
2	24	2000	2	2.66	.083	.111	.054	.229	.274
3	24	1000	8	8.00	.333	.333	.232	.478	.549
4	24	500	15	13.85	.625	.577	.482	.692	.750

THE HYPOTHESIS OF THE POISSON REGRESSION CANNOT BE REJECTED.

EXPERIMENT: TEST EXAMPLE, LDA; DATASET: TEST1
MAXIMUM LIKLIHOOD ESTIMATOR

ESTIMATED PHI: .001104179; STANDARDDEVIATION: .000178195

.95 CONFIDENCE LIMITS: .000754917 AND .001453440

ESTIM. 1/PHI WITH CONF.-LIMITS: 906 688 1325

P-VALUE OF CHI-SQUARE FIT: .93709

VALUE OF THE FIRST DERIVATION: .000000000008164
NUMBER OF ITERATIONS: 5

DOSE	CULTURES	CELLS PER CULTURE	NEG. CULTURES OBSERVED	NEG. CULTURES EXPECTED	FRACTION OF NEGATIVE CULTURES OBSERVED	FRACTION OF NEGATIVE CULTURES EXPECTED	.95 CONF.LIMITS		CLONAL-PROBAB.
1	24	8000	0	.00	.000	.000	.000	.002	.001
2	24	2000	2	2.64	.083	.110	.055	.221	.273
3	24	1000	8	7.96	.333	.331	.234	.470	.548
4	24	500	15	13.82	.625	.576	.483	.686	.749

THE HYPOTHESIS OF THE POISSON REGRESSION CANNOT BE REJECTED.

Table 1.

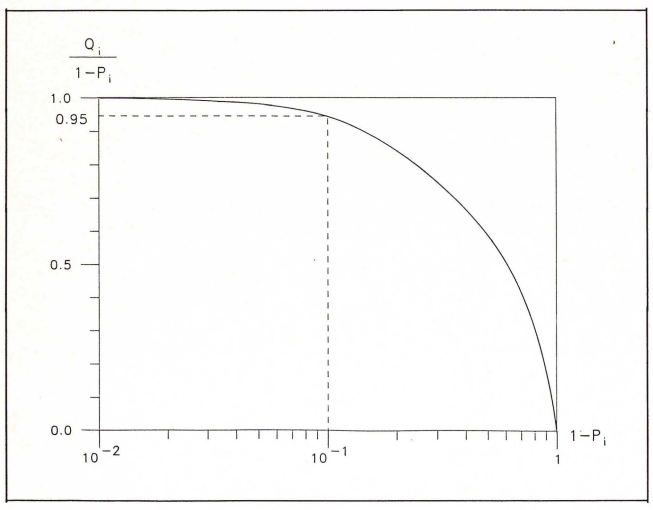
5. Final remark

This article presented two valid methods to estimate the unknown frequency Φ of CTL-precursors: the Maximum-Likelihood- and the Minimum-Chi-square-method.

TASWELL [1981] recommends the Minimum-Chi-square-method referring to BERKSON [1980].

But Berkson's statements are controversial. For example Pfanzagl, Rao and others (in BERKSON [1980]) judge the ML-method to be the better method. FAZEKAS [1981] who considers in special aspects of a Limiting Dilution Assay prefers the ML-method for LD experiments.

Fig. 2. The probability that a particular responding culture is clonal in dependence of the fraction of the observed positive responding cultures. (c.p. MILLER [1982], p. 224).



Appendix

(a) In order to calculate the ML-estimator $\hat{\Phi}_{ML}$ one has to maximize the function

$$\ln L = \sum_{i=1}^{I} \left[\ln \binom{n_i}{r_i} - r_i \Phi x_i + (n_i - r_i) \ln (1 - e^{-\Phi x_i}) \right] .$$

The first and the second derivative are:

$$\frac{d \ln L}{d\Phi} = \sum_{i=1}^{I} \left[-r_i x_i + \frac{(n_i - r_i) x_i e^{-\Phi x_i}}{(1 - e^{-\Phi x_i})} \right] ,$$

$$\frac{d^2 \ln L}{d\Phi^2} = \sum_{i=1}^{I} \left[\frac{-(n_i - r_i) x_i^2 e^{-\Phi x_i} (1 - e^{-\Phi x_i}) - (n_i - r_i) x_i^2 e^{-\Phi x_i} e^{-\Phi x_i}}{(1 - e^{-\Phi x_i})^2} \right] .$$

$$= \sum_{i=1}^{I} \left[\frac{-(n_i - r_i) x_i^2 e^{-\Phi x_i}}{(1 - e^{-\Phi x_i})^2} \right] .$$

(b) In order to calculate the Minimum-Chi-square-estimator $\hat{\Phi}_{MC}$ one has to minimize the function

$$\chi^2 = \sum_{i=1}^{I} \frac{(r_i - n_i e^{-\Phi x_i})^2}{n_i e^{-\Phi x_i}(1-e^{-\Phi x_i})} = \sum_{i=1}^{I} \frac{(r_i - n_i e^{-\Phi x_i})^2}{n_i (e^{-\Phi x_i} - e^{-2\Phi x_i})} .$$

The first and the second derivative are given by

$$\frac{d\chi^2}{d\Phi} = \sum_{i=1}^{I} \frac{2(r_i - n_i e^{-\Phi x_i}) n_i x_i e^{-\Phi x_i} n_i (e^{-\Phi x_i} - e^{-2\Phi x_i}) - \left[(r_i - n_i e^{-\Phi x_i})^2 n_i (e^{-\Phi x_i}(-x_i) - (-2x_i)e^{-2\Phi x_i})\right]}{n_i^2 e^{-2\Phi x_i}(1-e^{-\Phi x_i})^2}$$

$$= \sum_{i=1}^{I} \left[\frac{n_i x_i e^{-\Phi x_i}(2r_i - n_i) + r_i^2 x_i (e^{\Phi x_i} - 2)}{n_i (1-e^{-\Phi x_i})^2} \right],$$

$$\frac{d^2\chi^2}{d\Phi^2} = \sum_{i=1}^{I} \frac{\left[-n_i x_i^2 e^{-\Phi x_i}(2r_i - n_i) + r_i^2 x_i^2 e^{-\Phi x_i}\right]\left[n_i (1-e^{-\Phi x_i})^2\right] - \left[(2n_i x_i r_i e^{-\Phi x_i} - n_i^2 x_i e^{-\Phi x_i} + r_i^2 x_i e^{\Phi x_i} - 2r_i^2 x_i)(2n_i x_i e^{-\Phi x_i} - 2x_i n_i e^{-2\Phi x_i})\right]}{n_i^2 (1-e^{-\Phi x_i})^4}$$

$$= \sum_{i=1}^{I} \left[\frac{n_i^2 x_i^2 (e^{-\Phi x_i} - e^{-3\Phi x_i}) + n_i r_i x_i^2 (-2e^{-\Phi x_i} + 2e^{-3\Phi x_i}) + r_i^2 x_i^2 (e^{\Phi x_i} - 4 + 7e^{-\Phi x_i} - 4e^{-2\Phi x_i})}{n_i (1-e^{-\Phi x_i})^4} \right].$$

The APL-programme LIDIA including documentation can be requested by the 2nd author.

```
      LIDIA
 DATA ANALYSIS FOR LIMITING DILUTION ASSAY ACCORDING TO SINGLE HIT POISSON MODEL
 ------------------------------------------------------------------------------

 IDENTIFICATION OF EXPERIMENT (NOT MORE THAN 50 SYMBOLS):
TEST EXAMPLE, LDA
 NUMBER OF DOSES:
□:
      4
 PLEASE INPUT FOR EACH DOSIS:
 NUMBER OF CULTURES, CELLS PER CULTURE, NUMBER OF NEG. CULTURES
DOSE 1:
□:
      24 6000 0
DOSE 2:
□:
      24 2000 2
DOSE 3:
□:
      24 1000 8
DOSE 4:
□:
      24 500 15
 INPUT OF NAME OF THIS DATASET:
 ONLY LETTERS OR FIGURES, BUT BEGINNING WITH LETTERS!
TEST1
 MINIMUM CHI SQUARE ESTIMATOR ?, PLEASE INPUT  MCE
 MAXIMUM LIKLIHOOD ESTIMATOR ?,  PLEASE INPUT  MLE
MLE

means 'input'
```

Call of APL-programme LIDIA.

References

ARMITAGE, P. [1971]: Statistical methods in medical research. Blackwell Scientific Publications, Oxford and Edinburgh.

BERKSON, J. [1980]: Minimum chi-square, not maximum likelihood! (with discussion), Annals of Statistics **8,** 457–487.

FAZEKAS, S. [1982]: Review article: The evaluation of limiting dilution assay. Journal of Immunological Methods **49,** R11–R23.

FINNEY, D. J. [1978]: Statistical method in biological assay. 3rd ed. Charles Griffin & Company, London.

LINDER, A. [1964]: Statistische Methoden. Bd. 3, Birkhäuser Verlag Basel und Stuttgart.

MILLER, R. G. [1982]: An overview in »Isolation, characterisation, and utilization of t lymphocyte clones«. Eds. C. Garrison, F. Fathmann, Fitch, W.

OLSCHEWSKI, M., SCHACH, S., und M. SCHUMACHER [1979]: »Bioassay«, Arbeitsbericht der Abteilung Statistik, Universität Dortmund, F.R.G.

TASWELL, C. [1981]: Limiting dilution assays for the determination of immunocompetent cell frequencies. I. Data Analysis. Journal of Immunology **126,** 1614–1619.

Date of receipt: February 9th, 1983.

The author's address: Dipl.-Stat. Gisela Kynast and Prof. Dr. Ernst Weber, Abt. Biostatistik, Institut für Dokumentation, Information und Statistik DKFZ, Im Neuenheimer Feld 280, D-6900 Heidelberg 1.

EDV in Medizin und Biologie **14** (2), 58–61, ISSN 0300-8232

Refined estimation of the logarithmic series distribution

D. Böhning

Summary

This paper presents an improved estimation technique for the multivariate logarithmic series distribution. The technique is based on a stable fixed point principle. In addition, the problem of choosing the initial values is discussed.

Zusammenfassung

Diese Arbeit stellt ein verbessertes iteratives Verfahren zur Schätzung der multivariaten logarithmischen Reihenverteilung vor. Das Verfahren basiert auf einem numerisch stabilen Fixpunktprinzip. Zusätzlich wird das Problem diskutiert, in welcher Weise die Startwerte des Verfahrens gewählt werden können.

1. The Problem

We consider the density of the logarithmic series distribution (also called log series distribution)

$$\theta^x/\{-x \ln(1-\theta)\}, \quad x \in \mathbb{N} := \{1,2,\ldots\}; \tag{1}$$

$0 < \theta < 1$, and its multivariate version

$$\frac{\left(\sum_{i=1}^{k} x_i - 1\right)!}{-\ln(\theta_{k+1}) \prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} \theta_i^{x_i} \tag{2}$$

where $\theta = (\theta_1, \ldots, \theta_{k+1})^T \in S_> := \{\theta > 0 \mid \sum_{i=1}^{k+1} \theta_i = 1\}$; $\quad x = (x_1, \ldots, x_k)^T \in \mathbb{N}^k$.

This paper presents an improved technique for finding the maximum likelihood estimator of θ. For the sake of brevity we only consider (2) and note that $k=1$ leads to (1).

2. Historical Background

The logarithmic series distribution goes back to FISHER/CORBERT/WILLIAMS (1943) who used it to describe the distribution of certain species in a biological community. Since then, the logarithmic series distribution has experienced a variety of applications. Ecologists use it to fit the *species abundance distribution* of an ecological population; in addition,

$$\alpha := \{-\ln(1-\theta)\}^{-1}$$

is used as an index (diversity index) to describe the variety of species in that population (PIELOU 1975; KEMPTON 1979). Recently, in the field of epidemiology, ISZÁK/JUHÁSZ-NAGY (1982) fit the logarithmic series distribution to morbidity and mortality data.

The problem of estimating θ via the maximum likelihood method is primarily treated in the literature for the univariate case, for which the likelihood equation becomes

$$\bar{x}/\theta + 1/\{(1-\theta)\ln(1-\theta)\} = 0 \tag{3}$$

where $\bar{x}$ is the arithmetic mean of n i.i.d. observations $x_1,\ldots,x_n$. To solve (3) we could use the Newton-Raphson procedure. However, it is mentioned in BÖHNING (1983a) that Newton-Raphson works only with "good" initial choices of θ, and not at all in cases where $\bar{x}$ is small (near by 1). BIRCH (1963) introduces the parameter $\lambda := 1/(1-\theta)$ so that (3) becomes

$$1 + \bar{x} \ln(\lambda) - \lambda = 0 \tag{4}$$

The Birch algorithm – used thus far as a fitting procedure for the univariate log series distribution – is simply the Newton-Raphson procedure applied to (4). However, the Birch algorithm – although quadratically convergent, if at all – converges only with initial choices of λ, which are situated in the vicinity of the solution of (4). Iszák/Juhász-Nagy avoid the Birch algorithm for reasons similar to those mentioned above. They observe that solving (4) is equivalent to finding the fixed point of

$$G : (1, \infty) \to (1, \infty) \quad \text{defined by}$$

$$G(\lambda) := 1 + \bar{x} \ln(\lambda) \tag{5}$$

and use the iteration procedure

$$\lambda^{(j+1)} = G(\lambda^{(j)}), \lambda^{(1)} > 1 \tag{6}$$

for finding the fixed point. Unfortunately, they do not give a theoretical discussion of (6).

The results of Izsák/Juhász-Nagy neatly link up with work presented recently in BÖHNING (1983a, b). There, the multivariate case is treated and it is shown that – based on i.i.d. observations

$(x_1^{(1)},\ldots,x_k^{(1)})^T,\ldots,(x_1^{(n)}),\ldots,x_k^{(n)})^T \in IN^k$ – there exists a unique maximum likelihood estimate $\hat{\theta} \in S_>$, and that it is given by

$$\hat{\theta} = \frac{\ln(\hat{p})}{\Sigma \ln(\hat{p}) - 1} (\bar{x}_1,\ldots,\bar{x}_k, -1/\ln(\hat{p}))^T, \qquad (7)$$

where $\Sigma = \bar{x}_1 + \ldots + \bar{x}_k$, $\bar{x}_i = \frac{1}{n}(x_i^{(1)} + \ldots + x_i^{(n)})$, for $i = 1,\ldots,k$;

and $\hat{p}$ is the fixed point of $F : (0,1) \to (0,1)$ defined by

$$F(p) := 1/\{1 - \Sigma \ln(p)\} \qquad (8)$$

The characterization given in (7) is of interest, since a problem involving k variables is reduced to a problem involving only one variable.

Let us define G more generally as

$$G(\lambda) = 1 + \Sigma \ln(\lambda)$$

Then, $\hat{p}$ is the fixed point of F if and only if $\hat{\lambda} = 1/\hat{p}$ is the fixed point of G. Moreover,

$$\hat{\theta} = \frac{\ln(\hat{\lambda})}{\Sigma \ln(\hat{\lambda}) + 1} (\bar{x}_1,\ldots,\bar{x}_k, 1/\ln(\hat{\lambda}))^T.$$

We now give a discussion of the simpler mapping G.

3. The Operator G

It is fairly obvious that G is an operator on $[1,\infty)$. Because $\frac{dG}{d\lambda} = \Sigma/\lambda$, G is *not* a contraction operator.

However, if, for suitable $\varepsilon > 0$ and $\lambda \in IB_\varepsilon := \{\lambda \in IR \mid \Sigma + \varepsilon \leqq \lambda\}$, $G(\lambda)$ is also an element of IB_ε, then G is a contraction on IB_ε using the mean value theorem. Note that it is neither sufficient to consider $IB := \{\lambda \{\in IR \mid \Sigma < \lambda\}$ nor $I\bar{B} = \{\lambda \in IR \parallel \Sigma \leqq \lambda\}$, because IB is not complete and

$$\sup_{\lambda \in \overline{IB}} \frac{dG}{d\lambda}(\lambda) = 1$$

Lemma 1

If

$$\varepsilon = G(\Sigma) - \Sigma \qquad (9)$$

then $\varepsilon > 0$.

Proof

This is geometrically quite obvious (see Figure 1).

A formal argument runs as follows. We consider

$$g(\lambda) := 1 + \Sigma \ln(\lambda) - \lambda$$

which is a strictly concave function $\left(\frac{d^2g}{(d\lambda)^2} < 0\right)$

and uniquely maximized at $\lambda = \Sigma$ (see Figure 2). We show $g(\Sigma) > 0$.

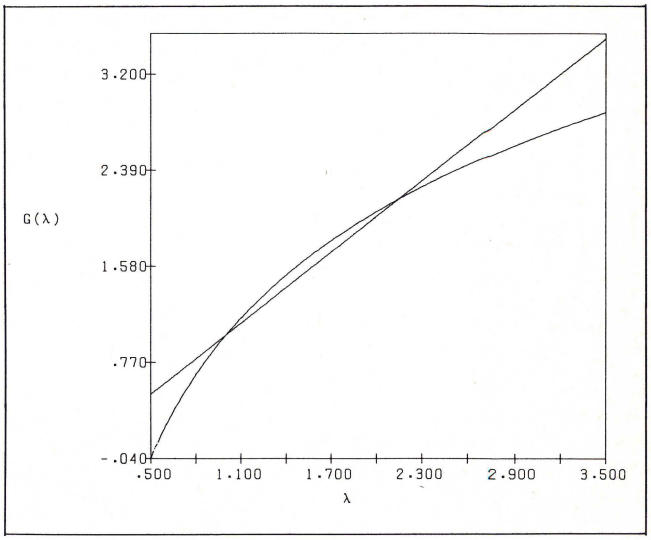


Figure 1. Graph of G for $\Sigma = 1.5$ with trivial fixed point $\lambda_1 = 1$ and fixed point λ_2 corresponding to the MLE$(1 < \Sigma < \lambda_2)$.

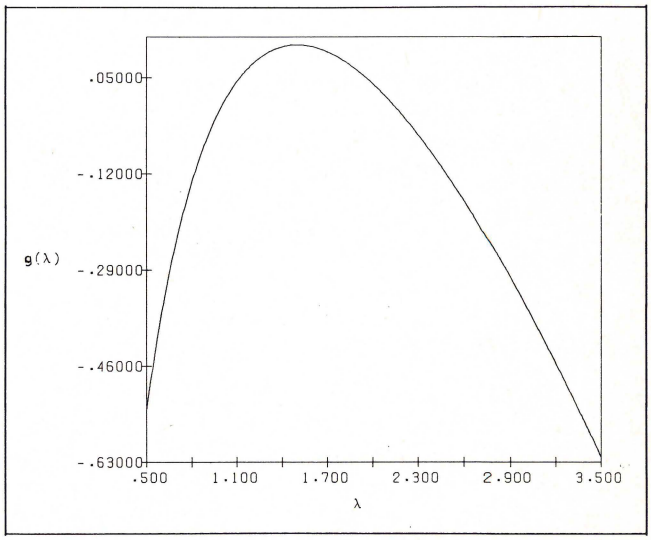


Figure 2. Graph of g for $\Sigma = 1.5$ with $g(\Sigma) > 0$.

Using the mean value theorem we have

$$\begin{aligned} g(\Sigma) &= g(1 + (\Sigma - 1)) \\ &= g(1) + (\Sigma - 1)\frac{dg}{d\lambda}(1 + \beta(\Sigma - 1)) \\ &= (\Sigma - 1)\left[\frac{\Sigma}{1 + \beta(\Sigma - 1)} - 1\right] \end{aligned}$$

for some $\beta \in (0,1)$. Since $\Sigma > 1$ we have

$$\Sigma > (1 - \beta)\,1 + \beta\,\Sigma$$

for all $\beta \in [0,1)$ and thus $g(\Sigma) > 0$ or $\varepsilon > 0$.

For the rest of this paper ε will be choosen according to (9).

Lemma 2

$$\lambda \in IB_\varepsilon \quad \text{implies } G(\lambda) \in IB_\varepsilon$$

Proof

Because $\frac{dG}{d\lambda} > 0$ on $(0,\infty)$, G is strictly monotone increasing.

Thus
$\Sigma < \Sigma + \varepsilon \leqq \lambda$ implies $G(\Sigma) < G(\Sigma+\varepsilon) \leqq G(\lambda)$, and therefore $\lambda \in IB_\varepsilon$ e.g. $\lambda \geqq \Sigma+\varepsilon$ implies $G(\lambda) \geqq G(\Sigma) = \Sigma+\varepsilon$

From Banach's contraction theorem we attain the following

Theorem 1

Let $\lambda^{(1)} \geqq G(\Sigma)$. Then $(\lambda^{(j)})$ defined by

$$\lambda^{(j+1)} = G(\lambda^{(j)}) \tag{10}$$

converges linearly to the unique fixed point of G in IB_ε.

Since statisticians have learned to be sensitive with respect to problems of numerical stability of statistical algorithms (see for example CHAN/GOLUB/LEVEQUE (1982) or CHAMBERS (1977)) the following theorem is also of importance.

Theorem 2

G is numerically stable on IB_ε, that is, the condition number

$$\left| \frac{\lambda}{G(\lambda)} \frac{dG}{d\lambda}(\lambda) \right| < 1$$

for all $\lambda \in IB_\varepsilon$.

Proof

Note that $\varphi(\lambda) := \frac{\lambda}{G(\lambda)} \frac{dG}{d\lambda}(\lambda) = \frac{\Sigma}{1 + \Sigma \ln(\lambda)}$ is

strictly monotone decreasing with $\varphi(\lambda) \to 0$ as $\lambda \to +\infty$. In addition, by Lemma 1 $\varphi(\Sigma) < 1$ concluding $\varphi(\lambda) < 1$ for all $\lambda \in IB_\varepsilon$.

4. Choice of the Initial Value

We have mentioned that the fixed point procedure based on G converges only with a linear rate. Particularly if Σ is small, convergence can be painfully slow. To improve the convergence behaviour we can try to choose the initial estimate, close to the final (maximum likelihood) estimate.

The reader might recall that we are interested in solving the Birch equation (4) $g(\lambda) = 0$ for some $\lambda > 1$. Note that $g(1) = 0$. Let us write the solution of (4) as $\lambda = 1+h$, $h>0$ and consider the Taylor approximations of $g(1+h)$

$$g(1) + h\frac{dg}{d\lambda}(1) + \frac{1}{2}h^2\frac{d^2g}{(d\lambda)^2}(1) \qquad \text{(quadratic)}$$

$$+ \frac{1}{6}h^3\frac{d^3g}{(d\lambda)^3}(1) \qquad \text{(cubic)}$$

For the quadratic case we arrive at

$$h(\Sigma-1) - \frac{1}{2}h^2\Sigma = 0$$

which leads to the non-zero solution

$$h_q = 2(\Sigma-1)/\Sigma$$

The cubic case

$$h(\Sigma-1) - \frac{1}{2}h^2\Sigma + \frac{1}{3}h^3\Sigma = 0$$

has only a real non-zero solution if

$$\Sigma \leqq \frac{16}{13}.$$

In this case

$$h_c = \frac{3}{4} - \sqrt{3}\sqrt{\frac{3}{16} - \frac{(\Sigma-1)}{\Sigma}}$$

This leads to initial estimates

$$\lambda_q^{(1)} = 3 - \frac{2}{\Sigma} \tag{11}$$

based on the quadratic approximation and

$$\lambda_c^{(1)} = \frac{7}{4} - \sqrt{3}\sqrt{\frac{1}{\Sigma} - \frac{13}{16}} \tag{12}$$

based on the cubic approximation. Table 1 compares initial estimates (out of parentheses $\lambda_c^{(1)}$, in parentheses $\lambda_q^{(1)}$) and final estimates. In addition the number of iteration steps needed to reach 4 significant numbers of accuracy is given. Obviously, $\lambda_c^{(1)}$ gives the better approximation. Note that the initial estimates $\lambda_c^{(1)}$ meet the condition $\lambda_c^{(1)} \geqq G(\Sigma)$ of Theorem 1, if $\lambda_c^{(1)}$ can be used. $\lambda_q^{(1)}$ does not necessarily satisfy this condition and has to be checked. If $\lambda_q^{(1)} < G(\Sigma)$ some other initial value, for example $G(\Sigma)$, ought to be used.

Σ	$G(\Sigma)$	Initial		Final	Number of steps
1.01	1.01005	1.02007	(1.01980)	1.02007	1
1.02	1.02020	1.04030	(1.03922)	1.04030	1
1.03	1.03045	1.06071	(1.05825)	1.06071	1
1.04	1.04079	1.08033	(1.07692)	1.08131	3
1.05	1.05123	1.10220	(1.09524)	1.10184	22
1.06	1.06177	1.12335	(1.11321)	1.12253	31
1.07	1.07239	1.14482	(1.13084)	1.14334	37
1.08	1.08312	1.16667	(1.14815)	1.16429	40
1.09	1.09393	1.18894	(1.16514)	1.18536	42
1.10	1.10484	1.21169	(1.18182)	1.20656	43
1.11	1.11584	1.23502	(1.19820)	1.22789	43
1.12	1.12693	1.25901	(1.21429)	1.24933	43
1.13	1.13811	1.28377	(1.23009)	1.27090	43
1.14	1.14937	1.30946	(1.24561)	1.29258	43
1.15	1.16073	1.33624	(1.26087)	1.31437	43
1.16	1.17217	1.36437	(1.27586)	1.33629	43
1.17	1.18369	1.39419	(1.29060)	1.35832	43
1.18	1.19531	1.42616	(1.30508)	1.38047	42
1.19	1.20700	1.46102	(1.31933)	1.40273	42
1.20	1.21879	1.5	(1.33333)	1.42510	42
1.21	1.23065	1.54545	(1.34711)	1.44758	42
1.22	1.24260	1.60332	(1.36066)	1.47017	42
1.23	1.25463	1.71096	(1.37398)	1.49288	42

Table 1: Comparison of initial and final estimates

Supplement

It was pointed out to me by Prof. Johansen (University of Copenhagen) that the condition $\lambda^{(1)} > 1 + \Sigma \ln(\Sigma)$ of theorem 1 can be relaxed to $\lambda^{(1)} > 1$. In fact, $G(\lambda) = 1 + \Sigma \ln(\lambda)$ is an increasing function ($G'(\lambda) = \Sigma/\lambda > 0$) and $G(\lambda) \geqslant \lambda$ for $\lambda \in (1, \hat{\lambda})$, $G(\lambda) \leqslant \lambda$ for $\lambda \in (\hat{\lambda}, \infty)$. Thus, $(\lambda^{(n)})$ converges monotonously to the unique fixed point $\hat{\lambda} \in (1, \infty)$ with local linear convergence rate.

In addition, investigation of the function g (Lemma 1) gives a deeper look into the Birch algorithm

$$\lambda^{(j+1)} = (1 - \Sigma + \Sigma \ln(\lambda^{(j)}))/(1 - \frac{\Sigma}{\lambda^{(j)}})$$

which is just the Newton-Raphson sequence applied to g. g is a strictly concave function maximised at $\lambda = \Sigma$. Thus, the Birch algorithm converges to $\hat{\lambda}$ if $\lambda^{(1)} > \Sigma$, and to 1 if $\lambda^{(1)} < \Sigma$. Convergence rate is quadratically.

References

BIRCH, M. W.: An algorithm for the logarithmic series distribution. Biometrics **19,** 1963, 651–652.

BÖHNING, D.: Maximum likelihood estimation of the logarithmic series distribution. Statistische Hefte **24,** 1983a, (to appear).

BÖHNING, D.: Iterative Construction of the MLE for the logarithmic Series Distribution. Rate of Convergence. Statistische Hefte, 1983b, (to appear).

CHAN, T. F., GOLUB, G. H., and LEVEQUE, R. J.: Updating Formulae and a Pairwise Algorithm for Computing Sample Variances. In: Proceedings of the 5th Compstat-Congress. Edited by H. Caussinus, P. Ettinger, and R. Tomassone. Wien 1982, p. 30–41.

CHAMBERS, J. M.: Computational Methods for Data Analysis. New York 1977.

FISHER, R. A., CORBET, A. S., and WILLIAMS, C. B.: The relation between the number of species and the number of individuals in a random sample of an animal population. Journal of Animal Ecology **12,** 42–57.

ISZÁK, J., JUHÁSZ-NAGY, P.: Studies of Lognormality on Mortality Statistics. Biometrical Journal **24,** 1982, 731–741.

KEMPTON, R. A.: The structure of species abundance and measurement of diversity. Biometrics **35,** 1979, 307–321.

PIELOU, E. C.: Ecological Diversity. New York 1975.

Date of receipt: May 5th, 1983.
The author's address: Dr. D. Böhning, Abt. Epidemiologie, Institut für Soziale Medizin, Freie Universität Berlin, Kelchstraße 31, D-1000 Berlin 41.

PROGRAMMINFORMATIONEN/INFORMATION CONCERNING PROGRAMS

EDV in Medizin und Biologie **14** (2), 61–64, ISSN 0300-8232

Computerprogramme zur Berechnung exakter Wahrscheinlichkeiten für den Mann-Whitney-U-Test mit Bindungen (Uleman-Test)

W. Buck

Zusammenfassung

Es werden FORTRAN-Programme zur Berechnung der exakten Punkt- und Übertretungswahrscheinlichkeiten unter H_0 bei ein- und zweiseitiger Fragestellung für den U-Test mit Bindungen (ULEMAN-Test) angegeben. Die Bindungen werden auf der Grundlage von Durchschnittsrängen (mid-ranks) behandelt. Da in der statistischen Auswertungspraxis bei der Durchführung von U-Testen fast immer Bindungen auftreten, ist die exakte Berücksichtigung von Bindungen von besonderer Bedeutung.

Summary

FORTRAN-programs are submitted for the calculation of the exact point and significance probabilities under H_0 with one-sided and two-sided statement of the problem for the U-test with ties (ULEMAN-test). The ties are handled on the basis of the mid-rank method. As in the practice of statistical analyses there appear ties in nearly any case of U-test performance the exact consideration of ties is particularly important.

1. Einleitung

BUCK hat (1976) einen Algorithmus zur Berechnung der exakten Punktwahrscheinlichkeiten unter H_0 für den U-Test mit Bindungen (ULEMAN-Test) beschrieben. Im folgenden sollen die Programme für diesen Algorithmus angegeben und erläutert werden. Mit dem gleichen Algorithmus wurden die Tabellen der kritischen Werte für den ULEMAN-Test brechnet (BUCK, 1976), welche sich auch auszugsweise in dem Tafelband von LIENERT (1975) finden.

2. Methode

Gegeben seien die unabhängigen Stichproben $X = (x_1, x_2, \ldots, x_m)$ und $Y = (y_1, y_2, \ldots, y_n)$ bestehend aus Rang- oder Meßwerten. Zur Testdurchführung werden die beiden Stichproben zusammengefaßt und nach steigenden Werten die Ränge 1, 2, 3, . . . ,C zugeordnet, wobei die gleichen Werte die gleichen Rangzahlen erhalten. Bezeichnen wir die Anzahl der Bindungsgruppen mit C, die Anzahl der Werte in der i. Bindungsgruppe mit T_i und die Anzahl der Werte in den Stichproben X (Y), welche der i. Bindungsgruppe angehören, mit u_i (v_i) für i = 1, 2, . . . ,C, so ergibt sich das folgende Marginalschema:

Tabelle 1: Marginalschema zum U-Test nach ULEMAN

Rangklassen	1 2	. . . C	
Ranghäufigkeiten zur Stichpr. X	u_1 u_2	. . . u_C	$\Sigma u_i = m$
Ranghäufigkeiten zur Stichpr. Y	v_1 v_2	. . . v_C	$\Sigma v_i = n$
Ranghäufigkeiten zur Stichpr. X+Y	T_1 T_2	. . . T_C	N=m+n
			$T_i = u_i + v_i$ für i=1,2, . . ., C

Zur Gewinnung der Nullverteilung des ULEMAN-Tests genügt es, alle möglichen Feldvektoren
$u = (u_1, u_2, \ldots, u_c)$ und $v = (v_1, v_2, \ldots, v_c)$
zu generieren, welche die folgenden Restriktionen erfüllen:
$\Sigma u_i = m$, $\Sigma v_i = n$, $0 \le u_i, v_i \le T_i$
$u_i + v_i = T_i$ mit i = 1, 2, . . . ,C (2.1)

Die C-Tupel u können mit dem Algorithmus zum »Generieren der Feldvektoren zu vorgegebenen Marginalsummen« (BUCK, 1976, Abb. 1) erzeugt werden. Der Vektor v läßt sich dann nach (2.1) durch $v_i = T_i - u_i$ berechnen. Ein FORTRAN-Unterprogramm dieses Algorithmus zeigt die Abb. 1.

Jeder Feldbesetzung u, v ist die Prüfgröße

$$U = \sum_{i=2}^{C} u_i (v_1 + v_2 + \ldots + v_{i-1}) + 1/2 \sum_{i=1}^{C} u_i \cdot v_i \qquad (2.2)$$

und dieser die Punktwahrscheinlichkeit

$$p(u\,|\,T;m) = \binom{T_1}{u_1} \cdot \binom{T_2}{u_2} \cdots \binom{T_C}{u_C} \Big/ \binom{N}{m} \qquad (2.3)$$

zugeordnet. Zur Berechnung der Binomialkoeffizienten in (2.3) wird das FORTRAN-Unterprogramm 'BINKOE' der Abb. 2 benutzt.

Die Routinen 'TUP2' und 'BINKOE' sind Unterprogramme für 'PU1'. Zur Verifizierung ist nur nötig, die 3 genannten Programme gemeinsam zu linken und 'PU1' von einem Haupt- oder weiteren Unterprogramm aus mit den in der Variablenliste von 'PU1' (vgl. Abb. 3) definierten Parametern aufzurufen.

Die führenden Zahlen in den Programmlisten von 'TUP2', 'BINKOE' und 'PU1' sind Zeilennummern und können bzw. müssen fortgelassen werden. Das nach einer Zeilennummer eventuell folgende Zeichen " (Anführungsstriche) ist das Kommentarzeichen in CALL-FORTRAN und muß eventuell durch ein anderes Kommentarzeichen (in der Regel: C) ersetzt werden.

Ein anderes Verfahren zur Berechnung der Nullverteilung des ULEMAN-Tests haben STUCKY & VOLLMAR (1976) angegeben, welches auf einer Methode von KRAUTH (1971) beruht.

3. Beispiel

Wir wollen das Beispiel 2 aus (BUCK, 1976) betrachten: Dafür gilt: $X = (x_1, x_2, x_3, x_4) = (7, 9, 9, 10)$ und $Y = (y_1, y_2, y_3, y_4, y_5) = (6, 7, 7, 8, 10)$. Die Zusammenfassung und aufsteigende Anordnung der Stichproben ergibt:
$6_y\ 7_y\ 7_y\ 7_x\ 8_y\ 9_x\ 9_x\ 10_y\ 10_x$
Die zugeordneten Ränge sind:
$1_y\ 2_y\ 2_y\ 2_x\ 3_y\ 4_x\ 4_x\ 5_y\ 5_x$

Damit nimmt das Marginalschema für dieses Beispiel folgende Werte an:

Rangklassen	1	2	3	4	5	
Ranghäufigkeiten zu X	u_1=0	u_2=1	u_3=0	u_4=2	u_5=1	m=4
Ranghäufigkeiten zu Y	v_1=1	v_2=2	v_3=1	v_4=0	v_5=1	n=5
Ranghäufigkeiten zu X+Y	T_1=1	T_2=3	T_3=1	T_4=2	T_5=2	N=9

Für die Prüfgröße UK ergibt sich für diese Feldbesetzungen nach (2.2):
$UK = u_2v_1 + +u_3 (v_1+v_2) + u_4 (v_1 + v_2 + v_3) + u_5 (v_1 + v_2 + v_3 + v_4) + (u_1v_1 + u_2v_2 + u_3v_3 + u_4v_4 + u_5v_5) = 14.5$
Somit ist ›PU1‹ mit den Parametern
N = 9, M = m = 4, C = 5, T = (1, 3, 1, 2, 2) und UK = 14.5
aufzurufen. Nach dem Rücksprung ergeben sich folgende Funktionswerte:

NTUPEL = 30
PR (U ≤ UK) = P(1) = 0.88889
UR = M(N-M)-UK = 5.5
UMIN = MINIMUM (UK, UR) = 5.5
UMAX = MAXIMUM (UK, UR) = 14.5
PR (U ≤ UR) = P(2) = 0.19841
PR (U ≤ UMIN) + PR (U ≥ UMAX) = P(3) = 0.30952
PR (U = UK) = P(4) = 0.06349

i	U_i	PD_i
1	0.0	0.00794
2	2.0	0.02381
3	3.5	0.04762
4	4.0	0.00794
5	5.5	0.11111
6	7.0	0.02381
7	7.5	0.11111
8	9.0	0.12698
9	9.5	0.04762
10	11.0	0.17460
11	12.5	0.01587
12	13.0	0.12698
13	14.5	0.06349
14	15.0	0.02381
15	16.5	0.06349
16	18.5	0.01587
17	20.0	0.00794

```
"     GENERIEREN DER FELDVEKTOREN ZU VORGEGEBENEN MARGINALSUMMEN
"     ==========================================================
"
"     IFUNC   = FUNKTIONSPARAMETER
"       "  =-1= ES HANDELT SICH UM DEN INITIALISIERUNGSAUFRUF. NACH RUECKKEHR AUS DEM UP 'TUP2'
"                GILT IFUNC=0, FALLS IN DIESEM AUFRUF EIN WEITERES TUPEL ERZEUGT WURDE.
"       "  = 0= ES WURDE GEGENUEBER DEM VORANGEHENDEN C-TUPEL EIN LEXIKOGRAPHISCH FOLG. C-TUPEL
"                ERZEUGT.
"       "  =+1= ES KANN KEIN WEITERES C-TUPEL ERZEUGT WERDEN
"     C       = DIMENSION DER VEKTOREN T UND U.
"     T       = VEKTOR DER SPALTENSUMMEN MIT T(I)≥1 UND T(1)+T(2)+...+T(C)=N
"     M       = U(1)+U(2)+ ... U(C) UND 0≤U(I)≤T(I)
"     U       = ZU GENERIERENDER FELDVEKTOR (= C-TUPEL). DIE FELDVEKTOREN WERDEN IN LEXIKO-
"               GRAFISCHER REIHENFOLGE ERZEUGT. DABEI WIRD DER VEKTOR
"               U=(U(1),U(2),...,U(C)) ALS ZAHL MIT DER BASIS M+1 UND DEN ZIFFERN
"               U(1), U(2), ... , U(C) AUFGEFASST. EIN AUFRUF VON 'TUP2' ERGIBT
"               DEN LEXIKOGRAFISCH FOLGENDEN FELDVEKTOR.
"
10      SUBROUTINE TUP2 (IFUNC, C, T, M, U)
20 "
30      INTEGER*4  C, S, T(1), U(1)
40 "
50      IF (IFUNC) 10, 20, 1000
60 "
70 "    INITIALISIERUNG
80 "    ---------------
90 "
100    10 I      = 0
110       S      = 0
120       IFUNC  = 0
130       GO TO 90
140 "
150 "    GENERIEREN DER C-TUPELE
160 "    -----------------------
170 "
180    20 I      = C
190       S      = M
200    40 S      = S - U(I)
210       I      = I - 1
220       IF (I) 50, 50, 60
230    50 IFUNC  = 1
240       GO TO 1000
250    60 IF (S  - M   ) 70, 40, 40
260    70 IF (U(I)-T(I)) 80, 40, 40
270    80 U(I)   = U(I) + 1
280       S      = S    + 1
290    90 J      = C
300   100 IF (S+T(J)-M) 120, 120, 110
310   110 U(J)   = M - S
320       GO TO 130
330   120 U(J)   = T(J)
340   130 S      = S + U(J)
350       J      = J - 1
360       IF (J-I) 1000, 1000, 100
370 "
380  1000 RETURN
390       END
```

Abb. 1. FORTRAN-Unterprogramm zum »Generieren der Feldvektoren zu vorgegebenen Marginalsummen«.

```
10 "    BERECHNUNG VON BINOMIALKOEFFIZIENTEN
20 "    ====================================
30 "
40 "    DAS UP BERECHNET DIE BINOMIALKOEFFIZIENTEN VON N UEBER K FUER N,K≥0
50 "    IDARST = 1 = MAN ERHAELT DEN ECHTEN FUNKTIONSWERT.
60 "                 FUER N<251 SIND ALLE BINOMIALKOEFFIZIENTEN <1.E+75
70 "    IDARST = 2 = MAN ERHAELT DEN DEKADISCHEN LOGARITHMUS DES FUNKTIONSWERTES.
80 "                 DIE LOGARITHMISCHE DARSTELLUNG IST NOTWENDIG, WENN DER ECHTE
90 "                 FUNKTIONSWERT >1.E+75 WIRD. DIES KANN NUR FUER N>250 AUFTRETEN.
100 "   DIE ARGUMENTE N UND K KOENNEN NICHTNEGATIVE GANZZAHLIGE WERTE ANNEHMEN,
110 "   WOBEI K=N ZUGELASSEN IST. K .GT. N IST NUR BEI IDARST=1 MOEGLICH.

10      REAL FUNCTION BINKOE*4 (IDARST, N, K)
20 "
30    10 N1     = N + 1
40 "    BERUECKSICHTIGEN DER SYMMETRIEEIGENSCHAFT DER BINOMIALKOEFFIZIENTEN
50      KK      = K
60      IF (K .GT. N/2) KK = N - K
70      IF (IDARST .EQ. 2) GO TO 30
80 "    ECHTE DARSTELLUNG
90      BINKOE = 0.
100      IF (KK .LT. 0) GO TO 100
110      BINKOE = 1.
120      IF (KK .EQ. 0) GO TO 100
130      DO 20 IK = 1, KK
140   20 BINKOE = BINKOE*(N1-IK)/IK
150      GO TO 100
160 "    LOGARITHMISCHE DARSTELLUNG
170   30 BINKOE = 0.
180      IF (KK .LE. 0) GO TO 100
190      DO 40 IK = 1, KK
200   40 BINKOE = BINKOE+ALOG10(FLOAT(N1-IK)/IK)
210 "
220  100 RETURN
230      END
```

Abb. 2. FORTRAN-Unterprogramm zur Berechnung von Binomialkoeffizienten.

```
C     BERECHNEN DER EXAKTEN WAHRSCHEINLICHKEITEN FUER DEN U-TEST NACH ULEMAN
C     ======================================================================
C
C     FUER JEDEN AUFRUF WIRD DURCH 'PU1' DIE VOLLSTAENDIGE EXAKTE DICHTEFUNKTION
C     ZUR ZUFALLSVARIABLEN U BERECHNET. DIESE IST IM ALLGEMEINEN ASYMMETRISCH.
C
C     VARIABLENLISTE
C
C     N       = WERTEANZAHL IN DER 1. UND 2. STICHPROBE
C      M      =      "       "   "  1. STICHPROBE
C      C      = ANZAHL BINDUNGSGRUPPEN (DIMENSION DES VEKTORS T)
C      T(I)   =   "    WERTE, DIE DER I. BINDUNGSGRUPPE ANGEHOEREN
C      UK     = AKTUALWERT DER ZUFALLSVARIABLEN U FUER DEN VORLIEGENDEN 2-STICHPROBENFALL
C               ES SEI UR=M*(N-M)-UK DAS KOMPLEMENT ZU UK UND UMIN=MINIMUM(UK,UR) SOWIE
C               UMAX=MAXIMUM(UK,UR)
C      P(1)   = EINSEITIGE  UEBERTRETUNGSWAHRSCHEINLICHKEIT PR(U≤UK)
C      P(2)   = EINSEITIGE  UEBERTRETUNGSWAHRSCHEINLICHKEIT PR(U≤UR)
C      P(3)   = ZWEISEITIGE UEBERTRETUNGSWAHRSCHEINLICHKEIT PR(U≤UMIN)+PR(U>UMAX)
C      P(4)   = WAHRSCHEINLICHKEITSDICHTE                   PR(U=UK)
C     NTUPEL  = ANZAHL DER MOEGLICHEN FELDVEKTOREN UV, V
C     PD      = VEKTOR DER WAHRSCHEINLICHKEITSDICHTEN MIT DER DIMENSION 2*M*(N-M)+1.
C               DER ZUFALLSVARIABLEN U ENTSPRICHT IN PD DIE PUNKTWAHRSCHEINLICHKEIT
C               PD(2*U+1). GILT PD(2*U+1)=0, SO EXISTIERT ZU DIESEM U-WERT KEINE
C               PUNKTWAHRSCHEINLICHKEIT. U IST DEFINIERT IN DEM INTERVALL 0≤U≤M*(N-M).
C
C     UV UND V SIND DIE MIT DEM ALGORITHMUS ZUM "GENERIEREN DER FELDVEKTOREN ZU VOR-
C     GEGEBENEN MARGINALSUMMEN" ERZEUGTEN FELDVEKTOREN UND MUESSEN FUER C>100 DIE
C     DIMENSION C HABEN.
C
      SUBROUTINE PU1 (N, M, C, T, UK, P, NTUPEL, PD)
C
      REAL*4     P(4), PD(1)
      INTEGER*4  C, T(C), UV(100), V(100)
C
C     SETZEN DES VEKTORS PD AUF 0.
      NU      = M*(N-M)*2 + 1
      DO 10 I = 1, NU
   10 PD(I)   = 0.
C     GENERIEREN DER FELDVEKTOREN UV ZU DEN VORGEGEBENEN MARGINALSUMMEN N, M, T
      NTUPEL = 0
      IFUNC  = -1
      BIKNN1 = BINKOE (1, N, M)
   20 CALL TUP2 (IFUNC, C, T, M, UV)
      IF (IFUNC) 100, 30, 60
   30 NTUPEL = NTUPEL + 1
C     BERECHNEN DER ZUFALLSVARIABLEN U
      DO 40 I = 1, C
   40 V(I)   = T(I) - UV(I)
      U      = 0.
      ISUM1  = 0
      ISUM2  = 0
      DO 45 I = 2, C
      ISUM1  = ISUM1 + V(I-1)
      ISUM2  = ISUM2 + V(I)*UV(I)
   45 U      = U + UV(I)*ISUM1
      U      = U + 0.5*(ISUM2+V(1)*UV(1))
C     BERECHNEN DER PUNKTWAHRSCHEINLICHKEITEN ZUM VEKTOR UV
      PR     = 1.
      DO 50 I = 1, C
   50 PR     = PR * BINKOE (1, T(I), UV(I))
      INU    = U*2. + 1.
      PD(INU)= PD(INU) + PR/BIKNN1
      GO TO 20
C
C     BERECHNEN DES VEKTORS P
C     -----------------------
C
   60 P(1)   = 0.
      P(2)   = 0.
      P(3)   = 0.
      UR     = M*(N-M) - UK
      UMIN   = AMIN1 (UK, UR)
      UMAX   = AMAX1 (UK, UR)
      DO 90 I = 1, NU
      U      = 0.5 * (I-1.)
      IF (U .EQ. UK) P(4) = PD(I)
      IF (U .GT. UK) GO TO 70
      P(1)   = P(1) + PD(I)
   70 IF (U .GT. UR) GO TO 80
      P(2)   = P(2) + PD(I)
   80 IF (U .GT. UMIN .AND. U .LE. UMAX) GO TO 90
      P(3)   = P(3) + PD(I)
   90 CONTINUE
C
  100 RETURN
      END
```

Abb. 3. FORTRAN-Unterprogramm zur Berechnung der exakten Wahrscheinlichkeiten für den U-Test mit Bindungen.

Literatur

BUCK, W. (1976): Der U-Test nach ULEMAN. EDV in Medizin und Biologie, **2,** 65–75.

LIENERT, G. A. (1973): Verteilungsfreie Methoden in der Biostatistik. Anton Hain · Meisenheim am Glan

LIENERT, G. A. (1975): Verteilungsfreie Methoden in der Biostatistik. Tafelband. Anton Hain · Meisenheim am Glan

KRAUTH, J. (1971): A locally most powerful tied rank test in the WILCOXON situation. Ann. Math. Stat., **42,** 1949–1956

STUCKY, W. und J. VOLLMAR (1976): Exact Probabilities for Tied Linear Rank Tests. J. Stat. Comp. Simul., **5,** 73–81

Eingegangen am 17. Dezember 1982.

Anschrift des Verfassers: W. Buck, Institut für Biostatistik, Kleiner Griechenmarkt 28–30, D-5000 Köln 1.