Journal of the

VOLUME 78
Number 4
December, 1988

WASHINGTON ACADEMY OF SCIENCES

ISSN 0043-0439

Issued Quarterly
at Washington, D.C.

CONTENTS

Washington Academy of Sciences

Founded in 1898

ACADEMY OFFICE
1101 N. Highland St.
Arlington, Va. 22201
Telephone: (703) 527-4800

The Journal

This journal, the official organ of the Washington Academy of Sciences, publishes historical articles, critical reviews, and scholarly scientific articles; proceedings of meetings of the Academy and its Executive Committee; and other items of interest to Academy members. The *Journal* appears four times a year (March, June, September, and December)—the December issue contains a directory of the Academy membership.

Subscription Rates

Members, fellows, and life members in good standing receive the *Journal* without charge. Subscriptions are available on a calendar year basis only, payable in advance. Payment must be made in U.S. currency at the following rates:

U.S. and Canada $19.00
Foreign..... 22.00
Single Copy Price..... 7.50

Back Issues

Obtainable from the Academy office (address at bottom of opposite column): **Proceedings:** Vols. 1–13 (1898–1910) **Index:** To Vols. 1–13 of the *Proceedings* and Vols. 1–40 of the *Journal* **Journal:** Back issues, volumes, and sets (Vols. 1–75 1911–1985) and all current issues.

Claims for Missing Numbers

Claims will not be allowed if received more than 60 days after date of mailing plus time normally required for postal delivery and claim. No claims will be allowed because of failure to notify the Academy of a change in address.

Change of Address

Address changes should be sent promptly to the Academy office. Such notification should show both old and new addresses and zip number.

Published quarterly in March, June, September, and December of each year by the Washington Academy of Sciences, 1101 N. Highland St., Arlington, Va. 22201. Second class postage paid at Arlington, Va. and additional mailing offices.

Editor's Note

The Washington Statistical Society (WSS) was founded in 1926 and joined the American Statistical Association (ASA) in 1935. The American Statistical Association, founded in Boston in 1839 at No. 15 Cornhill, will celebrate its sesquicentennial in 1989. The ASA was founded because of concerns with the inadequate and inaccurate national statistics. An excellent account of the early history of the American Statistical Association can be found in its 1918 publication *The History of Statistics, Their Development and Progress in Many Countries*, collected and edited by John Koren and published for the American Statistical Association by the Macmillan Company of New York in 1918. John Koren recounts some of the early history of the ASA (1839–1914) and its concern with issues of national statistics such as the Census and vital statistics, interests clearly consistent with the provisions of their bylaws which state (Koren, pp. 4–5) that "the operation of this Association shall principally be directed to the statistics of the United States; and they shall be as general and as extensive as possible and not confined to any particular part of the country . . ." I also highly recommend *Revolution in United States Government Statistics 1926–1976*, for an overview of statistical issues as they evolved with the Federal Statistical System. It was prepared by Joseph W. Duncan and William C. Shelton and issued October 1978 by the U.S. Department of Commerce, Office of Federal Statistical Policy and Standards.

With the forthcoming sesquicentennial of the ASA, many special publications are anticipated. My purpose in this brief note is to let the reader know of the roots of the statistical profession in the U.S. The articles in this special issue demonstrate in a very limited way the range and richness of the science of statistical application and methodology.

This special issue of the WAS Journal is devoted to a collection of articles by members of the Washington Statistical Society, an affiliate of the Washington Academy of Sciences. This special issue appears as the American Statistical Association enters its 150th year.

The articles in this issue convey the broad scope of statistical activities. The lead article by Richard Allen, President of the Washington Statistical Society provides an overview of WSS activities and its history.

The article by Miles Davis and Donald Wolfe applies the classical statistical methodology of principal components for multivariate data to the voting records of the U.S. Senate. Their insights and graphs will capture your imagination and encourage you to ask of the data your own thought provoking questions.

Edward Wegman's article examines the implications of computational statistics on statistical theory, education, and practice. This article provides an elegant description of some recent advances in graphical techniques for exploring multivariate data of many dimensions.

Keith Eberhart's article on ignition propensities of cigarettes follows in a long tradition of the National Bureau of Standards (NBS) recently renamed the National Institute of Standards and Technology (NIST). Traditionally, Bureau statisticians work closely with their colleagues to design and analyze statistical studies. The tradition at

NBS is long and includes such notaries as W. J. Youden, John Mandel, Joseph Cameron, Churchill Eisenhart, Harry Ku, Mary Natrella, Joan Rosenblatt and many others who have had the good fortune to work there. The article by Joan Rosenblatt in the *Encyclopedia of Statistics* (Volume 6, pp. 148–155, John Wiley and Sons, New York) is a valuable reference to this tradition.

The article by N. Philip Ross and Gilah Langner examines the tradition of statistical agencies in government and argues the case for credible centralized environmental statistics for the U.S.

My own article addresses some methodological issues in medical statistics and discusses some uses of administrative data and its augmentation for special studies.

It has been my pleasure to bring these articles together for this special issue. I hope the Washington Academy regularly will feature articles devoted to statistical science and its applications.

Journal of the Washington Academy of Sciences,
Volume 78, Number 4, Pages 291–295, December 1988

The Washington Statistical Society and its History

Richard Allen

President, Washington Statistical Society

ABSTRACT

The Washington Statistical Society (WSS) serves 1500 statisticians in this Washington, D.C. based chapter of the American Statistical Association (ASA). Each year, WSS organizes and sponsors 35–40 technical sessions, a tradition that began when WSS was organized in 1926. Accomplishments over the years are described.

About the Washington Statistical Society

The Washington Statistical Society (WSS) is the Washington, D.C. based chapter of the American Statistical Association (ASA), the leading professional society for statisticians in the world. The WSS normally has about 1,500 members which is approximately 13 percent of the members of all chapters of ASA.

While one of the strengths of the WSS is the contributions of members from the Federal statistical agencies, its membership is broad based. Approximately 55 percent of members are from government, 10 percent from academia, 30 percent from private and nonprofit organizations, and 5 percent self-employed or retired. Members of the WSS form the backbone of many ASA activities and committees. WSS members can be found in leadership roles for nearly every ASA section, committee, and publication. For example, two of the last three presidents of the ASA have been WSS members.

One of the major goals of the WSS is to organize and sponsor a wide variety of technical sessions relating to statistics each year. These sessions are advertised through the monthly WSS Newsletter and newsletters of organizations which might be interested in a specific session. Usually 35–40 sessions are offered each year ranging from very technical sessions on work in progress to presentations of general interest from renowned professionals. Many of these sessions are cosponsored with other professional organizations and university departments. Sessions are planned by a number of program committees which currently include agriculture and natural resources, economics, methodology, public health and biostatistics, physical sciences and engineering, social and demographic statistics, computing technology, and quality assurance. In addition to the regular program sessions, the WSS has organized and presented a few short

Correspondence should be sent to: Richard Allen, President WSS, c/o USDA/NASS/DAP Room 4133 So. Bldg., Washington, D.C. 20250.

courses in each of the recent years. These short courses have brought in leading speakers and video tape presentations on developing statistical methodologies. Attendance has been upwards to 200 participants for some of these short courses.

One special activity of the WSS Newsletter is a monthly employment column for both employers and prospective employees. The society participates in Washington, D.C. area science fairs by judging all projects for noteworthy statistical content. Prizes and other recognition are provided. The society also sponsors an annual award at all local universities for outstanding graduate students with interests in statistics. This award consists of a year's membership in both WSS and the ASA. The Society also cooperates with six other government agencies and professional associations in the annual Julius Shiskin memorial award for economic statistics.

Most members of WSS are also members of the ASA. However, WSS does offer associate memberships for individuals who may be professionals in related fields and who wish to keep up on information in statistical activities in the Washington area.

Below is listed a brief description of the history and development of WSS.

Brief History of the Washington Statistical Society[1]

The WSS was organized in 1926 with a short constitution that proclaimed it was to be a chapter of the American Statistical Association. Officers established were President, Vice-President, Secretary, and two Representatives-at-Large. However, a charter for ASA membershp was not applied for until November 1935.

The names of most early Society officers are lost in the mists of history. William M. Stewart, Director of the Census Bureau, was President in 1928 and in 1930. The 1930 Secretary-Treasurer was E. A. Goldenweiser.

Although the 1926 constitution set the annual dues at $1, during the Great Depression no serious effort was made to collect dues. Instead, the Secretary, who by 1935 was the ASA District Representative, stood at the meeting house door and collected 15 cents from each attendee (25 cents at luncheon meetings). Dinner meetings cost $1.25 for dinner.

Apparently, in the earliest years, WSS program meetings were held about once a month. Information on topics and speakers for meetings before 1940 is scarce but there is a reference that in a 1939 meeting titled "Irrelevant Remarks on Trivial Matters in Modern Statistical Theory," addressed by Leon Henderson, Arne Fisher and Bassett Jones, the discussion got so ribald that "the doors were closed."

The informal financial management of the Society's early days was simply inadequate, and in 1941 WSS was insolvent (a $10 deficit). The regular meeting door fee of 15 cents could not be raised because most of the meetings were held at George Washington University, where a charge greater than 15 cents would make the University liable for taxes as conducting meetings for pay. WSS discontinued door collections and ordered collection of $1 from each member at the beginning of the season. Heretofore, chapter membership had been defined as every ASA member living in the Washington area. With the increasing number of statisticians migrating to Washington because of the "People's War of Survival," WSS trumpeted to the world that the membership in 1942 was a thousand persons—approximately one third of the total membership of ASA. However, when the Board of Directors sought to collect its dues only about 200 persons responded to the call.

WSS membership grew steadily after World War II. The 1953 report of the Secretary-Treasurer stated that more than

[1]Acknowledgement: Credit goes to Al Mindlin for the original history version.

700 of the nearly 900 ASA members in the Washington area were WSS members. Membership increased particularly rapidly in the period 1958 to 1964 and continued inching up to a temporary peak of 1,500 in 1967. Membership declined in the period 1969–1974 to about 1,350. However, the trend reversed in 1975 and a probable membership peak of 1,773 was reached in 1979.

The first Annual Dinner meeting was held in 1947, with Isadore Lubin as speaker. The event has been held each year thereafter with the exception of five years.

Through the 1950's and early 1960's attention of the Society's Board of Directors focused heavily on scheduling meeting. It was mainly a program committee, leaving little time to develop additional activities. For the most part an evening meeting was held each month thru a seven or eight month season. In 1963 a special committee was established to develop over the summer the entire year's set of seven meetings. Under the leadership of T. D. Woolsey the first program committees were established in 1964—Economics under Hyman Kaitz, Public Health and Biostatistics under Oswald Sagen and Monroe Sirken, and Social and Demographic under David Kaplan. The program committee system flourished from the beginning, rapidly taking over the main burden of meetings and freeing the Board of Directors for further improvements.

As it became apparent that daytime meetings were more popular than evening meetings, more and more meetings were held during the day. In the 1969–70 season there were no evening meetings at all, except the Annual Dinner.

The 1960's saw other dramatic expansions of Society activities. In 1962 an award of one-year membership in ASA and WSS to an outstanding graduate student in each of seven universities in the Washington-Baltimore area was initiated. Also in 1962 a Methodology Committee was established under Jerome Cornfield and Seymour Geisser. Under the subsequent leadership of Samuel Greenhouse, and with an enabling revision of the constitution, this evolved by 1963 into a semi-autonomous Section of the Society with its own elected officials and separate programming. In 1966 a Baltimore Committee was established under Harold Grossman. WSS financed the growth of a statistical program in the Baltimore area, until by 1969 it felt sturdy enough to cut the parental tie and strike out as a "Maryland" chapter of ASA. In 1967 an Employment Service was established under Marie Eldridge. Also in 1967 an Outstanding Paper Award Committee for young statisticians was established under Churchill Eisenhart. In 1968 a WSS Committee on ASA Fellows was established under Margaret Martin.

The increase in Society activities during the 60's and rapidly rising costs spelled the end of the $1.00 dues in 1967 even though a number of economy measures such as switching to bulk rate newsletter mailings were taken. Dues were raised to $2.00 a year in 1967 and remained at that level for 10 years. Increases in mailing costs necessitated raises to $3.00 in 1977, $4.00 in 1979, and $6.00 in 1984 which has been maintained since then.

A major membership survey was conducted in 1969 which pointed out a generation gap. The majority of members of WSS were age 40 or above and favored nontechnical meetings. Nontechnical meetings tended to have higher attendance, but technical meetings drew more of the younger members. The WSS program attempted to provide a balance of the two types of sessions.

The 70's saw WSS attempt a number of new programs. Some of them were envisioned as annual programs but did not prove to have such permanence. In 1973 a W. J. Youden memorial scholarship was established in conjunction with the American Society for Quality Control. The program was to provide scholarship assistance for a worthy student at the Washington Technical Institute. Appar-

ently the scholarship was awarded only in 1973 and in 1976.

In 1974 WSS participated with a dozen other private and government sponsors in a three-day symposium on Statistics and the Environment. This symposium followed other similar presentations in California and was very successful, with over 200 participants. There was considerable interest in establishing such a symposium as an annual event but it was not to be.

In 1976, to commemorate the 50th anniversary of the Society, five past presidents of the American Statistical Association addressed the WSS annual dinner on the topic of "The Past as Prologue to the Future." This proved to be one of the most interesting annual dinner meetings of all time.

Another notable accomplishment of the 1975–76 Society activities was the establishment of a "local associate membership" program. That feature allows a non-ASA member such as a retired statistician or an individual working in some other field to receive the WSS newsletter and keep abreast of statistical activities in the Washington, D.C. area.

The 1977–78 program year was marked by two very popular activities. A short course on variance estimation was planned which ran on six consecutive Fridays. Over 140 people applied for the 50 available spots and a random process was used to select a lucky 50 participants. A reception was held for visiting statisticians from Latin America. Since the room would hold only 135 people over 60 individuals had to be turned away.

A major new award program was started during the 1978–79 program year, the Julius Shiskin Memorial Award for Economic Statistics. WSS is joined in this program by the Bureau of Labor Statistics, the Bureau of the Census, the National Association of Business Economists, the Bureau of Economic Analysis, the National Bureau of Economic Research, and the Office of Management and Budget, all of which Julius Shiskin was associated with during his career. This award has been presented annually with the selection made by a committee of representatives from each participating organization.

The fundraising activities for the Shiskin Award led WSS to separate the two positions of secretary and treasurer, an arrangement which had been provided for in the WSS Constitutions. Other changes over time in the structure of the elected positions of the Society was the establishment of the Vice President as the President Elect in 1966 and the establishment of two-year terms for Representatives at Large, Secretary, and Treasurer in 1984. The emphasis on maintaining continuity of WSS activities has been reflected in the encouragement of two-person teams for organizing various program sessions.

The thawing of relations with China led to one of the most successful activities of the Society. Five statisticians from the National Statistical Society of China visited Washington, D.C. for nearly a full week in conjunction with their visit to the 1981 ASA annual meetings. WSS served as the host for the D.C. part of the trip. Activities arranged included visits with statistical organizations, a reception at the National Academy of Sciences, an evening "pot luck" dinner, chaperoned sightseeing, and a most successful evening dinner in honor of the visitors.

The success of China delegation visit led to the sponsoring of other special presentations by the Society. In the spring of 1982, an all day program was scheduled to commemorate the 200th anniversary of the birth of S. D. Poisson. That fall, W. E. Deming's birthday was celebrated in an evening social session. Morris Hansen's 73rd birthday was recognized in 1983 with a similar evening social.

The 1984–86 time period led to several important developments in the operations of the Society. The interest in the Deming and Hansen birthday parties was institutionalized into an annual WSS holiday party. This early evening event held every December provides a second social activity to accompany the June annual dinner.

During the 1984–85 program year, Terry Ireland urged WSS to try some WSS sponsored short courses. These were not to compete with universities or private vendors but were intended to provide instruction or knowledge on statistical topics of broad interest. A short course might involve the use of video tapes from an ASA tutorial, with a knowledgeable individual available as a resource person. These short courses were to be provided on a fee basis, with fees covering rentals, printing of materials, and travel expenses of speakers.

The short courses were an instant success. All had been well researched and participants felt they got a fair return for their investment. From the simple beginnings of low key one-day sessions in the Martin Luther King Library the complexity of courses and arrangements increased. The 1987–88 program year was marked by a very ambitious two-day symposium on quality assurance in the government. This proved to be a landmark session bringing together a unique combination of quality professionals and agency staff members. Very professional materials were available and registration included meals and receptions for speakers. This session, planned for 150 participants, drew over 200 people.

Another key development during the 1984–86 time period was the beginning of WSS' involvement in judging local science fairs in the metropolitan area. Susan Ellenberg, in her position as a Representative-at-Large to the WSS Board, initiated contact of all local school systems and found volunteers to do the judging. Since no statistics category was offered, WSS members judge all projects for statistical content. Prizes have included books on statistics appropriate to high school age individuals.

The success of the short course program led to another change in WSS program philosophy. WSS normally does not pay any honoria for presentations. Many of the name speakers each year from outside the Washington, D.C. area have spoken for free when they are in the area for other commitments. In the 1986–87 program year, the concept of one invited lecture was originated. Based on the fact that some extra proceeds were available from short course operations, one outside speaker a year would be brought in. The concept was successful in 1987 and was repeated in 1988. In that case, the science fair winners were invited to display their projects at the invited lecturer.

In 1984, WSS started its own internal award program. This Presidents' Award goes annually to a member or members for outstanding contributions to the Society. The award carries no monetary value but since it is presented at the Annual Dinner, it does provide good recognition to the recipient.

Many WSS members are actively involved in the celebration of the 150th anniversary of the American Statistical Society. This celebration will last from August 1988 through December of 1989. WSS has planned a series of 10 monthly special presentations on broad topics such as statistics and the law and statistics and the media.

Journal of the Washington Academy of Sciences,
Volume 78, Number 4, Pages 296–310, December 1988

Clustering of Senators and Their Votes

Miles Davis and Donald Wolfe

Loyola College, Baltimore, Maryland

ABSTRACT

Votes by Senators of the United States of America on major bills are analyzed by principal components and clustering techniques. Clusters of Senators and bills are shown in graphs. The three classes of Senators elected in successive Senatorial elections are studied to detect systematic differences in their voting patterns. A dimension of liberal-conservative gradation and a pattern of change over time are found from the voting records.

Introduction

Voting in the Senate of the United States of America is an interesting and important source of complex data. Although it is highly structured by party affiliation, by ideology and by regional interest, it is still sufficiently unpredictable to be interesting. We analyze voting on key bills by all senators from 1969 to 1986, using principal component analysis. We offer the data as a readily understood and important body of illustrative but real data for experimentation with statistical methods.

Voting Data

We abstracted data from the complete record of voting in the Senate appearing regularly in the *Congressional Quarterly*[2] for key bills as selected by Michael Barone in the *Almanac of American Politics*.[1] We identify the 144 key bills in Table 1. They are selected to reflect the decisive action on issues that often come to several votes. They are thus more closely contested than the entire record of votes and represent the high drama of decision.

The actors in the drama are the 217 senators who served from 1969 to 1986. Their classes, party affiliations, states, names, and the bills on which they voted appear in Table 2. The term "class" refers to the three divisions of the Senate determined by the year of election. Since senators are elected for six year terms, and one-third of the Senate is elected every two years, each Senator holds a seat in one of the three classes.

The Senators' votes on the key bills appear in Table 3 as the data for our study. To show the votes compactly, the rows of Table 3 are identified by class and state, and the columns are identified by bill number. For example, the first row of Ta-

Correspondence should be sent to: Dr. Miles Davis, 1214 Bolton Street, Baltimore, Maryland 21217-4111.

Table 1

	Senate Bills
1○69/54	Prevent funds for Safeguard.
2○69/142	Reduce depletion allowance on oil & gas.
3●69/245	Reject Philadelphia Plan amendment.
4○70/11	Drug Control, striking "no-knock" provision.
5○70/19	Drug Control. Hughes amendment reducing further marijuana penalties.
6○70/89	Voting Rights Act—Voting at 18.
7●70/103	Corundum disposal.
8○70/112	Carswell Nomination to Supreme Court.
9●70/157	School bus in desegregated districts.
10●70/180	Bar U.S. military in Cambodia.
11●70/195	Consumer Products Warranty and Guaranty Act.
12●70/206	Limit agricultural subsidy to $20,000 to any producer in a single year.
13○70/211	Bar price support for tobacco.
14○70/240	Require weapons systems be tested.
15○70/249	Increase funds for Bureau of Prisons.
16○70/251	Create volunteer army.
17○70/252	Prohibit military use of defoliants.
18○70/256	Reduce authorizations for defense.
19○70/328	Elect Senate committee chairmen.
20○70/380	Reduce defense public information.
21○71/23	Restore funds for supersonic transport
22●71/354	Presidential campaign fund from tax.
23○71/355	Extend to single persons tax rates applicable to married persons.
24○71/361	Limit US military in Europe to 250,000
25●71/417	Rehnquist Nomination to Supreme Court.
26●72/54	Bar school busing on basis of race.
27●72/97	Table National Voter Registration Act.
28●72/110	Equal Rights Amendment.
29○72/144	Unfair Billing Practices.
30○72/215	Delete criminal penalties in 1972 Food, Drug and Consumer Product Act.
31○72/219	One-year authorization for Corporation for Public Broadcasting.
32○72/226	Delete National Legal Services Corporation from Economic Opportunity Amendments.
33○72/262	Delete minimum wage for domestics.
34○72/292	Give states exclusive authority to manage fish and wildlife, unless endangered species.
35○72/296	Reduce annual crop subsidy maximum.
36●72/334	Refer no-fault auto insurance to Judiciary Committee.
37●72/339	Outlaw Saturday-night special guns.
38●72/383	Table reduction of exemption on preference income for minimum tax.
39●72/391	General Revenue Sharing.
40○73/27	Senate committee meetings closed only by public vote at start.
41●73/34	Highway Trust Fund for bus or rail.
42●73/154	Prohibit U.S. combat in Cambodia or Laos.
43●73/286	Permit Alaska Pipeline.
44○73/372	Restrict limousines for officials.
45○73/400	Reduce military headquarters overseas.
46○73/551	New northeast rail labor agreements.
47●73/571	Halt import of Rhodesian chrome.
48●74/66	Table required registration of handguns.
49●74/69	New standards for death penalty.
50●74/138	Federal Election Campaign Financing.
51●74/156	No-Fault Auto Insurance.
52○74/187	Prohibit required student testing.
53○74/194	Table prohibition of school busing.
54●74/212	Freedom of Information Amendment.
55○74/225	Demobilize 76,000 U.S. oversea military.
56●74/286	Repeal no-knock provision of Drug Abuse Control Act of 1970.

Table 1.—Continued

Senate Bills	
57•74/395	Consumer Protection Agency.
58○74/479	Prohibit food stamps for strikers.
59•74/496	Foreign Aid Authorization.
60•75/55	Amend Cloture Rule.
61•75/67	Rescind F-111 fighter-bombers.
62•75/130	Table barring Medicaid abortions.
63•75/190	Resume military aid to Turkey.
64•75/382	"Redlining" Disclosure.
65○75/431	Emergency Natural Gas.
66•75/516	Common-Site Picketing.
67•76/27	Prohibit arms sales to Chile.
68•76/65	Federal Employees' Political Activities.
69○76/93	Prohibit federal supersonic air funds.
70•76/103	Recommit no-fault auto insurance.
71○76/141	Reduce national defense budget.
72○76/180	Bar production of B-1 bomber.
73○76/333	Utilities pay for Clinch River reactor.
74○76/471	Retain nitrogen oxide standards.
75•76/521	Delete House ban on abortion funding.
76○76/554	Water Pollution Control Amendment.
77•77/11	Presidential Pardon for Draft Resisters.
78•77/41	Warnke SALT Nomination.
79•77/42	Warnke as Director of Arms Control and Disarmament Agency.
80•77/59	Halt Rhodesian Chrome Imports.
81○77/164	Reduce target price for wheat.
82○77/232	User fees for water resources.
83•77/263	Prohibit federal funds for abortion.
84○77/275	Limit Clinch River reactor spending.
85○77/280	Prohibit production of neutron bomb.
86○77/320	Federal tax fund for Senate elections.
87•77/523	Natural Gas Pricing.
88•78/66	Ratify Panama Canal Treaty.
89○78/161	Disapprove mideast figher plane sales.
90○78/166	Amend National Labor Relations Act.
91○78/435	Extend ERA ratification deadline.
92○78/447	Revenue Act of 1978.
93○78/480	Medicare-Medicaid Cost Containment.
94•79/70	Establish Department of Education.
95○79/169	Defer new nuclear power plants.
96○79/206	Food Stamps.
97•79/438	Windfall Profits Tax.
98○79/445	Indexing individual income tax.
99•79/490	Chrysler Loan Guarantees.
100○80/60	Spending Limits.
101•80/101	Fiscal 1981 Budget Targets.
102•80/197	Draft Registration Funding.
103•80/272	Aid to Nicaragua.
104•80/315	Exempt small business from OSHA.
105•80/345	Invoke cloture on Alaska Lands.
106○80/441	Kill federal funds for abortion.
107○80/466	Cut MX missile funds.
108○80/496	Fair Housing Act Amendments.
109○81/140	Require paying for food stamps.
110•81/182	Budget reconciliation.
111•81/239	Cut individual income tax rates.
112○81/275	Helms (R-NC) Foreign Aid Amendment.
113○81/335	Disapprove AWACS sale.

Table 1.—Continued

Senate Bills	
114•81/368	Legal Services Corporation.
115•82/2	Bar court-ordered school busing.
116•82/118	Chemical weapons.
117•82/288	Balanced Budget/Tax Limitation Amendment to the Constitution.
118•82/420	Bar MX missile procurement.
119•82/422	Eliminate Clinch River reactor funds.
120•82/463	1982 Transportation Assistance Act.
121○83/35	Social Security Disability.
122○83/65	Postpone date for resident status.
123•83/101	Immigration Reform and Control Act.
124○83/169	Human Life Federalism Amendment.
125○83/170	Tax Rate Equity Act.
126•83/293	Martin Luther King, Jr. Holiday.
127•83/355	Table tuition tax credits.
128○84/34	School Prayer Amendment.
129•84/51	Table combat troops in El Salvador.
130•84/132	Table ban on new MX missiles.
131○84/252	Prohibit activities against Nicaragua.
132•84/266	Table freeze on nuclear weapons.
133•85/142	Firearm Owners' Protection.
134•85/191	Immigration Reform and Control Act.
135•85/300	Textile Import Quotas.
136•85/310	Discount sales of tobacco.
137○85/371	Sequester funds for national defense.
138•86/51	Aid to Nicaraguan "contras".
139•86/176	Table limit to "star wars".
140•86/209	Reduce limits on PAC's.
141•86/266	Rehnquist Nomination as Supreme Court Chief Justice.
142•86/296	Tax Overhaul.
143○86/300	Omnibus Drug Bill.
144•86/311	South Africa Sanctions.

ble 3 shows votes by the senators occupying the seat reserved for Arizona in Class I. Reference to Table 2 shows that this seat was held by Senator Fannin (R, AZ) while voting on bills 1 through 76 and by Senator DeConcini (D, AZ) for bills 77 through 144.

Votes are recorded as letters or symbols:

y = yea vote
n = nay vote
# = paired for
x = paired against
f = CQ poll for
• = CQ poll against
+ = announced for
− = announced against
p = voting present
c = not voting to avoid conflict of interest
? = unknown
space means not a senator for that vote.

Bills are identified in Table 1. Our sequential bill numbers are followed by the years of the *Congressional Quarterly* (CQ) volumes and the numbers assigned by CQ to the bills. Bills marked • were passed, but those marked ○ were rejected. Short descriptions of the bills follow the numbers. Much more thorough descriptions are in CQ.

Analysis of the Data

We began to analyze the data by concentrating on the process of voting. We formed a matrix of 217 rows and 144 columns containing a +1 wherever a yea vote (y) was cast and a −1 wherever a nay vote (n) was cast. Elsewhere, the ma-

Table 2

Senators listed by class, party, state and name, with the numbers of the earliest and latest bills in their terms of office.

1	R AZ	Fannin	1	76
1	D AZ	DeConcini	77	144
1	R CA	Murphy	1	20
1	D CA	Tunney	21	76
1	R CA	Hayakawa	77	120
1	R CA	Wilson	121	144
1	D CT	Dodd, T.	1	20
1	R CT	Weicker	21	144
1	R DE	Williams, J.	1	20
1	R DE	Roth	21	144
1	D FL	Holland	1	20
1	D FL	Chiles	21	144
1	R HI	Fong	1	76
1	D HI	Matsunaga	77	144
1	D IN	Hartke	1	76
1	R IN	Lugar	77	144
1	D ME	Muskie	1	101
1	D ME	Mitchell	102	144
1	D MD	Tydings	1	20
1	R MD	Beall	21	76
1	D MD	Sarbanes	77	144
1	D MA	Kennedy	1	144
1	D MI	Hart	1	76
1	D MI	Riegle	77	144
1	D MN	McCarthy	1	20
1	D MN	Humphrey, H.	21	87
1	D MN	Humphrey, M.	88	93
1	R MN	Durenberger	94	144
1	D MS	Stennis	1	144
1	D MO	Symington	1	76
1	R MO	Danforth	77	144
1	D MT	Mansfield	1	76
1	D MT	Melcher	77	144
1	R NE	Hruska	1	76
1	D NE	Zorinsky	77	144
1	D NV	Cannon	1	120
1	R NV	Hecht	121	144
1	D NJ	Williams, H.	1	115
1	R NJ	Brady	116	120
1	D NJ	Lautenberg	121	144
1	D NM	Montoya	1	76
1	R NM	Schmitt	77	120
1	D NM	Bingaman	121	144
1	R NY	Goodell	1	20
1	R NY	Buckley	21	76
1	D NY	Moynihan	77	144
1	D ND	Burdick	1	144
1	R OH	Young, S.	1	20
1	R OH	Taft, Jr.	21	76
1	D OH	Metzenbaum(II)	77	144
1	R PA	Scott, H.	1	76
1	R PA	Heinz	77	144
1	D RI	Pastore	1	76
1	R RI	Chafee	77	144
1	D TN	Gore, A., Sr.	1	20
1	R TN	Brock	21	76
1	D TN	Sasser	77	144
1	D TX	Yarborough	1	20
1	D TX	Bentsen	21	144
1	D UT	Moss	1	76
1	R UT	Hatch	77	144
1	R VT	Prouty	1	21
1	R VT	Stafford	22	144
1	D VA	Byrd, H., Jr.	1	120
1	R VA	Trible	121	144
1	D WA	Jackson	1	125
1	R WA	Evans	126	144
1	D WV	Byrd, R.	1	144
1	D WI	Proxmire	1	144
1	D WY	McGee	1	76
1	R WY	Wallop	77	144
2	D AL	Sparkman	1	93
2	D AL	Heflin	94	144
2	R AK	Stevens	1	144
2	D AR	McClellan	1	87
2	D AR	Hodges	88	93
2	D AR	Pryor	94	144
2	R CO	Allott	1	39
2	R CO	Haskell	40	93
2	R CO	Armstrong	94	144
2	R DE	Boggs	1	39
2	D DE	Biden	40	144
2	D GA	Russell	1	20
2	D GA	Gambrell	21	39
2	D GA	Nunn	40	144
2	R ID	Jordan	1	39
2	R ID	McClure	40	144
2	R IL	Percy	1	132
2	D IL	Simon	133	144
2	R IA	Miller	1	39
2	D IA	Clark	40	93
2	R IA	Jepsen	94	132
2	D IA	Harkin	133	144
2	R KS	Pearson	1	93
2	R KS	Kassebaum	94	144
2	R KY	Cooper	1	39
2	D KY	Huddleston	40	132
2	R KY	McConnell	133	144
2	D LA	Ellender	1	35
2	D LA	Edwards, E.	36	39
2	D LA	Johnston	40	144
2	R ME	Smith, M. C.	1	39
2	D ME	Hathaway	40	93
2	R ME	Cohen	94	144
2	R MA	Brooke	1	93
2	D MA	Tsongas	94	132
2	D MA	Kerry	133	144
2	R MI	Griffin	1	93
2	D MI	Levin	94	144
2	D MN	Mondale	1	76
2	D MN	Anderson, W.	77	93
2	R MN	Boschwitz	94	144
2	D MS	Eastland	1	93
2	R MS	Cochran	94	144
2	D MT	Metcalf	1	87
2	D MT	Hatfield, P.	88	93
2	D MT	Baucus	94	144
2	R NE	Curtis	1	93

Table 2.—Continued

2	D NE	Exon	94	144
2	D NH	McIntyre	1	93
2	R NH	Humphrey, G.	94	144
2	R NJ	Case	1	93
2	D NJ	Bradley	94	144
2	D NM	Anderson, C.	1	39
2	R NM	Domenici	40	144
2	D NC	Jordan	1	39
2	R NC	Helms	40	144
2	D OK	Harris	1	39
2	R OK	Bartlett, D.	40	93
2	D OK	Boren	94	144
2	R OR	Hatfield, M.	1	144
2	D RI	Pell	1	144
2	R SC	Thurmond	1	144
2	R SD	Mundt	1	39
2	R SD	Abourezk	40	93
2	R SD	Pressler	94	144
2	R TN	Baker	1	132
2	D TN	Gore, A., Jr.	133	144
2	R TX	Tower	1	132
2	R TX	Gramm	133	144
2	D VA	Spong	1	39
2	R VA	Scott, W.	40	93
2	R VA	Warner	94	144
2	D WV	Randolph	1	132
2	D WV	Rockefeller	133	144
2	R WY	Hansen	1	93
2	R WY	Simpson	94	144
3	D AL	Allen, J.	1	89
3	D AL	Allen, M.	90	93
3	D AL	Stewart	94	108
3	R AL	Denton	109	144
3	D AK	Gravel	1	108
3	R AK	Murkowski	109	144
3	R AZ	Goldwater	1	144
3	D AR	Fulbright	1	59
3	D AR	Bumpers	60	144
3	D CA	Cranston	1	144
3	R CO	Dominick	1	59
3	D CO	Hart	60	144
3	D CT	Ribicoff	1	108
3	D CT	Dodd, C.	109	144
3	R FL	Gurney	1	59
3	D FL	Stone	60	108
3	R FL	Hawkins	109	144
3	D GA	Talmadge	1	108
3	R GA	Mattingly	109	144
3	D HI	Inouye	1	144
3	D ID	Church	1	108
3	R ID	Symms	109	144
3	R IL	Dirksen	1	1
3	R IL	Smith, R. T.	2	19
3	D IL	Stevenson	20	108
3	D IL	Dixon	109	144
3	D IN	Bayh	1	108
3	R IN	Quayle	109	144
3	D IA	Hughes	1	59
3	D IA	Culver	60	108
3	R IA	Grassley	109	144
3	R KS	Dole	1	144
3	D KY	Cook	1	59
3	D KY	Ford	60	144
3	D LA	Long	1	144
3	R MD	Mathias	1	144
3	D MO	Eagleton	1	144
3	D NV	Bible	1	59
3	R NV	Laxalt	60	144
3	R NH	Cotton(I)	1	59
3	R NH	Cotton(II)	64	64
3	D NH	Durkin	65	108
3	R NH	Rudman	109	144
3	R NY	Javits	1	108
3	R NY	D'Amato	109	144
3	D NC	Ervin	1	59
3	D NC	Morgan	60	108
3	R NC	East	109	138
3	R NC	Broyhill	139	144
3	R ND	Young, M.	1	108
3	R ND	Andrews	109	144
3	R OH	Saxbe	1	47
3	D OH	Metzenbaum(I)	48	59
3	D OH	Glenn	60	144
3	R OK	Bellmon	1	108
3	R OK	Nickles	109	144
3	R OR	Packwood	1	144
3	R PA	Schweiker	1	108
3	R PA	Specter	109	144
3	D SC	Hollings	1	144
3	D SD	McGovern	1	108
3	R SD	Abdnor	109	144
3	R UT	Bennett	1	59
3	R UT	Garn	60	144
3	R VT	Aiken	1	59
3	D VT	Leahy	60	144
3	D WA	Magnuson	1	108
3	R WA	Gorton	109	144
3	D WI	Nelson	1	108
3	R WI	Kasten	109	144

trix contains zeroes. This matrix was condensed into Table 3 by placing the votes of all Senators into rows corresponding to the Senate seats that they hold, identified by class and state. Rows in the data matrix, unlike Table 3, correspond to senators. The columns of the data matrix correspond to bills, as in Table 3. Gradations of support short of voting are ignored by this choice, and an alternative analysis might well take them into account.

A spectral analysis of the matrix was done, following the ideas of Good[8]. The

TABLE 3.1

Votes in the Senate. Senate seats arranged by class and state identify the rows, and senate bill numbers identify the columns.

```
              111111111122222222223333333333444444444455555555556666666666777
     123456789012345678901234567890123456789012345678901234567890123456789012
1 AZ nnnnnyyy?n?n?xnyn-?nynnnyyynnyyyyynyny??nnyynynyynnyn?nnnyn?nnynnnnnnynn
1 CA nn?n-yyyxnnx+n?#--xnnynnn-nyy?nnnnnnyny+yyn.nnynnyyynyyyyynyyyynyyyy+y??x
1 CT nnynn-+n+-++-n+nn-??nnynynyyn?nnyn?nnyyyy#ynxnyyynnyyyy?yynnnnynynyyynynyn
1 DE nynnnnyynnnyyynynynynnnnyyyynyyyynynyyyyynnnyyyyynyynynyyyynnynnnn?nnnyn?
1 FL nnnynynynnnnnxnnnnnnnyynyy-#y???nnyyynnyyynn?nnyyynynynnyynnyyynynyyynynn
1 HI nn?nnn#nyn+nn-ny?-?nynynyyyynynyynnyyyyny+y?nyy???cynynyynynyyynnn.nnc.n
1 IN yyy??nyn?yy???yy?+?ynyy+n.nyynnnnnynyyynnyynynyyynynnn?yyyn?y?nn?yyyy?y?y
1 ME yyyyynnn#yyyn?ynyy?yny??n-?y????nn?ny?yyy?nnnnynnyyynyyyyynyyyyyyyyy?ynny
1 MD yyyn+nnn?yy+n+yn-#??ynnnyyyynyyyynynyyyyynyynynyyyyynynnynyyynynynnynnynn
1 MA yyyy?n?nyyyyn??nny#ynynynnnyynnnnnyny.fyyynnfnynnyyynyyyy+nyyyynfyyyyynyy
1 MI yyyyyn?n+yyy#+yyyynynyyynnnyfnnnnn#nynyyyynnfnynnyynyyyyyynyyfyyfyyfynny?
1 MN yy?yy--n+y++-yyn+y-ynyynnx-y?nnnnn#nyxyynynnynynnyynyyyyyyny+yy.+yyyyxnyy
1 MS nnnynynynnynnnnnnnnnynnnyyynnynyyynnynynxx??yn?nyyxnynnnnnyn?nnynnnnnnynn
1 MO y#?nnxnnyyy--??yny-nyyyyynnyynnnnnnnynnnyynny?yny+y?yy#yynnynyyynyyyxynn
1 MT yn#yynnn#yyyx?y#yyyynynyxnny????nnnnnnnxnynnyny??y#yyyynynny?yyyyyyxxnny
1 NE nnnnny+ynnnnnnnnnnnnynnnyyyynyyyyynyyyynnnyy.ynyynnynnnnny.nnnynnnn-nynn
1 NV ynnnnnnnny?y#x?x?x?nyynnyyynyynnnyfnynnynn+ynnnnyyyynnnnnynnyn?yyynyynynn
1 NJ yyyyynnnyy+yyy?yyyyynyyynxnyyn.nn.ynyyy+#ynnynynnyynyyyyyynyyyy-yyyyyynyy
1 NM ynynyn?nyyynn?????ynynyynnynnnnnnnynyynnynnyn?yyynyyyyyyynnfynnynyy?yyn?
1 NY yyynynyn#yyyy#yyyy##ynnnyyynnyyyy-nyyyynynnyfynyyynnyn?nnn??n?nynn?nnyynx
1 ND ynnynnynyyynnnyy-ynynyyyynnyynnnyynnyynnnynnfnyynyynyyyyynnyyynynyyyyynnn
1 OH yy?yynnn+yy++y-yyynyyynnny+yyny.xyn#nnynyyny??ynyynyyyyynynny#f+yynynnnyn-
1 PA nnynnyxy#nnynn+ynnnnynnnynyy.yynyy##yyynynyyynyyyyyyyynnnynyyy+yynnnn-nnn
1 RI nyyyynynyy+yyy-yny-nnynyynny+nnnnn#nyyynyyynnnnynnyyyy+yyynnyynnyyy+#yny#
1 TN yynynnnnny??x??n???yynnnyyyynyyyyyn?yyyyyny..ynyynnynnnynynnnyynnnnn.nn.
1 TX ynx#nnnn?y?nnn-ynnnynynnyyynynnnnnnnynynynyyn?nyyyyn?nnnyyyyynynnnnn?nyn?
1 UT ynyy-n+nyyy+y-yyy++yyynynnnyynnnnnnnnnyyyyn..nyyyyy.yyfyynyyyyxfyyyy.yn.
1 VT nyy??n?nyn?yn-nynn-nnnnyynyyyynnnn#yn+yyyynynnyyyyynyynyynyynyyyyynxnyyn
1 VA nnnyyyyynnyyn?nnnnnnnnnnyyyynyyyynyyyynnnyynnynyyxnynnnynynnynnnnnnnyynn
1 WA nyynynnnyyyyyn+nnnnnyyynn--++nnnnnnnynynynnnnnyyyyynyynyynyynynyyy+y-xyn
1 WV nynnxnyynyyynnnnnynyyynyyynyynnynnnnyynnnyyynynnyynyyynnynynnynyyynnynynnny
1 WI yynynnynyyyynynyyynynyyyyynyynnnnnynynnyyynnynyyynyyyynyyyynnyynnyyyynynny
1 WY nn?ynnnnyny-??ynnn?nyyynynnyynn?nynnxy+nn?yynnyyy+ynyyxyynynnn#ynyny?n??
2 AL nnnyyy+ynny--nnnnn?nyynny?yy.nnyynnyyf+nnxyynnnyffn.nf.ynnynnnyynnyynyn.
2 AK nn?+nnnyyyy??--#-nynynynynyyfynnnynnnyynnyyynnyyyyyynnynyyynyyny++nyn#nn.-
2 AR nnnnny+ynnnnnnnnnnnnnynyyy#?#???nynnynyynnyyynnnyynnynnnynnyn?nnynnnnnn#?n
2 CO nnxnny#ynnxynnnnnnnnnynynyyyyxyyynyxyy+#yy+nyfnyyyyynnyyyyn?yyy-yyyyyynyy
2 DE nyynnnyyynnynnnynnnnynnnynyynyyyynyynyyyyyynnynfynyynyyyyy?nyynnyyyyy#nyy
2 GA ny?nn#?yx-????nnnn?nynyyyyy++--y-x+#ynnyyynnnnyynnnnnnnnynnyyynnnnynynn
2 ID nnnnnn+ynnnynnnynnnnnnnnyyyyynyyyyny#yyyx#nynnnynyynnyn?nnnynnnnynnnn?ny-n
2 IL y-yxnnxnyy###nyynnyynnnn#n-ynnnnn-ynyynyyyny#nynnyy?yynyy?yynynynyn-nnn.
2 IA nnynnyyynnnnnnnnnnnnn+nnnny.yynynyynnyy+#yyynnynyyynyyynyyyyynyyyy-yyyyyynyy
2 KS yn?nnnyyyyynnnxynynyynynynyyynn?yynnyyynyyny.yyynyyyyynyynyynyyyynnynynnn
2 KY yy?ynnyyyynnnyynnynynn?nynyyyyyynnnyyyynnyyn?nnyyynynnnyynyyyyypnnyynyny
2 LA ynxnnyyynn+nnnnnnnnnyynyyyyyy#ny+?xyyyyynyynn?yyynnynnn?nynynnyynnnynyn.
2 ME yynnnnnnnny++y-nnnnnnnnn+ynynyyynnnnyyyyyynnyyynnyynyyyyynynyy?yyyyynyyy
2 MA yyynynynyyyyynyynnynnnynnnnyynnnnnynyyyyyyyynnynnyynyynyynyyyynyyyyy?yy
2 MI n#ynnyyyyn?y#nynnnnnnnynyyyyn?y#ynyxynynynyyynyyyyynyynnnnnyyyynnynnnnx?nnn
2 MN yyyyynynyyyy??ynyyyynyyynnnyynnnnn#n#nyyyynnyny??yynyyyyynyy?y?yyyyyynyy
2 MS nn?ynynynn?nn?nnnxnnynnnyyyxnyn#ynny-yynnnyy?nny?nnynnnnn?n??n?nnnn?nynn
```

TABLE 3.2

Votes in the Senate. Senate seats arranged by class and state identify the rows, and senate bill numbers identify the columns.

```
                 1111111111222222222223333333334444444444555555555566666666677 7
     12345678901234567890123456789012345678901234567890123456789012345678901234 5 6 7 8 9 0 1 2
2 MT yy#yyx+n+ynn+y--yynynyyynn.yy..nnynnxnyxynnynynnyyyyyyyynyyy.fyyyyyynyn
2 NE nnnnnyyynnnx-nnnnnnnynnnyyyynynyyynyyyxnnnyyxynyynn+nnnnnynxnnynnnnnnnynn
2 NH yyynnn?nyyyynnnnnnnnnnyyyyn?+y?.nnnynyn+n?ynynnyynyynyyyyynyy?ynyyn?y?nnn
2 NJ yy+yynynyyyyyyyynyyyynynnnnnyynnnnnynyyyyyynnynynnyyyyyynyyyyyyyyyyyyyyy
2 NM n-?n?n??yy???nnnn?nnnyny?nnyn?n??nnyyy?nnyyynyyyynnnyyyynynnnnnnnynn
2 NC nnn#ny+ynyy-xnnnnn?n-ynnyyyy?yny??xyyyynnnynnyyynnynnnnnynnynnnynnnynn
2 OK ynyyynnnyyy??#ynyy+ynyyynn-y?????nyny??ynyyn.ynyynnynnnnnynnynynnnnnnynn
2 OR yy#nnnxnyyyyy#yyyyy#nnyyynyy#--nyy+yyynyyyn.#yy+-yynyyyyynnyyny-ynyyyny#
2 RI yyyyynyxyy+yyynnny?ynyynynnyynn?nn#n??ynyynnynynnyyny?yyynnyyynyyyyynnyy
2 SC nnnnny#ynnnnnxnnnnnnynynyyyynyyyynnyyy#xnnyynynyynnyn-nnnyynnnynnnnnnnynn
2 SD n-?x-?++--x----x--x--??????????????????#yyn?ynyynyynyyyyynn??ynyyyyfynfy
2 TN n-nnn-+ynnnnx?+ynnnnynynyyyynyy???xyn?yny?y?nynyynnynnnnynyynn??nnnnnnyn?
2 TX nn?xn+nynnxnnnnnnnnxynnnyyyynyyyyynynyynnnyynynyynnnnnnyynnyynnnnnnnynn
2 VA nynynnynnyyyny?nnnnynynnyyyyyyynnyn#ny#ynnnynny?yy?nynnynnnynnnnnnnnynn
2 WV ynnynnnnynyyynynnnyyyyynyyynyynnnnnxyyyynnyynynyy#n-ynyyynnynx.nnyyynyyn
2 WY nnnnnnyynnnnn-nnnnnnnnnnyyynnyyyyynynfynnnyynynyynnynnnnnynnnnynnnnnnynn
3 AL nnnynyyynnynnnnnnn?nnynnyyyynynyynnynyynn.ynnynyynnnnnnnynynnynn.nnnnnynn
3 AK yn##?-nnyyy??#+++#?yyyyynnny+.-xnyxnnny#nyyny??ynyy.f+yfynynyynynyyfc-?n
3 AZ n???n??y?n?n?n?yn?ynyn?nyyyn???y?yxyn?xnnxynnyyy-n?nnnnnyn?nnynnnn-nynn
3 AR yxnn?nyn?y?nyynx?ynynynyny-yynnn?nnyyyynnynnyny.??x????y????yyynynyyyynn
3 CA ynyy?nnnyyynynyyy#yynyyynnnynnnnnnnny?yyyy?nynynnyynyyyyynyyyynyyyy+ynyn
3 CO nnnx-nyyynnynnnnnnnxynnnyyyynyyyyyxynyyny+yn.yn??nnnn?nnnyyyyynyyy?yynny
3 CT yy#y#n#nyy++?#yyy#nynyyynn-yynnnnnynyyyyyynn+nynyyynyyyyy?yynyn?yyy-ynyy
3 FL nnnnn+yynnnnnxnynxnnynnnyyyynyyyynnyyyyynynn.nnyyy??nnnnn?nyynnnnnyynynn
3 GA nn?ynynynnynnnnnnnnnyynyyyyyyyynyynnyyyynn#ynnnnyynnnnnynny??yy?nnnnynynn
3 HI yyyyy-nnyy+nnn+++-yyyyyyn.nyyffnnnxnyyyxyyynynynnfy.fffyynyyn#fyyyy??n.n
3 ID ynynyn+nyyyynnnyny-ynyn+nnnyynnnnnfynnxny+nnynfyy#ny#yyyy?nnfnnyyyf#ynf+
3 IL nxyn-#yyny-yyyyyyx#ynyynynnyynnnnnynynnyyynnynynnnynyyyynyyyynyyyfynnny
3 IN yyynnnnnyyyynyyny?yy-yyyn-nynnnnnnynyny#xynnynyyyyynyyyynnyfynfy+fy#nyy
3 IA yyyyyn+nyyyynyy#yyyynyyynnnyyx.nnnynynyynyn.fnynnyynfyyynnyyyyyynyyy
3 KS nnnnnnyyyynnnn+ynnnnynynyyyy?yyyyynyyyynnnynyynnynnnyyynyynnnnnnnnnnn
3 KY yn?nnn+nynn?xnnynny?ynnnyyyynnnynynyynyynnnynyynn?nynyynyyyyynyyyyy?y
3 LA nnxnn#?yxxynnnnnn-?yyynyy#nynnnyynnnyyynnnynynyy+nnnny?nynyyn??nynnnnny
3 MD yny#?n?nyynyn?#?nyyyyynnynnyynnnnn#yyyyyyynnfyyynyynyyyyy?yyyyyy?yynny?y
3 MO yyynnnnnyyynn??nyyyynyyyynnyynnnn??yynn#?yn?ynyynynyyy?ynn??nnynyyy#y?n
3 NV nnnnnnnynyyyn-nnnnnnyynyyyny.nnnnnnynynnn+ynynnyyynnnnynnnnnnynnnn?nnynn
3 NH nn?nnnyyxnnynn?nn?nnynn?yyynnyyyyy#yyyynn+y??y???nnnnnnnnyy nyyyyyynyy
3 NY yyy++nynyy##?nynny?ynnnnnnnyynnnnnynyyynyynnynnyy?yynyy?yynynfnyyyynyy
3 NC nnxynynynnynnnnnnnnnnnnnnyyyn#ynyynnyyynxn+y.ny-yynnynnnyny.?y??xnnnyny-n
3 ND nnnnnnnynnnnnnnnnnnnnynyyyynynyn.yynyyynnnyyynnnyyynyny.nnynynnynnn.nnyn.
3 OH yn?nn??yyyyynn?x?#yyynn?y-yyyynnyyyynynn?yyy?ynnn#ynyyyyynyyyynfnnnynnnn
3 OK nnynnxnyynnnn?nnnn?nynnnynnynyyynynyy?yynyyy?ynyyn?y?nn?ny?nynynnnnnnynn
3 OR nny??nynyynyyyyy??ynnnyyy?y+?nyyynxynyy#yynynyy??yy?yy?yyynyyyy?yynnnyny
3 PA ynynnnnnyyyyyyyyynnyynnynynnyynnnnnnynnyyyyyynnnyyyyynnyyyyny+ynnyyyyyynny
3 SC n?xynnnynyynnn-nn-xyyynyyynynfnnynnnyynyynyynyy???xn?x??nyynn.yn.ynny?ynn
3 SD yyy?ynnnyyyy?y?yyyyynyy+n-n+xnn?n?+ny?#yyynnynyynynny?yyynnyyynyyyy??nyy
3 UT nnynnyn#nnnnnnnnnnnnnynnn+yynnyyyyyxynyyny.yy.y.ffnyf..nnn?yynnynnn.nnynn
3 VT yyynnnyyyynnny-nnn?nnnnyynyyyyyynnnyyyynynnnnyynnynyyyyynyyyy.yyyyy?nny
3 WA yyynynxnyyy?#nnnyxnnyyyynnny+nnnnnnnynyn#y-nnnyyyyynyyyyyn-yn+n#yyyynnnn
3 WI yy#yy-yny#?ynyyyyyyynyyynnnyynnnnnynynnyyynnyny??yyyyyyyynn?yynyyn?yynyy
```

TABLE 3.3

Votes in the Senate. Senate seats arranged by class and state identify the rows, and senate bill numbers identify the columns.

```
                                         11111111111111111111111111111111111111111111
     77777778888888888999999999900000000001111111111222222222233333333344444
     345678901234567890123456789012345678901234567890123456789012345678901234
1 AZ nnyyyyyyyynnyyyyyyynnyyyyyyynnnyyyyyyny?#ynyyynyynyyyxyyynn?#ynynyynnnyynny
1 CA ?++-nnnnynynnnyynnyyynnynynyyyynyynnnyyyyn?yyyyyynynnnynyyynyyyynyyyyyyyny
1 CT nyy.nnnyynynynyyyynn.ynnynnnynyy?yn?ynyyynyynnnnyynnynnyynnnynyyyynnnnnnnyy
1 DE ynnnnny?yynnnnynynyyyynyyyyyynyyyynnyyyyynynyyynynnynnynyyyynyyyyynnyyynynny
1 FL nnyynnyyynnnnyyyynynnynnynynyyynyynnynyyynynynyyynyyynyyyyyyyynnyyy??#ynyyynny
1 HI nnyfyyyyn?yyyynyyynnnyynynynnnyny?nynyy?yynnnnnyyyynyyynnnynnnynynnnynyyy
1 IN n?xynnnnnnnnnynnnyyynnynyyyyyyyyyynnyyyynnyyyyynynnyynynyyyynyyyynyyyyynyyny
1 ME nnynyyyyy-nyyynynynnn?ynynyn?nynyynynyynyynnnnnnynnnyyynnnynynynynnnynyyy
1 MD nn?ynyyyynyyyynyyynnnyynynynnnynynnynnnynyyn?nnnny?ynyyynnnynnyynnnnynyyy
1 MA yyynyyyynnyyyynyyyynnnyyny?y??n??ynyynnnnyynxnnn-yynnyyynnnynnnyyynnyny?y
1 MI f.ynyyyy-ny+yynyyynnnyynyyynnnynynyynny-yynnnnnnynnnyyynnnynynynnnnynyny
1 MN nyynyyyynyyynyxynynnnyynyyynyyyyyyyynynyyyyynnynnynnyynynnynnyyynnynynyyyy
1 MS nnnynnnnnynnnnynnnynyyny?nnnyyy?n?nnyyynnnyyyyyynnyyyynyyyyny?yyyyyyyyyy
1 MO ??ynynnyyynnynyyynnyyyynnyyyyynnyyynnnyyyyyyynyyyynnyynynnyynyyyynyyyyynynyy
1 MT yyy#yyyynynnyyyynyynnyyynnnynnn??nynynyynn?ynynnyyyyyyyyyyynnynyyynynnnynnyy
1 NE nnnynnnynnnnny#nynyynynynynyyyyyyynnnyynnnyyyyynynnynnnyyyyyynnnynnnynyyny
1 NV nny+nnyxynnnnynynnnynnnynyny#?ynnn-n?yny+yyyyy?nnyynnnyyynyyyynyyyyynyyynn
1 NJ yyynyyyyyyyyyyynyyynnnyynynynny?ny?yynnynyy?yyyyyyyyynyyynnnynnnynnnnyny?y
1 NM nnn?nnnnnnynnnynnnnyynnynyyyyy?ynn?nnyyynyyyyyyyyyyynnyyynnnynynnnnnnyynyny
1 NY ?nn?fnyyyyyynynyyynnnnnnynynnnyynynnnynnyny+nnnnn?nyynyynnnnynnyynnnnynyny
1 ND .nnyyyyynyynnyynyynnnnynnynnnny?nyn?ynyynyynnnynyyy?ynyyyn?nynynnnnnnnynyyy
1 OH .nyyyyyyynyyyyynyyynnnyy?ynynnnynyn?yn?yny#nnnnnnyyynyyyynnnynnnyynnnnnynyyy
1 PA nnynynyyyyyynynyyyyynyyyyynyyn?ynnyyynynyyynyynnnny?nnynyy?nyynyyyyy#yynyyny
1 RI nyxnyyyyynyyyynyynynynyynynnnynyynynnnynyynnynnnnnnynyynnyynyyynnnyynynnnyyyyy
1 TN .?#yyyyyyyynnynyyyyynn+nny?ynny?yynnynyynyyynynynynynyyyynnynyyyyyynnynnny
1 TX nnnynyyyn-ynnnyynnnynyyynynnnynyyy?ynnnnyynyyyyyyynynnnynyyyynxnyyyyyyyynyyy?y
1 UT nynnnnnn-yynnnnynnnyyyynynynyyn?yny?nyyyynnyyyyy-nnnynnnyyynyyynynyyyynynnn
1 VT nnynnyyyyny.nyyynynxn?nnyyyynyyynynnnynyyynnnnynnyn?ynnyyn?nynyyynnnnnyyy?y
1 VA nnnynnnnynnnnnynnnnynynnynnnnyyynynnnnyyyynnyyyyynynnyyynynyyyynyyyyyyyyyyyyyny
1 WA nyynnnyyynyyynnynyyyyynnynnynynyyyynynnnynyyynyyynnnyy+yyynyyyynyynyyyynyynnnyyyy
1 WV yynyyyyynynyyynynynynnnynnnynynyyyynynnnynyyynynyyyyynnnynynyyyyynyynyyyyynnyynyny
1 WI yynnyyyyynnyyyynyyyynynnyyyynyynnnyyyyyyynynynynnnnyyyynynnnynyyyynynnnyyyyy
1 WY nny+nnnnnnynnnynnnnyyyn-ynynyyyynynn?nnyyynnyyyyyn?nynnnnyyynyyynyyyyynyn?n
2 AL nnnyyyyynyyyynnnynnnynynyyyynyyynnnynnnyyyyyynyyyyynyn?ynyyyyyynyynyyyyyyyyynny
2 AK nn#ynnyynyyynnnynnnyyyyynyn?yyyyyynn?nnyyynyyyyyyynnynnnynyyynyyyyyyyyynyyynn
2 AR ?nnynynn?y???-yynnnnny?nnnnnyyyyynnynyyyny?ynynnnnnnynyyyyynny?ynyyynnnyy??y
2 CO yyy..yyynyyyyynyyyyn??nnynynyynnyny?nyyyynnyyyynnnnnynynyyyyny?nnyyyynyyynn
2 DE y?nnnyyyynn?yynyyyynynyyn+nyny?ynyy?yynynyyynnnnny?ynyyynnnynynyn?nn+nyny
2 GA nnyynnyyynyyynnnyynyyyyynyynnnyyyyynnnyyyynnnyyyynynnnyyyyyyyyynyyyyyyyyyyyynny
2 ID ?nn?nnnnnnnnnnynnnyy+ynynynyynny?y?nyyyynny?yyyyn?nynnnyyynyynnyyyynyyynn
2 IL ?nynyyyyyyyyy?#ynynyyyynyyyyyyyyyynnynyynnynnyynynnnynnnyyyy?n??nynynnyn-?y
2 IA yyynyyyyynyyyyynyyyynn?nnyn?nyynnyyynnyyyynnyyyyyyynnyynnnyyynyyyynynnnynyyy
2 KS nnyyyyyyynyynnyyyyny+nynnynynyyny?ynynnyynnnynnnnynnynnyynyyyyyyynynynyyyyny
2 KY nyyyynyyynynnnyny?yynyynnynynyyyyyyynnnyynnyynynyynnyyyynyynynyyyyyyyyynyyyny
2 LA nnnyyyyyyyyynnnnynnnyyyyynnnnynyyy?nynnnyynnyyyyyyynn?yynynyynnyyyyyynynyyyyny
2 ME ynyn.yyyynyyyynnyyynn.nynyynyynnyyn?nn?yynynynynnnnnnyyynyyyyyynynyyyyyyyy
2 MA nyynyyyyynyyyyynyyyynnny?nyyynnnynynyynnnnyynnnnnyyyynyyyn?nynnnynynnnyny?y
2 MI nnnynnnyynn?nnynnnnnyxyynynynnn?nynyynnnny+nnnnnyyynyyyynnnynnnynynnnynnny
2 MN y???yyyynyyyyynyyyn?nyyynynyyyyyyyynynyynyyynyyynynnyynyynyynyyynyyyynyyyy
2 MS nnnynyynnyn?nnynnnnynyynnnnynyyynyyynnnyyynnnynyyynnnyyynyyyyyynyyyyyyyyynyy?n
```

TABLE 3.4

Votes in the Senate. Senate seats arranged by class and state identify the rows, and senate bill numbers identify the columns.

		111111111111111111111111111111111111111111111
		777777888888888899999999990000000000111111111122222222223333333333344444
		345678901234567890123456789012345678901234567890123456789012345678901234
2	MT	yyynyyy#nnyffynynynnny??ynynny?nynnynyyny+nnnnnynnynyyynnnynyynnynnynyny
2	NE	nnnynnnnnnnnnnynnnyyynnyyynnnyynyyynnnyynn-yyyynnnnyyynyynnyyyynynnyynnny
2	NH	ynyx+yyyynyy?ynyyynn?nnynynyyynynynnyyyynnnyyynnnnnynnnyynnyynnyyyynyyynn
2	NJ	yyynyyyyynyyynnyyynnnyynynynnn?n?nnynnnnyynnnnnnynynyyynnnn?n?nynnynynyny
2	NM	nnnynnnnnnnnnnynynyy?ynyn?nyyyyyyynnnyyynyyyyyyy?nnynyyyyynyynnyyyynyyny
2	NC	ynnynnnnnnnnnnynnnyyynnynynyyynynynnyyyynnyyyyynnnnpnnnyyynyynyyyyynynnn
2	OK	nnny?nnnn???nnynnnyyyy?nnnnnyynynynnyyynn?ynyyynnnynnyyyynn?yynyyyynyyn?y
2	OR	+ynnyyyynynyynyyyynnyyyny+nynnnyyyy+nyynyyynnynx?n-yynyynnnyn+ynnnnnyyyyy
2	RI	nyynyyyyynnyyynyyynnnyynynnnn+y?ynyynnynyynnnnnyyyynyyynnnynnyynnnnynyyy
2	SC	nnxynnnnnynnnnynnnyyyynynynyyynyyynnyyyynnyyyyyynnyyny+yyynyyyyyyyyyyynn
2	SD	yyynfyyyn.fyyynynynnnyn?nyny?y?y???nnyynnyyyynyy?nyynnyynnyyyynynyyyyn?n
2	TN	nnynnnnynnynnnyynnyyyy?y???yyynynxnnyyyyny?yyyyynn?ynyyyyynyyyyyynnynyny
2	TX	.nyynnnnnyynnnynnny##nnynyyyyyn?n?nnyyyynnyyyyyn?nnnnnnyyynyynnyyyynyynn
2	VA	?nyynnnnyy?nnnynnnyy?nnynynyyynyyynnyyyynyyyyyyyynnyynyyyyynyyyyyyyyyyynny
2	WV	nnnnyyynynnyyyynnynnnyynynynnynnyynynnynnyyynnnyynyyyynyynnynyyyyynnynyny
2	WY	nnyynnnnnnynnnynnnyyyynynynyyyn?nnnnnyyynnyyyyynnnynnyyyyynyyynyyyyyyynnn
3	AL	nnnynnnnnynnnnynynynyyynyyynynyyyn??nyyyyynnyyyyynnnyynynyyyynyyyyyynyynyynn
3	AK	ynyyyyyy.nf+yyyynynyn?y?n?n???ynn??ynyyynyyyyyyynnyynnnyyynyyyyyyyyyyyny
3	AZ	?nn+?nnnnnnnnnynnnyyynny??-yyy??y?nnyyyyn??yy?#?nnynnnnnyynyyynyy?yyy?y?n
3	AR	yyynnyyyyyyyyynynnnnnyynynnnyyyyynnynynnyynynnnnnnynyyynnnynyyyyynnyyyny
3	CA	yyynyyyynnyyyynyyynnnyynynynnnynynnynxyny?nnnnnx?ynnyy+nnnynnnnynnnnnyyy
3	CO	yyy-yyyynnyyyynyyynnnyynyyynnyn?y?n?ynnnny+nnnnny???nyyyn?nynnnnnnnnynyyy
3	CT	yyy-yyyyyyyynyxynynnnyynynnnyy+nyn?ynnnnyynnnnnyy?ynyyynnnynnyynynnynnny
3	FL	nynynnynnynnnyyyynnnnynnyyyyyyynyy?nynyyyyn+?yyynnnyynynyyyynyyyyyyny+yyyny
3	GA	?nyyyny-nny.nnnyynynyyynn??yyyy???n??yyyynnyyyyynnnyynyyyyynyyyyyyyyyyyny
3	HI	.nynyyyfnyy.nynynynny?n?ynynyyyn?nnynyyynyynnnnnnyy#nyyynnnynnnyyxnnynnyy
3	ID	yny?yyyy.nnnyynyyyyny+nn??ynnxy????yyyyynnyyyyyynynynnnyyynyynnyyyynyyynn
3	IL	nnynyyyyyyynnynynyynnynnynnnn#y?y?yynyyynyyn#ynnynyynyynnnnnnynynyyynyyyny
3	IN	fy#ny#+ynnynyynyyynnn?ynynynnyyny?nyyyyynnyyyyny?nyynynyyynyyynyyyyynyy?y
3	IA	yyy-yyyyynyyyynynynnnyynynynnn?n??nyyyyynnynyyyynnyynnnyynnyyynyyyynyyyny
3	KS	nnx+nnn-nnnnnnynynyyynnnnyyyynnynynnnyyynnyyyyyynnyynynyyyynyyynyyyyynyyynn
3	KY	nnnynyyynynynyynyyyynyyyyynynyyyynyynynyyynyyynnnnnnynyyyyyyynnynyyyyyynnyyyny
3	LA	nnyyyyyyynyyn?nyynnynyy?ynnyny?n?n?nnyyyyn?yyyyyynnyyyyynyynny?yyyyynnyyny
3	MD	yyy?yyyyynyyyyyyynynn?ynnynyy?nynyn?ynynnn?nnnnyy?nynnyynynnnnnyynxnnynyyy
3	MO	nnnnyyyynynnnynyny?nny?nynynnnnny?yynnnnyynnnnnnynyyyyynnnynyyyynnnnynnnyy
3	NV	nnn?nnnnnyn?nnyn?nyyynn?nynyyynyn??nyyyynnyyyyyynnyynynynyyyynyyyyy?yynyyynn
3	NH	yynnnyyyyynynynyyynnnyynyyynyynn?ynynyyynynyyyyynnynnnynyyynyyyyyyyyyyyynn
3	NY	nyynyyyyyny?ynnyyynnnynnynyyny?n+??ynyyyyyyyyyyynnyynynyyynyyyyyyyyyyyyyny
3	NC	nnnynyynynyynnnyn-ynynnyynnnyyyyn??nyyyyn-yyyyynnnnynnnyyynyy?yy?yynyyynn
3	ND	nnnynyynnynnnnynnnyyyynynyyyyyn?nynnyyyynyynynyynnyynyynn?ynyyynynnnyyyyy
3	OH	nnynyyyyynyynynynynynnnynnnnynyy?nynnynyyny?nyn?-yyyynyyynynynyyynnnyynyyy
3	OK	nnnynnnyynynnnyynnyynynynnnyynyynnnnyyyynnyyyynnnnyynnyyyyynyyyynyyyyyynnn
3	OR	yyynnnyynyynnnyyyynyyyynnyynyyn?yynnnnyyyy?nnynny?nynnynnynynyyynyynnyyyyny
3	PA	y?nnnnnyyynnnnynyyyyynynyyynyynny?#nnnyynyynnynyynnynyyynnnnnyyynynnyyyyny
3	SC	ynynnyy.yyynnnnynnynyyynnyn?nyynyyn?ynynnyyy?ynn?y??nyyyynnnnyyyyyyyyyyyny
3	SD	y?ynyyyynnyyyynynynnnyyn??ynnn?yy??yyyyynyyyyyyyynnyynnyyyynyyyynyyyyyyyynny
3	UT	?n.ynnnnyynnnnynnnyyyynynynyyn?ynynnyyyynnyyyyynnnnynnyyyyynyynynyyyn+-??
3	VT	yyynyyyynnnyyynynynnnyynyyyn?n?nynyynnnny+nnnnnynyynyyynnnynyyyynynnynyny
3	WA	-y#?nxyynyyynnynynynnnynnyyynnxynynnynyyynyynnyyynnynnyyynyyynyyynyyyyy?y
3	WI	yyynyyyynnyyyynyyynnnyynynynnnynynnyyyyyynyyyyynnnyynynyyyynyyyynnyyyyyyyny

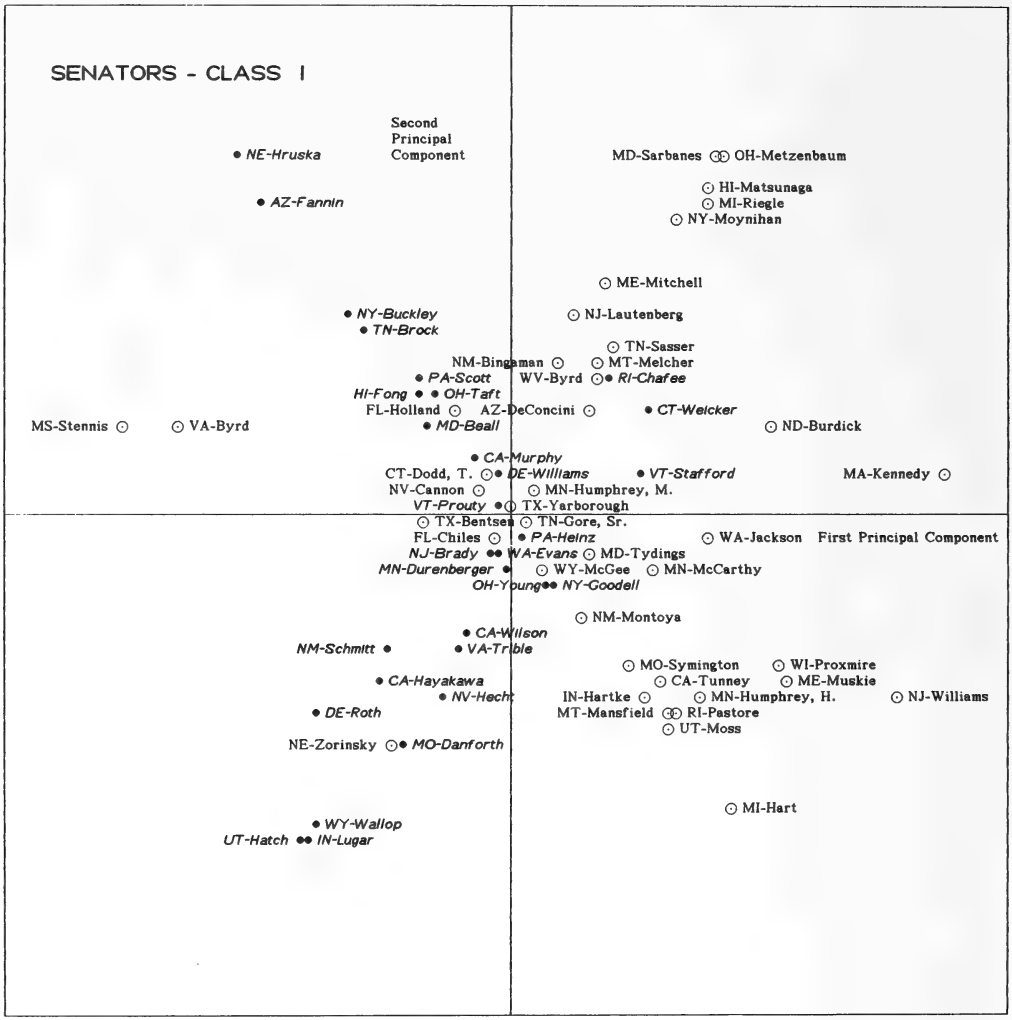

Fig. 1. Principal components for Senators in Class I.

mathematics involves a singular value decomposition by the Jacobi method[10]. Perhaps more efficient methods would be found in *LINPACK*[4] and *EISPACK*[11]. Faddeeva[6] wrote about older but interesting methods. MacRae[9] and Easterling[5] used the related method of factor analysis to analyze voting in the Senate. Davis and McCoy[3] analyzed survey responses in a similar way. The analysis by principal components is far from new, but it seems useful as a first look at the data.

Results

In the figures, the principal component vectors corresponding to the largest and the next largest principal values are plotted, following the widely used methods named "biplots" by Gabriel[7]. In Figures 1, 2 and 3, the senators are divided among the classes to which they belong. The three classes of the Senate differ in their characteristics, which may be seen in the figures. Figure 4 shows the Senate bills,

which are also identified by name in Table 1.

Inspection of the figures allows us to interpret the first, or largest, principal component as a liberal-conservative gradation. The second is not so easily understood, but the sequence of bills in Figure 4 helps us to see a pattern. Early bills, from 1969 and the early '70's, seem to fall along a diagonal line from upper left to lower right. As time passes, the line pivots about the origin until, in the mid-1980's, the line passes from upper right to lower left. A further rotation of the axes might well simplify the picture.

In Figures 1–3, the party affiliation of the Senator is shown by ● *Republican* or ⊙ Democratic symbols. More liberal Senators are more Democratic or north-eastern; more conservative Senators are more Republican or south-western. These observations are well known, but it is interesting that the spectral analysis was performed with no reference to state or

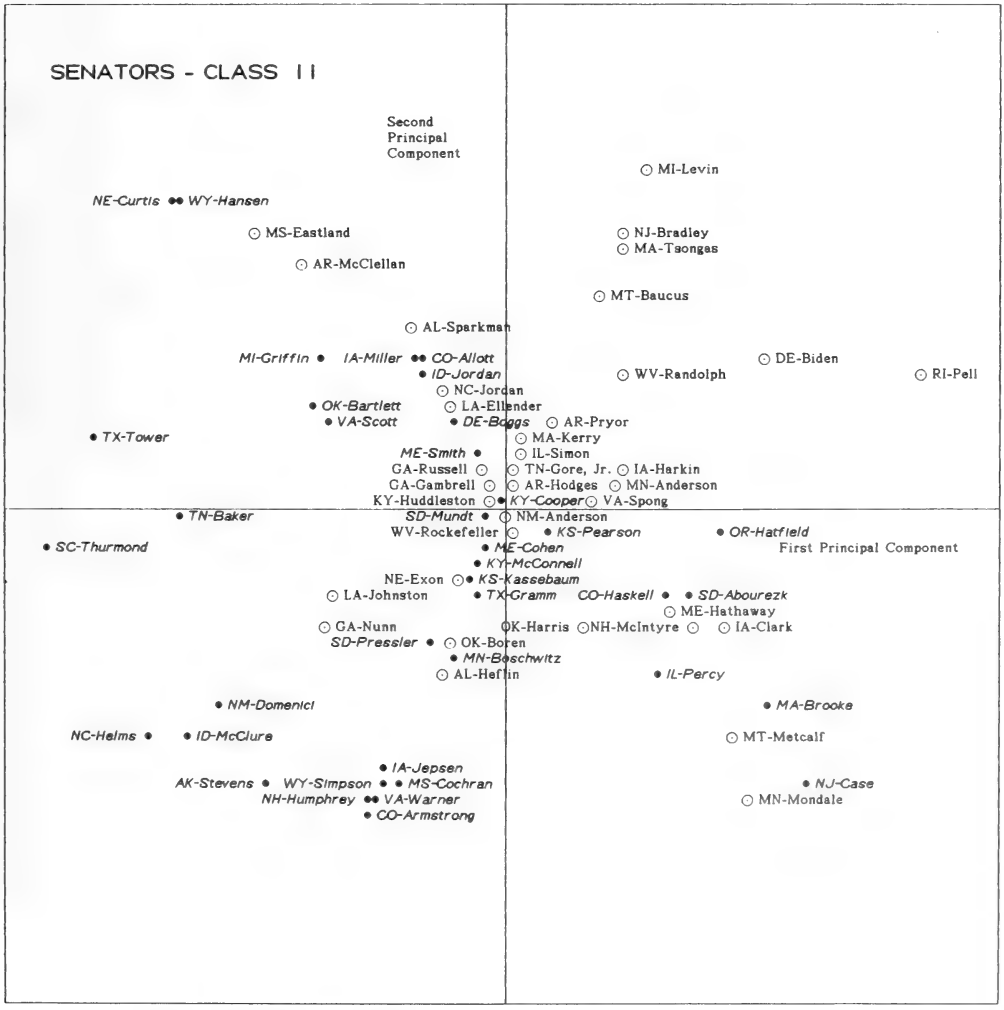

Fig. 2. Principal components for Senators in Class II.

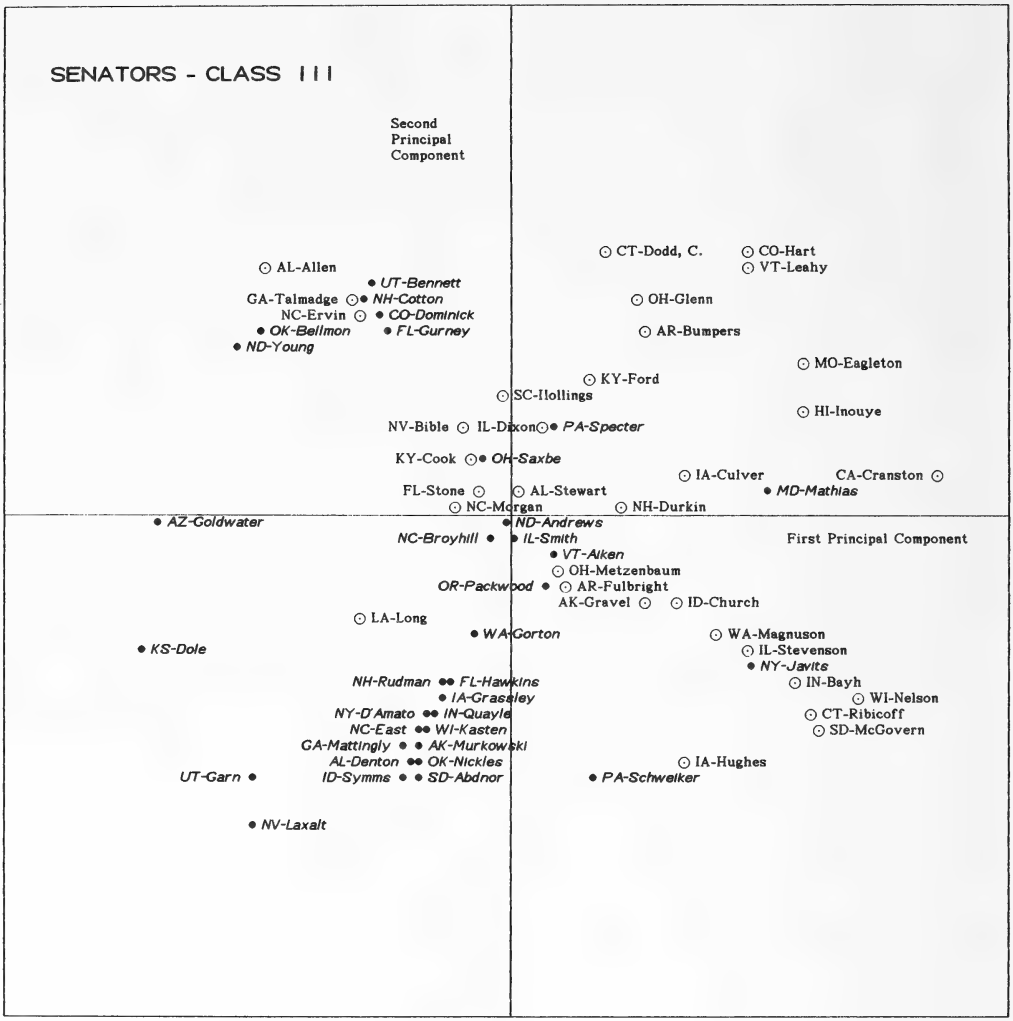

Fig. 3. Principal components for Senators in Class III.

party affiliation. Only the voting record contributed to the results.

In Figure 4, bills that were adopted are shown as solid circles ● and those that were not are shown as open circles ○. There is a pattern reflecting the legislative leadership at varying times. In the earlier years, more liberal bills, appearing in the lower right corner, are passed more often; but in later years, more conservative bills, appearing in the lower left corner are more often carried. Bills toward the edges of the patterns are not often passed, perhaps reflecting a moderating tendency of the Senate to avoid support of extreme legislation.

As Colin Mallows remarked on seeing the graphs in the poster session at the Joint Statistical Meetings in New Orleans, the data would be usefully explored further by computer graphics and shown on videotapes. We are working on that possibility, and we invite others to do so. The data are available on a floppy disc in an ASCII file in IBM-PC compatible form.

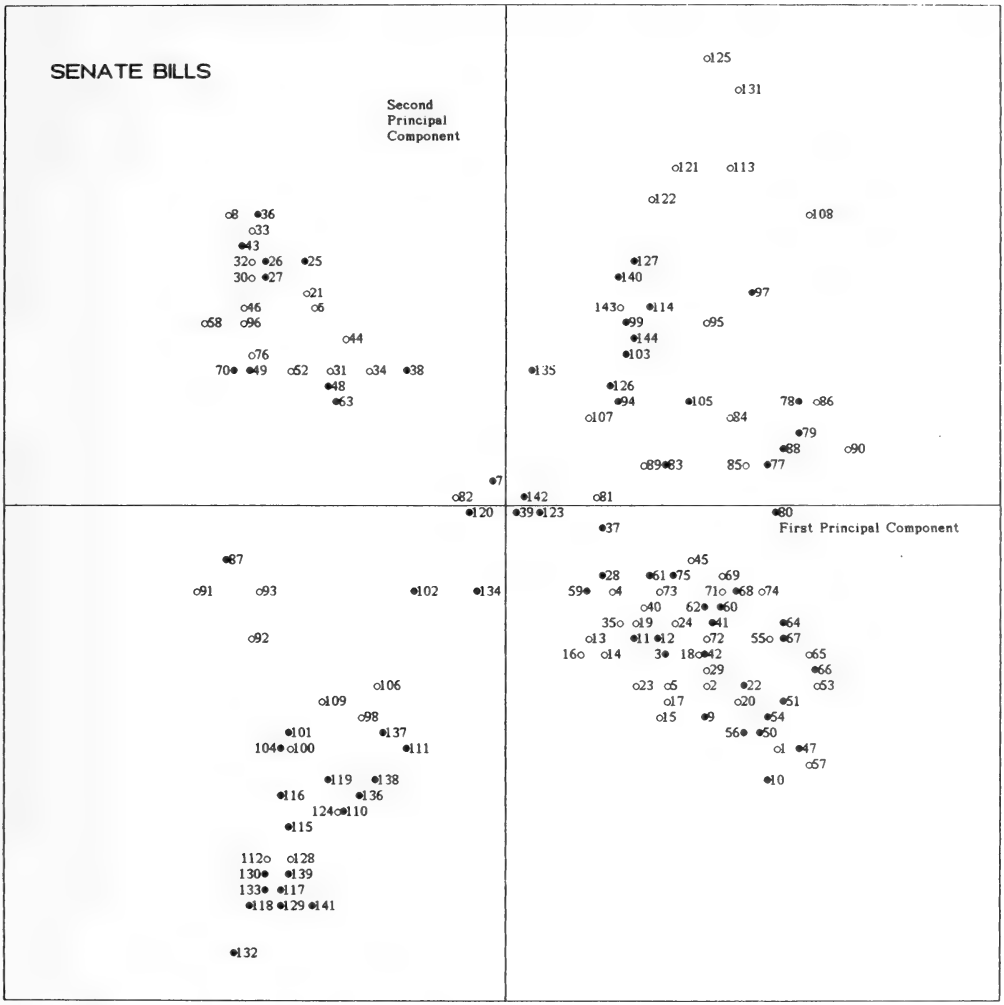

Fig. 4. Principal components for Senate Bills.

Acknowledgment

We thank Charles Stembler for preparing a draft of the voting data in Table 3.

References Cited

1. **Barone, Michael:** Almanac of American Politics, Washington, DC, National Journal, 1972–1987.
2. **Congressional Quarterly:** Washington, DC, Congressional Quarterly, Inc., 1970–1986.
3. **Davis, M. and McCoy, J.:** "A multivariate model for response reliability in surveys," Proceedings of the Section on Survey Research Methods of the American Statistical Association, 603–608, 1978.
4. **Dongarra, J. J., Moler, C. B., Bunch, J. R. and Stewart, G. W.:** LINPACK Users' Guide, Philadelphia, Society for Industrial and Applied Mathematics, 1979.
5. **Easterling, Doublas V.:** Political Science: Using the Generalized Euclidean Model to Study Ideological Shifts in the U. S. Senate, Chapter 10 in Young, Forrest W.: Multidimensional Scaling: History, Theory and Applications, Hillsdale, NJ, Lawrence Erlbaum Associates, 1987.
6. **Faddeeva, V. N.:** Computational Methods of Linear Algebra, New York, Dover, 1959.

7. **Gabriel, K. R.:** The biplot graphic display of matrices with application to principal component analysis, Biometrika, 58:453–467, 1971.
8. **Good, I. J.:** The Estimation of Probabilities, Cambridge, MA, The MIT Press, 1965.
9. **MacRae, D.:** Issues and Parties in Legislative Voting-Methods of Statistical Analysis, New York, Harper and Row, 1970.
10. **Ralston, A. and Wilf, H. S.:** Mathematical Methods for Digital Computers, New York, John Wiley, 1960.
11. **Smith, B. T., Boyle, J. M., Dongarra, J. J., Garbow, B. S., Ikebe, Y., Klema, V. C., Moler, C. B.:** Matrix Eigensystem Routines-EISPACK Guide, 2nd ed., New York, Springer-Verlag, 1976.

Journal of the Washington Academy of Sciences,
Volume 78, Number 4, Pages 310–322, December 1988

Computational Statistics: A New Agenda for Statistical Theory and Practice

Edward J. Wegman*

Center for Computational Statistics
4400 University Drive
George Mason University, Fairfax, VA 22030

ABSTRACT

The impact of workstation and personal computing has important implications for the future of statistics. We argue that the capabilities of new computing environments will change methodological focus because computationally intensive algorithms free of onerous restrictions are feasible in place mathematically tractable but potentially nonrobust algorithms. Moreover electronic instrumentation allows us to collect data substantially different from traditional data collection. In particular, we argue that in place of small, low dimensional homogeneous data sets chosen according to a well designed experiment, we are more likely to see very large, high dimensional nonhomogeneous data sets collected opportunistically. We outline a comparison between traditional statistics and what we call computational statistics. We give several examples a computational statistics and complete our thesis with a discussion of the implications for graduate curricula.

1. Introduction

The spectacular growth in the field of computing science is obvious to all. Indeed, the most obvious manifestations, the ubiquitous microcomputer, is in some ways perhaps the least significant aspect of this revolution. The new pipeline and parallel architectures including systolic arrays and hypercubes, the emergence of

*Correspondence

artificial intelligence, the cheap availability of RAM and color graphics, the impact of high resolution graphics, the potential for optical, biological and chemical computing machines and the pressing need for software/language models for parallel computing are all aspects of the computation revolution which figure prominently in the other sciences.

While not as obvious to the casual observer, the fields of statistics and probability have experienced equally spectacular technical achievements including weak convergence theory, the almost sure invariance principle, exploratory data analysis, nonlinear time series methods, bootstrapping, semiparametric methods, percolation theory, simulated annealing and the like principally within the last decade. The computing and statistical technologies both figure prominently in the manipulation and analysis of data and information. It is natural then to consider the linkage between these two discipline areas. The interface between these two discipline areas has been labeled by the phrase, "statistical computing," and more recently, "computational statistics." We should like to distinguish between the two and, indeed, argue that the latter embodies a rather significantly different approach to statistical inference.

In thinking carefully about the relationship between computing science and statistical science it is possible to describe a large number of linkages. Two that come immediately to mind are the stochastic description of data flow through a computer and the characterization of uncertainty in expert systems. In the former case, we can view a computing architecture, particularly a parallel or distributed architecture as a network with messages being passed from node to node. This is essentially a queueing network and the characterization of the distribution state of the network becomes a problem in stochastic model building and estimation. In the latter example, a rule-based expert system is invariably characterized by imprecision in the specification of the basic predicates derived from the expert or experts. This may be because either the rule is a "rule-of-thumb" and hence inherently imprecise or because the rule is a not yet fully formulated and verified inference and hence exogenously stochastic. The assignment of probabilities (or such alternatives to probabilities as belief functions or fuzzy set functions) in a useful way is another application of statistical methodologies to computing science. In both of these examples statistical methodology is employed in the development of computing science. Both of these examples might legitimately be called statistical computing since the focus is on computing with statistics as an adjectival modifier.

Traditionally, of course, this is not at all what statistical computing means. Statistics is fundamentally an applied science, hence, a computationally oriented science. A statistical theory is useless without a suitable algorithm to go with it. Statistical computing has traditionally meant the conversion of statistical algorithms into a reasonably friendly computer code. This enterprise became feasible with the development of the IBM 360 series in the early 1960s and the slightly later development of the DEC Vax series of computers. Of course, the trend has accelerated with the ubiquitous PC. Packages such as SAS, BMDP, Minitab, SPSS and the like represent a very high evolution in statistical computing. When we use the phrase, "computational statistics," we have in mind a rather stronger focus on the exploitation of computing in the creation of new statistical methodology. We shall give some explicit examples shortly.

2. Statistics as an Information Technology

Statistics is fundamentally about the transformation of raw data into useful information. As such statistics is a fundamental information technology. It is,

therefore, appropriate to see statistical science as intimately related not only to computing science, but also communication technology, electrical engineering and systems engineering as part of the spectrum of modern information processing and handling technologies. Statistics is perhaps the oldest and most theoretically well developed of these information technologies.

It is thus important to understand how the changing face of these technologies affect statistics and, more importantly, how they offer new opportunity for the development of statistical methodologies. In particular, it is important to understand how the computer revolution is affecting the accumulation of data. Electronic instrumentation implies an ability to acquire a large amount of high dimensional data very rapidly. While such capabilities have existed for some time, the emergence of cheap RAM in the 1980's has given us the ability to store and access that data in an active computer memory. Satellite-based remote sensing, weather and pollution monitoring, data base transactions, computer controlled industrial automation and computer controlled laboratory instrumentation as well as computer simulations are all sources of such complex data sets. We contend that this new class of data represents a challenge for statisticians which is substantially different in kind. In many ways the characteristics of automated data are different from traditional data. Automated data is generally untouched by human brain. While there are fewer transcription errors, there are also fewer checks for reasonableness. There are likewise different economic considerations. In many traditional data collection regimes, the cost per item of data is expensive. In the automated mode, set-up costs are expensive, but once accomplished there is low incremental cost for taking additional data. Thus it is often easy to replicate, hence, increased sample size, and it is easy to take additional measurements on each sample item (variables), hence an increase in dimensionality. As an adjunct, it is worth pointing out that when an individual datum is expensive, we collect no more than absolutely needed to complete the inference. However, when the marginal cost of additional data is low, we tend to collect much more both in sample size and dimension motivated by the belief that it might be useful for some not-yet-precisely-specified purpose. Thus, the computing revolution often implies that there are less sharply focused reasons for collecting data.

The majority of existing methodology is focused on the univariate, IID random variable model. Even in the circumstance that a multivariate model is entertained, it is usually assumed to be multivariate normal. We contend, in addition, that while arbitrary sample size is frequently assumed, the truth of the matter is that these techniques implicity assume small to moderate sample sizes. For example, a regression problem with 5 design variables and 1000 observations would represent no problem for traditional techniques. By contrast, a regression problem with 40,000 design variables and 8 million observations would. The reason is clear. In the former case the emphasis is on statistical efficiency which is the operational goal for most current statistical technology. However, with massive replications the need for statistical efficiency is much less pervasive. Indeed, we may find it desireable to exchange highly efficient, parametric procedures for less efficient, but more robust, nonparametric procedures to guard against violations of (possibly untestable) model assumptions. The emphasis on parsimony in many contemporary books and papers is a further reflection of the mind-set that implicitly focuses on small to moderate sample sizes since few parameters do not necessarily make sense in the context of very large sample sizes. Finally, we note that the very fact of largeness in sample size implies that it is unlikely we would see IID homogeneity.

Thus, the computer revolution implies a revolution in the type of data we are

Table 1.—Comparison of Traditional Statistics

Traditional Statistics	Computational Statistics
Small to Moderate Sample Size	Large to Very Large Sample size
IID Data Sets	Nonhomogeneous Data Sets
One or Low Dimensional	High Dimensional
Manually Computational	Computationally Intensive
Mathematically Tractable	Numerically Tractable
Well Focused Questions	Imprecise Questions
Strong Unverifiable Assumptions relationships (linearity, additivity) error structures (normality)	Weak or No Assumptions relationships (nonlinearity) error structures (distribution free)
Statistical Inference	Structural Inference
Predominantly Closed Form Algorithms	Iterative Algorithms Possible
Statistical Optimality	Statistical Robustness

able to accumulate, i.e. large, high-dimensional nonhomogeneous data sets. More importantly, however, the richness of these new data structures suggest that we are more demanding of the data. In a simple univariate IID setting we primarily are concerned with variability of a single random variable and questions related to this variability. In a more complex data set, however, it is natural to ask more involved questions while simultaneously having less a priori insight into the structure of the data. When more variables are available, we would certainly ask about the functional relationship among them. We have used the phrase, structural inference, for methodologies aimed at describing such relationships, deliberately contrasting this phrase with the phrase, statistical inference, which has traditionally been about determining variability (i.e. probability distributions). With a more complex structure, we are open to ask more of the data than just simple decisions. We may want to ask forecasting questions, e.g. about weather, economic projections, medical diagnosis, university admissions and tax audits, or questions about automated pattern recognition, e.g. printed or hand-written characters, speech recognition and robotic vision, or questions about system optimization, e.g., process and quality control, drug effectiveness, chemical yields and physical design.

The contrast between traditional statistics and the new statistics is marked and strong. We summarize in Table 1. We think that the implications of the computer revolution on statistics are sufficient to warrant new terminology, specifically computational statistics. This terminology not only provides the linkage of statistical science with computing science, but also puts the focus on statistics rather than on computing.

3. Some Examples

In this section we should like to illustrate these notions of computational statistics with three examples of approaches to data that flow from thinking about statistics in light of contemporary computation. It goes without saying that such techniques as bootstrapping, cross validation and high interaction graphics are clear well-known examples of what we call computational statistics. We wish to describe some less well known ideas: 1. high dimensional graphical representation, 2. functional inference, and 3. data set mapping.

3.1 Example: High Dimensional Graphical Representation

Large, high dimensional data sets often have a complex, non-linear structure. For

this reason, exploratory analysis is even more important for such data sets than it is in more traditional well-structured data sets. Visualization of data structures in higher dimensions is, however, even more difficult than in low dimensional cases since geometric representation of data with cartesian coordinates is impossible in dimensions higher than three.

One interesting approach, the parallel coordinate representation, has been pursued by Inselberg (1985)[3] and Wegman (1986).[8] A point in n-dimensional space is represented in Figure 3.1. A parallel coordinate representation consists of n parallel axes, each axis meant to represent one dimension of the n-dimensional vector. A point in n-space is plotted by marking x_i on the ith axis and joining x_1 through x_n by a broken line segment. Thus a point in Euclidean n-space is mapped into a broken line segment in the parallel representation. Figure 3.1 represents two points coinciding in the 4th coordinate. Figure 3.2 represents a more complex but artificially contrived data set. While we will not try to develop intuition for the parallel coordinate representation in this paper, our experience has been that with a very modest amount of training, people rapidly become adept at interpreting these diagrams. Figure 3.2 has several features of interest including a normal marginal density in dimension one, a negative chi-square of dimension two, a three dimensional cluster in dimensions 3 to 5, a five dimensional mode as well as linear and nonlinear functional relationships.

The mathematical structure of parallel coordinate diagrams turns out to be ex-

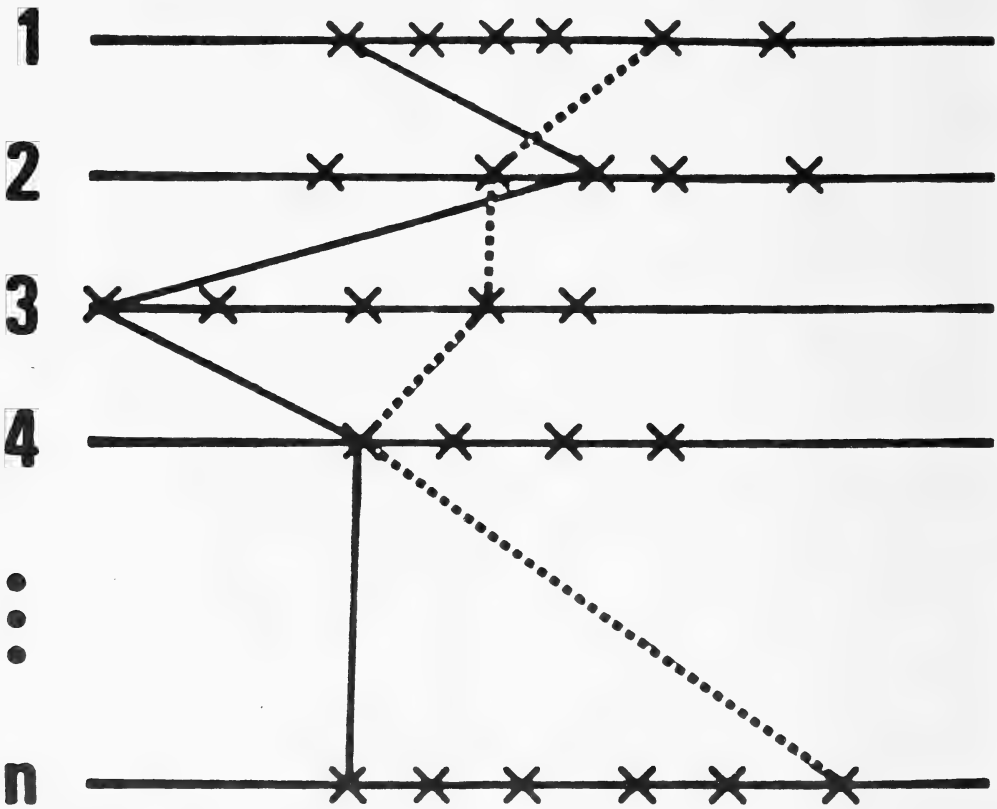

Fig. 3.1. Two points in n-dimensional space plotted in parallel coordinates. These points agree in the 4th coordinate.

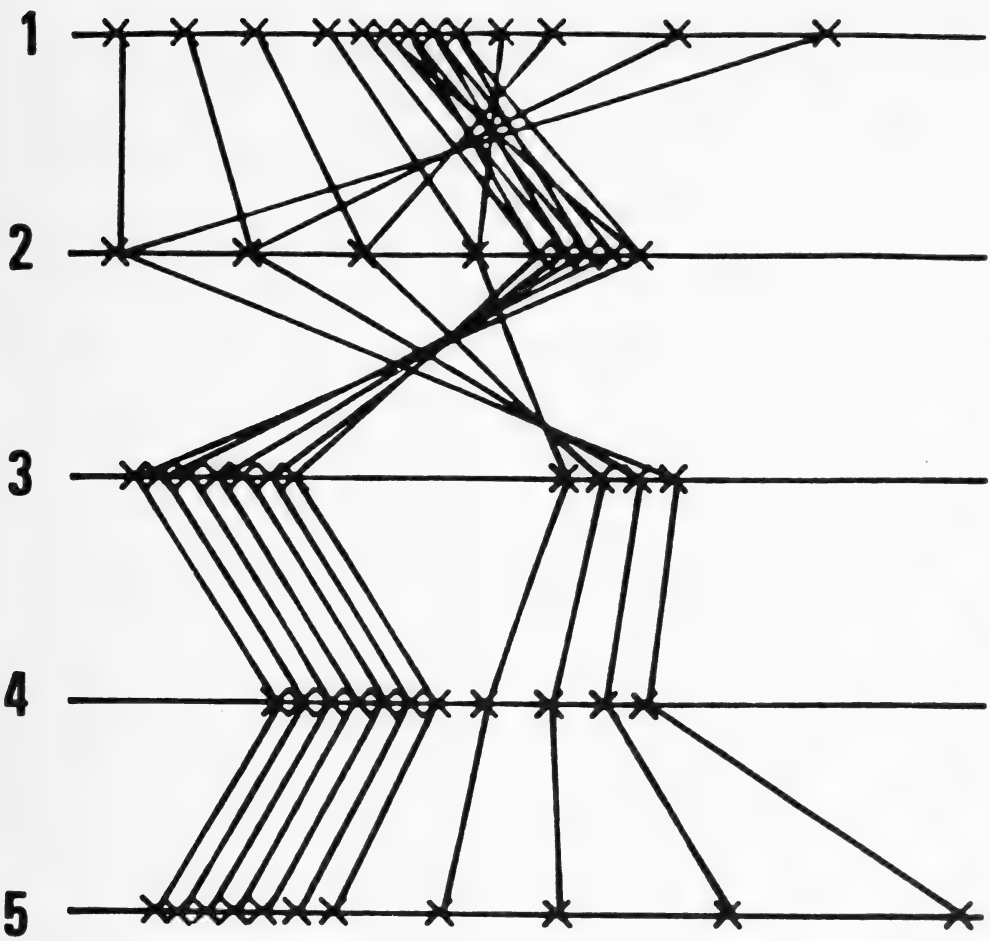

Fig. 3.2. Illustration of five dimensional data set represented in parallel coordinates.

tremely interesting. Focusing on the first two dimensions of Figure 3.1 for a moment, one observes that a point in ordinary cartesian coordinates maps into a line (segment) in parallel coordinates. This point-line mapping is suggestive of the duality found in projective geometry. Indeed, projective geometry plays a key role in the development of the parallel coordinate representation. If both the cartesian plane and two dimensional parallel coordinate plane are augmented with the appropriate ideal points so that they are both projective planes, the transformation from the cartesian coordinates to the parallel coordinates becomes a projective transformation or projectivity. A projectivity can be represented as a matrix transformation on the so-called natural homogeneous coordinates. The particular cartesian-to-parallel-coordinates transformation induces a number of interesting dualities.

Not only do points map into lines, but lines map into points as well. Conics map into conics, rotations map into translations, translations into rotations and interestingly enough points of inflection map into cusps and vice versa. There are several interesting implications of these facts beyond just the graphic display value of parallel coordinates. For one thing, since

translations are relatively easy to compute while rotations are relatively harder, there is a computational advantage to a parallel coodinate representation when heavy use of rotations is expected. It is also clear that points of inflection are relatively difficult to detect while cusps are relatively easy. We believe that the parallel coordinate representation will be an advantageous one for computational geometry as well as statistical data analysis.

The parallel coordinate diagram serves as a hyperdimensional analogue to the traditional scatter diagram and thus may be used as a fundamental tool for high dimensional exploratory data analysis. Several notions have already been explored, but these are certainly preliminary. It is known, for example, that the number of crossings of line segments between adjacent pairs of parallel axes is invariant with scale transformation. Such invariance suggests that features of the parallel coordinate axes depend only on ranks and thus may hve robustness features common to rank-based statistical methods. Yet another question of significant interest relates to dimensionality reduction. Is it possible, for example, to find a simple graphical algorithm for dimensionality reduction based on rotations, translations and nonlinear scaling? If such a procedure were available either automatically or with heuristic guidance, it would be a fundamental tool in model building.

3.2 Example: Functional Inference

Our fundamental premise is that we are interested in the structural relationship among a set of random variates. We formulate this as follows. Given the random variables $X_1, X_2, \ldots, X_n$ which are functionally related by an equation, $f(X_1, \ldots, X_n) = \epsilon$, determine f. This is a generalization of the problem of finding the prediction equation in the standard regression problem, $Y - x\beta = \epsilon$, but in a nonparametric, nonlinear setting. We suggest the following notion. Let $M = \{(x_1, \ldots, x_n): Ef(x_1, \ldots, x_n) = 0\}$. M in general is an algebraic variety and under reasonable regularity conditions a manifold. Thus, there is a fundamental equivalency between estimating the geometric manifold M and estimating the function f.

By turning our attention to the manifold M, we make the problem a geometric one, one whose structure is easier to visualize using computer graphics. For this reason we believe there is an intimate connection between the structural estimation problem and the visualization of high dimensional manifolds. While graphical methods for looking at point clouds have proven stimulating to the imagination, it is extremely difficult to understand true hyperdimensional structure, particularly when rotating about an invisible axis. We believe that a solid structure as opposed to a point cloud would provide the visual continuity to alleviate much of this problem. This solid structure is what we identify with the manifold M. To get a handle on the procedure for estimating a manifold we note a d-ridge is the extremal d-dimensional feature on a hyperspace structure of dimension greater than d. The 0-ridge corresponds to the usual mode. For some d, we contend that a reasonable estimate of M is the d-ridge of the n-dimensional density function of $(X_1, \ldots, X_n)$. This in effect is estimating the d-dimensional summary manifold with a mode-like estimator. In essence what we are suggesting to skeletonize hyperdimensional structures. This type of process has been done in the image processing context with good computational efficiency.

The inference technique then is to estimate M nonparametrically. This reduces the scatter diagram to a geometrically described hyperdimensional solid. This can then be explored geometrically using computer graphics. Features can then be parametrized and a composite parametric model constructed. The inference can be completed by a confirmatory analysis on the (perhaps nonlinear) parametric model.

Techniques presently in use often assume linearity or special forms of nonlinearity, e.g. polynomial or spline fits, and often assume additivity a low dimensional subcomponents, e.g. projection pursuit. We would not like to take this perspective a priori. While this methodology may consequently seem complex, the premise is that geometric-based structural analysis will offer tools superior to traditional purely analytic methods for building high dimensional functional models.

The key to this development is to appreciate the role of ridges in describing relationships between random variables. A simple two-dimensional example is illustrated in Figure 3.3. Note that the contours represent the density and the 1-ridge represents the functional relationship between x and y in a traditional linear regression. Since densities are key, a fast, efficient multidimensional density estimation technique is important.

The slowness of the traditional kernel estimators in a high dimensional setting arises from the fact that they are essentially point estimators. That is to say, to compute $f(x)$ one needs to do smoothing in a neighborhood of x. For a satisfying visual representation, the x's must be chosen reasonably dense. Moreover, traditional kernels are frequently nonlinear functions and thus the computation involves the repeated evaluation of a nonlinear function at each of the observations of a potentially large data set on a dense

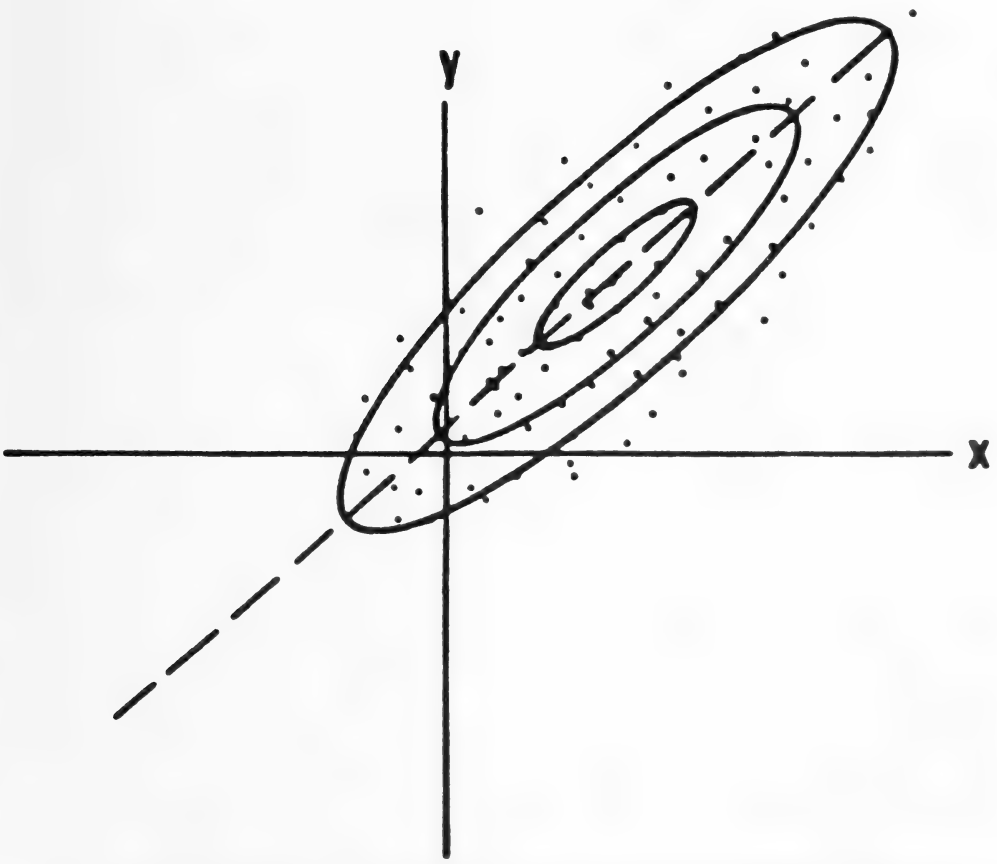

Fig. 3.3. Two dimensional scatter diagram with the contours of the two dimensional density function and the 1-ridge summary line.

set of points in the domain. Furthermore, as dimension increases, exponentially more domain points are required to maintain a constant number of points per unit hypervolume. Traditional kernel estimators become essentially useless for even relatively low dimensions.

The traditional histogram provides an alternative strategy. The histogram is a two-step procedure. The first step is a tesselation of the line. The second step is an assignment of each observation to a tile of that tesselation. The computation of the actual density estimator amounts to a simple rescaling of tile- (cell-) count. The histogram is a global estimator since the function is constant on the tiles which are finite in number and, indeed, relatively few in number compared with the denseness of points required for the kernel estimator. The traditional histogram, of course, operates with fixed equally spaced uniform tiles. There is no reason why the tiles must be fixed or uniform. Wegman (1975)[7] suggests a data-dependent tesselation in the one dimensional setting and shows that, if the number of tiles is allowed to grow at an appropriate rate with the increase of sample size, then asymptotic consistency can be achieved.

The ingenious papers by Scott (1985, 1986)[5,6] introduces the notion of the average shifted histogram, ASH. Scott recognizes the computational speed of a global estimator such as the histogram. His algorithm computes the histogram for a variety of tesselations and then averages these together to obtain smoothing properties. In this paper we are suggesting a combination of these two ideas. We propose a data-driven tesselation of the following sort. Take an $\alpha\%$ (10%) subsample of the sample. Use these points to form a Dirichlet tesselation of n-space. A two dimensional example is given in Figure 3.4. The tiles of the Dirichlet tesselation form the data-dependent convex regions upon which to base the density estimator. One pass through the data will be sufficient to classify each point according to tile and thence a simple rescaling to compute the estimator. Repeated subsampling will yield additional estimators which can then be averaged in the manner of Scott's ASH. The details of this algorithm need to are being explored, but the following conjectures are made. Asymptotic properties similar to those found in Wegman (1975)[7] hold. Maximum likelihood and nearest neighbor properties will hold. Computational efficiency will be substantially better than with kernal methods. Because of the repeated sampling, bootstrap-type behavior will hold.

It should be clear that the notions we are suggesting rely heavily on the generalization of present geometric algorithms to higher dimensional space. The Dirichlet tesselation, for example, has useful algorithms in 2 and 3 dimensions (see Bowyer, 1981, Green and Gibson, 1978).[1,2] but the analogues in higher dimensions are poorly developed (see Preparata and Shamos, 1986).[4] Thus a fundamental exploration of algorithms for these tesselations in hyperspace must still be done. Interestingly enough, the computation of tasselations in hyperspace is closely related to the computation of convex hulls (again see Preparata and Shamos, 1986, p. 246).[4]

The construction of the n-dimensional Dirichlet tesselation (Voronoi diagram) is a key element of the density estimation technique we suggest. A related issue is the assignment problem, that is, given the tiles, what is the best algorithm for determining to which of the tiles a given observation belongs. In part the 2-dimensional Voronoi diagrams were constructed to answer nearest-neighbor-type questions. That is, given n points, what is the best algorithm (minimum compute time) for finding the nearest neighbor. In general, the answer is known to be $O(nd)$ where d is the dimension of the space. There is some reasonable expectation that, since the construction of the tesselation and the assignment problem are closely linked, there is an efficient one-step algorithm for sorting the observa-

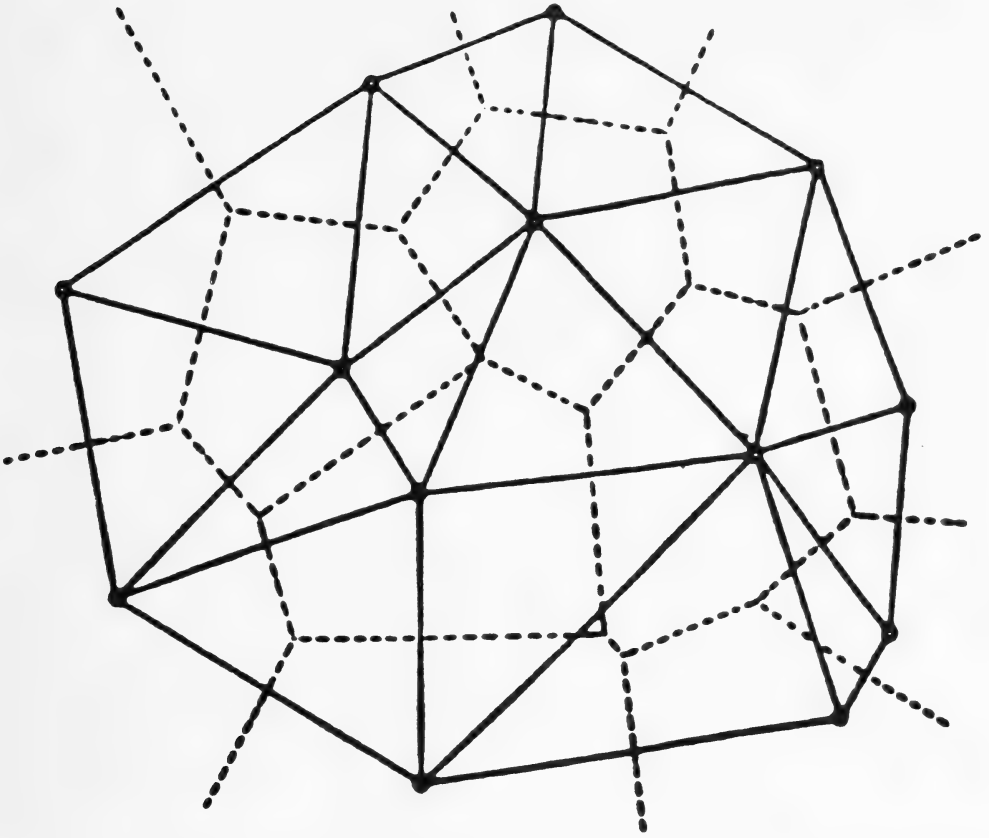

Fig. 3.4. Dirichlet tesselation (dashed lines) and Delauney triangulation (solid lines) of the plane based on 13 points.

tions in convex regions about certain nearest neighbors. If this could be done in linear, near linear or even polynomial time, the density estimation technique we are suggesting may be computationally feasible for relatively high-dimensional cases. In any case, the sorting, clustering and classification results which form the core of computational geometry also form the core of this approach to higher dimensional data structures.

We hope that this example makes clear the mathematical complexity inherent in computational statistics. It is our view that there is an extremely important role for mathematical statistics under the general rubric of computational statistics.

3.3 Example: Data Set Mapping

A traditional way of thinking about the model building process is that we begin with a fixed data set and apply a number of exploratory procedures to it in search of structure within the data set. The data set is regarded as fixed and the analysis procedures as variable. Of course, the model is iteratively refined by checking the residual structure until a suitable model reduces the residuals to a unstructured set of "random numbers." We suggest an alternative way of thinking.

Normal Mode of Analysis:

Data Set Fixed ← Try Different Techniques on It

Alternative Mode of Analysis:
Techniques Set Fixed ← Try Different Data on It

We have in mind the following. With the cost or availability of computational resources essentially not a significant consideration in the analysis procedure (i.e., they are essentially a free good), we can afford to standardize on say a dozen or more techniques which are always computed no matter what data are presented. Others, of course, would still be optionally available. Such standard techniques might include for example standard descriptive statistics, smoothers, spectral estimators, probability density estimators and graphical displays including 3-D projections, scatter diagram matrices, parallel coordinate plots, grand tours, Q-Q plots, variable aspect ratio plots and so on. Each of these might be implemented on a different node of a parallel computing device and displayed in a window of a high resolution graphics workstation, analogous to having a set of papers on our desk through which we might shuffle at will, the difference being that each sheet of paper would, in effect, contain a dynamic, possibly multidimensional display with which the analyst might interact. We have in mind viewing each of these data representations as an attribute of the data set (object orientation) so that if we modify the data set representation in one window, the fundamental data set is modified and consequently its representations in all of the windows are modified simultaneously.

With the set of techniques fixed, a data analysis proceeds through an iterative mapping of the data set, i.e. the data set is iteratively re-expressed. This is done by a series of techniques. A discriminant procedure or a graphical brushing procedure allows us to transform one data set into a number of more homogeneous data subsets. Data transformations, rescaling either linear or nonlinear, clustering, removing outliers, transforming to ranks, bootstrapping, spline fitting and model building are all techniques for mapping an old data set into one or more new ones. Notice that we treat model building as a simple data map. It is our perspective that a model fit is just a transformation of one data set to another (specifically the residuals) similar to any of the others mentioned.

Two points of interest can be made. First, the analysis of a data set can be viewed as the development of a data tree structure—each node is a data set and each edge is a transformation or re-expression of that data set to a new data set. The data tree structure preserves the record of the data analysis, indeed, the data tree is the data analysis. At the bottom of the data tree presumably we will have data sets with no remaining structure for, if not, then another iteration of our analysis and another edge in the data tree are required. The second point to be made is that thinking in the terms just described helps clarify our thinking by conceptually separating the representational methods (e.g. graphics, desciptive statistics) from the re-expression methods (transformations, brushing, outlier removal, model building). These are really two separate functions of our statistical methodology which are not commonly distinguished, but when distinguished, aid in clearer thinking. We particularly think it is helpful to understand, for example, that a square-root data transformation and a ARMA-model fitting are really quite similar operations each resulting in a new data set save that in the latter case we usually call the new data set the set of residuals. Indeed, when the data analysis is completely laid out as a data tree, the full model is really accumulated by starting at the root node (original data set) and following the edges all the way to the ending node (unstructure residual data set).

4. Curriculum Implications

First of all, it is important to recognize that operating from the perspective of

computational statistics does not imply that traditional statistical methodologies are obsolete. While some automated data sets will have the characteristics described earlier, many will not. In addition there will, no doubt, continue to be many carefully designed experiments in which each data point is costly to acquire, and, hence, traditional methods will continue to be used. Nowhere is this probably more true than in the case of bio-medical clinical trials. We believe there are certain shifts in emphasis, however, that may be useful to recognize.

In terms of mathematical preparation, real and complex analysis and measure theory are frequently emphasized as elements of a graduate curriculum. These are tied to probability theory and the more-or-less standard IID parametric assumptions. To the extent that the questions we ask of data are less structured, much of the need for the standard frameworks and probability models is lessened. In their place we will tend to have a more geometric analysis and a more function-oriented, nonparametric framework. The first and second examples were deliberately chosen to illustrate elements of projective geometry, differential geometry and computational geometry. In addition, there is a strong element of nonparametric functional inference in example 3.2 suggesting a heavier reliance on functional analysis. We believe therefore that functional analysis and geometric analysis should become part of the mathematical precursors to a statistics curriculum.

That computation will play a role is obvious, exactly what form it will take is not so obvious. Certain elements of computing science would appear to be candidates. Computational literacy means more than FORTRAN or C programming or familiarity with the statistical packages. Clearly such concepts as object-oriented programming, parallel architectures, computer graphics and numerical methods will play a significant role in future curricula.

In terms of the statistical core material itself, I believe the obsession with the parametric framework (be it classical or Bayesian) must end. It is important to recognize that we are often dealing with opportunistically collected data and that classical formulations are too rigid. Moreover, as indicated earlier, we are asking more of our data than simply distributional questions. Indeed the more interesting questions are the structural questions (i.e., in the face of uncertainty, how are two or more random variables related to each other?). Thus, it would seem that there should be some de-emphasis of traditional mathematical statistics (testing and estimation) and more emphasis on exploratory and structural inference. A contrast between a traditional first year curriculum and a revised curriculum is laid out in Figure 4.1.

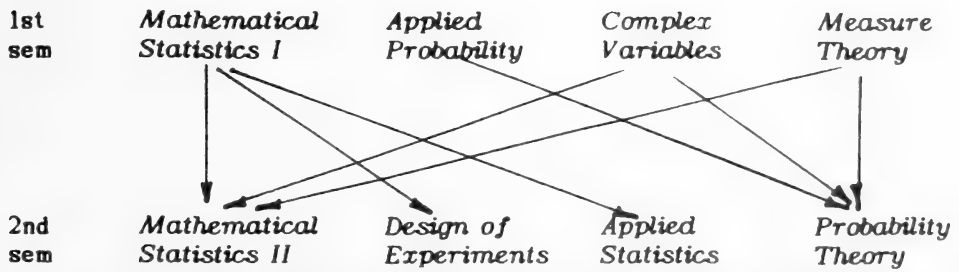

Fig. 4.1.a. A straw-man traditional first year curriculum for a Ph.D. Program in Statistics. Linkages between courses are indicated by directed edges.

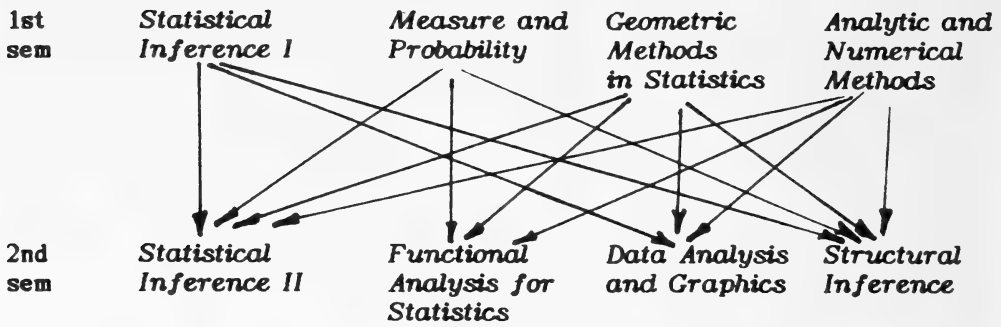

Fig. 4.1.b. A straw-man first year graduate curriculum in Computational Statistics. Additional course work should include data structures and programming, and computing architectures

Acknowledgements

This paper has benefitted from several conversations with Jerry Friedman. I would like to thank him for not only these discussions, but also his enthusiasm for the computational statistics. This research was supported by the Air Force Office of Scientific Research under Grant AFOSR-87-0179 and by the Army Research Office under Grant DAAL03-87-G-0070.

References Cited

1. **Bowyer, A.** (1981), "Computing Dirichlet Tesselations," Computer J. **24,** 164–166.
2. **Green, P. J. and Gibson, R.** (1978), "Computing Dirichlet Tesselations in the Plane," Computer J. **21,** 168–173.
3. **Inselberg, A.** (1985), "The Plane with Parallel Coordinates," The Visual Computer **1,** 69–91.
4. **Preparata, F. P. and Shamos, M. L.** (1986), Computational Geometry: An Introduction, New York: Springer-Verlag, Inc.
5. **Scott, D. W.** (1985), "Average Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions," Ann. Statist. **13,** 1024–1040.
6. **Scott, D. W.** (1986), "Data Analysis in 3 and 4 dimensions with Nonparametric Density Estimation," in Statistical Image Processing and Graphics, (Wegman, E. and DePriest, D. eds.) New York: Marcel Dekker, Inc.
7. **Wegman, E. J.** (1975), "Maximum Likelihood Estimation of a Probability Density, Sankhya (A) **37,** 211–224.
8. **Wegman, E. J.** (1986), "Hyperdimensional Data Analysis Using Parallel Coordinates," Technical Report 1, Center for Computational Statistics and Probability, George Mason University, Fairfax, VA, July 1986.

Journal of the Washington Academy of Sciences,
Volume 78, Number 4, Pages 323–332, December 1988

Statistical Analysis of Experiments to Measure Ignition of Cigarettes

Keith R. Eberhardt

Statistical Engineering Division
Center for Computing and Applied Mathematics
National Institute of Standards and Technology
Administration Building, Room A337
Gaithersburg, Maryland 20899

ABSTRACT

Under the Cigarette Safety Act of 1984, NIST was given the task of studying several types of commercial and experimental cigarettes to determine their relative propensities to ignite soft furnishings. The analysis of the data came under close scrutiny by the Technical Study Group appointed to oversee the research. In one experiment where the usual chi-squared test could not be readiliy justified, an extension of Fisher's Exact Test to 2×12 contingency tables was adopted. In another experiment, a modification of the angular transformation for count data was used along with normal probability plots of the effects to analyze a 2^5 factorial experiment.

Key Words: angular transformation, chi-squared test, contingency table, factorial experiment, Fisher's exact test, normal probability plot, statistical analysis

1. Introduction

This report describes some statistical data analysis aspects of two related research projects[1,2] concerned with the propensity of commercial and experimental cigarettes to ignite upholstered furniture. These projects were conducted in the Center for Fire Research at the National Institute of Standards and Technology (NIST, formerly the National Bureau of Standards) during 1986 and 1987.

Cigarette ignition of furniture is by far the leading cause of fire deaths and injuries in the United States. While the ignition resistance of manufactured furnishings has been greatly improved over the last decade, fire casualties could be further reduced if cigarettes were manufactured to cause fewer ignitions. In response to this situation, Public Law 98-567, the "Cigarette Safety Act of 1984," established a Technical Study Group on Cigarette and Little Cigar Fire Safety composed of representatives of the tobacco industry, the furniture industry, the fire service, the public health advocacy, and concerned Federal agencies. As part of their charge to design and oversee a research program on cigarette ignition propensity, the Technical Study Group engaged NIST to conduct the laboratory experiments described here.

The following two sections describe experiments designed to compare the performance of various types of commercial and experimental cigarettes under a variety of test conditions. To measure the ignition propensity of cigarettes, test mockups were constructed to simulate conditions corresponding to what happens when a lighted cigarette is dropped on an upholstered chair. The mockups were constructed using a variety of upholstery fabrics and padding types which were chosen to represent a range of substrates (i.e. fabric and padding combinations) that can ignite with commercial cigarettes. For each test condition, four or five cigarettes were lighted and placed on mockups, and the number of cigarettes that ignited the substrate was recorded. Thus, the basic data in these experiments consist of counts of the number of ignitions in a given number of trials. A common probability model, the binomial distribution, applies to the data for both cases to be described, but different statistical methods were required due to differences in the questions asked and in the experiment designs used.

2. Differences among Commercial Cigarettes

An important part of the first NIST project for the Technical Study Group was an experiment to determine whether there are measurable differences in the ignition propensities of different types of commercial cigarettes. To study this question, a test protocol was developed under which 12 types of commercial cigarettes were tested on 18 different mockup configurations. The experiment results are displayed in Table 1. The 18 mockup configurations, which are described fully in [1], differ in type of substrate used, whether or not the cigarette was placed in a crevice, and whether or not the cigarette under test was covered with a piece of cotton sheeting.

Inspection of Table 1 reveals that many of the substrates were either too ignition prone (columns 1–3), or too resistant (columns 12–18), to show any differences among the cigarettes tested. This was disappointing to the experimenters because the fabric and padding combinations used for the mockups had been pretested and were chosen carefully to represent cases where differences could occur. However, the pretest samples came from different lots of material than the larger quantities which were procured for construction of the mockups, and although both were nominally the same, their behavior in cigarette ignition tests was quite different.

When the test data of Table 1 were presented to the Technical Study Group, a wide variety of statistical analyses—with contradictory intepretations—were proposed. The reader can readily imagine how the vested interests of the organizations represented on the Technical Study Group would lead some members to favor analyses with conclusions opposite to those preferred by other members. When the collection of competing statistical analyses was presented to this author for his opinion, it was clear that whatever analysis he proposed would be carefully, and perhaps critically, reviewed by at least half of the members of the Technical Study Group. Thus careful attention was given to choosing statistical methods based on assumptions and mathematical approximations that would be readily accepted.

Some of the proposed analyses of Table 1 treated pairwise comparisons between cigarettes, within a given test configuration, as constituting a 2×2 contingency table with four trials for each of two cigarettes. The most extreme difference in such a table has 4 ignitions for one cigarette and 0 for the other. Applying the chi-squared test to a cigarette pair showing a 4 vs. 0 difference in ignitions yields a significance level of 0.005, which would indicate strong evidence of a true difference. Use of the chi-squared test for this situation was criticized because the sample sizes (4 cigarettes of each type tested)

Table 1.—Comparison of Ignition Propensities for Commercial Cigarettes: Numbers of Ignitions in 4 Trials

Cigarette Number†	Test Configuration*									
	1–3	4–5	6	7	8	9	10	11	12–18	TOTAL
2	4	3	4	1	0	1	0	0	0	24
7	4	4	1	4	1	1	0	0	0	27
1	4	4	4	4	0	2	0	0	0	30
4	4	4	4	3	3	0	0	0	0	30
12	4	4	4	4	4	1	0	0	0	33
11	4	4	4	4	3	3	0	0	0	34
3	4	4	4	4	4	2	0	0	0	34
5	4	4	4	4	4	2	0	0	0	34
9	4	4	4	4	4	2	0	0	0	34
10	4	4	4	4	4	3	0	0	0	35
8	4	4	4	4	4	2	2	0	0	36
6	4	4	4	4	4	4	1	1	0	38
TOTAL	48	47	45	44	35	23	3	1	0	389

*The 18 test configurations represented here are fully described in reference [1].

†The 12 cigarette types represent 12 different commercial cigarette packings which are distinguished by name, length, sometimes diameter, whether menthol or non-methol, whether filter or non-filter, and by package type (e.g. soft pack.)

are too small to justify the approximation implied by use of chi-squared tables. To properly accommodate the small sample sizes, use of Fisher's Exact Test had also been suggested. By this criterion, a 4 vs. 0 difference in ignitions corresponds to a significance level of only $p = 0.029$ – still less than the often-used figure of 0.05, but substantially larger than 0.005. While a result with $p = 0.029$ might be considered statistically significant were there only one pair of cigarettes under consideration, the fact that there are 66 possible pairs among the 12 cigarette types further diminishes the strength of evidence implied by obtaining one, or a few, p-values less than 0.05. One commenter pointed out that one should expect to obtain an average of 3.3 "significant differences" by chance when the 0.05 level is used for 66 comparisons.

Another proposed analysis was based on the "TOTAL" column of Table 1. In this analysis, the chi-squared test was used to test the global hypothesis that all cigarettes have equal probabilities of ignition. The result is $\chi^2 = 9.7$ on 11 degrees of freedom, a value that does not approach significance. A fault with this approach is that the row totals from the table do not satisfy the conditions required for validity of the chi-squared test. A practical indication of the problem can be seen by noting that the computed value of χ^2 would change by a large amount if columns 1–3 and 12–18 were not included in the row totals. Alternately, if a sufficiently large number of "flat" columns like 1–3 and 12–18 were *appended* to the table, the computed value of χ^2 for the TOTAL column could be made to approach zero. [Probabilistically, the situation can be characterized by noting that the row totals do not have binomial distributions, as is assumed by use of χ^2, even though the individual entries within the rows are binomial. The large differences in ignition behavior across the 18 test configurations imply that the row totals are distributed like sums of binomial variables, but with *different* probabilities of ignition for each term in the sum. Such a sum is not binomial.]

The approach finally adopted for Table 1 performs a separate analysis for each column (test configuration) of the table.

Lewontin and Felsenstein[3] have shown that the chi-squared test for $2 \times N$ contingency tables is valid when the expected frequencies exceed 1.0 in all cells. Since the data for configurations 8 and 9 (only) satisfy this condition, the chi-squared test was used to analyze each of these columns as a 2×12 contingency table. Configuration 8 shows highly significant differences between cigarettes ($\chi^2 = 36.6, p < 0.001$) while no significant differences are indicated for configuration 9 ($\chi^2 = 12.9$, $p = 0.30$). These two tests can be combined by adding the respective χ^2 values to obtain $\chi^2 = 49.6$ on 22 degrees of freedom, indicating highly significant differences for the combined tests ($p < 0.001$).

What about the remainder of the table? Clearly, no differences between cigarettes are indicated for configurations 1–3, where all cigarettes ignited, or 12–18, where none ignited. The remaining columns, 4–7, 10 and 11, can be evaluated for significant differences in the numbers of ignitions by an exact calculation, analogous to Fisher's Exact Test for a 2×2 contingency table. This procedure is based on the conditional probability distribution of the data given the total number of ignitions,[4] and it requires enumeration of essentially all possible patterns of ignitions and non-ignitions that are consistent with the marginal totals of the observed data. Applying this exact test to columns like 8 and 9, which have relatively large numbers of both ignitions and non-ignitions, would have been a very tedious process. Fortunately the chi-squared test could be used in those cases, and it is known that the two procedures give practically identical results in situations where use of χ^2 is valid.

Of these remaining configurations, columns 6 ($p = 0.003$) and 7 ($p = 0.011$) show significant differences between the cigarettes tested. None of the other configurations show differences that approach statistical signifiance.

Overall, the table shows some signs of interaction between test configurations and cigarette types. But to conclude that these data do not provide evidence that commercial cigarettes differ in ignition propensity would seem unreasonable.

3. A Factorial Experiment

To identify characteristics of cigarettes that could lead to a reduction in ignition propensity, a second experiment was designed using cigarettes of well-characterized composition and construction supplied by the cigarette industry. The experimental cigarettes were custom-made to differ on five design characteristics, each at two levels: tobacco packing density (low and high), cigarette circumference (21 mm and 25 mm), paper permeability (low and high), paper citrate concentration (0.0 and 0.8%), and tobacco type (burley and flue-cured). Small lots of cigarettes were produced corresponding to each of the $2^5 = 32$ possible combinations of the five design factors, and five cigarettes of each type were tested for ignition propensity on upholstered furniture mockups, as described above. Four test configurations were used for the experiments; these were chosen, based on the experience gained from testing commercial cigarettes, to be conditions where differences between cigarettes could be shown. The data resulting from the tests, shown in Table 2, constitute a complete 2^5 factorial experiment for each of the test configurations.

The standard statistical technique for analyzing data from a factorial design is the analysis of variance (ANOVA). However, the hypothesis tests and estimation procedures of ANOVA produce valid statistical inferences only if the data for each cell of the table (defined by the cigarette design and test configuration) come from populations that are at least approximately normal with equal variances. In fact, the data in Table 2 are binomial count data, and thus not well-modelled by equal variance normal distributions: the binomial distribution is asymmetric and the variance depends strongly on the mean. To

Table 2.—Ignition Propensity Results for Experimental Cigarettes

Cigarette Design*					Number of Ignitions, Test Configuration†			
Packing Density	Permeability	Circumference	Citrate Conc.	Tobacco Type	1	2	3	4
E	L	21	N	B	0	1	0	0
E	L	21	N	F	1	3	0	0
E	L	21	C	B	0	3	3	0
E	L	21	C	F	0	5	1	0
E	L	25	N	B	3	2	2	0
E	L	25	N	F	3	1	0	0
E	L	25	C	B	5	5	4	0
E	L	25	C	F	5	3	2	0
E	H	21	N	B	3	4	0	0
E	H	21	N	F	4	5	3	0
E	H	21	C	B	4	5	2	0
E	H	21	C	F	4	5	5	0
E	H	25	N	B	5	5	5	0
E	H	25	N	F	5	5	2	0
E	H	25	C	B	5	5	5	0
E	H	25	C	F	5	5	5	0
N	L	21	N	B	2	5	5	0
N	L	21	N	F	5	5	5	1
N	L	21	C	B	3	5	5	0
N	L	21	C	F	5	5	5	0
N	L	25	N	B	5	5	5	3
N	L	25	N	F	5	5	5	2
N	L	25	C	B	5	5	5	3
N	L	25	C	F	5	5	5	3
N	H	21	N	B	5	5	5	4
N	H	21	N	F	5	5	5	5
N	H	21	C	B	5	5	5	2
N	H	21	C	F	5	5	5	4
N	H	25	N	B	5	5	5	5
N	H	25	N	F	5	5	5	5
N	H	25	C	B	5	5	5	5
N	H	25	C	F	5	5	5	5

*Design Factors:
Packing Density: E = low (expanded tobacco),
N = high (non-expanded tobacco)
Permeability: L = low, H = high
Circumference: 21 = 21 mm, 25 = 25 mm
Citrate Concentration: N = 0.0%, C = 0.8%
Tobacco Type: B = Burley, F = Flue-cured
†Test configurations are described in reference 2.

address this situation, the basic data were transformed using the Freeman-Tukey modification of the commonly used angular transformation.[5] This transformation produces a response variable having a more nearly symmetric distribution with nearly constant variance. The formula for obtaining the transformed response variable is

$$Y = (0.5)\{\text{ARCSIN}[\text{SQRT}(X/6)\,] + \text{ARCSIN}[\text{SQRT}((X+1)/6)\,]\},$$

Table 3.—Significance Probabilities (in Percent) of Design Factors for Experimental Cigarettes (Experimental error was estimated from 4- and 5-way interactions.)

Design Factors	Test Configuration* 1	2	3	4
D (Packing Density)	0.05%	0.10%	0.01%	0.01%
P (Permeability)	0.08	0.23	0.69	0.01
R (Circumference)	0.03	85	2.7	0.01
C (Citrate Conc.)	25	2.0	0.69	7.6
T (Tobacco Type)	8.9	52	56	7.6
Factor Interactions				
D × P	1.55	0.23	0.69	0.01
D × R	0.45	85	2.7	0.01
D × C	43	2.0	0.69	7.6
D × T	45	52	56	7.6
P × R	0.93	52	39	5.5
P × C	45	6.6	85	21
P × T	17	85	7.4	21
R × C	29	85	62	2.2
R × T	8.9	3.8	2.8	2.2
C × T	45	52	62	81

*Test configurations are described in reference 2.

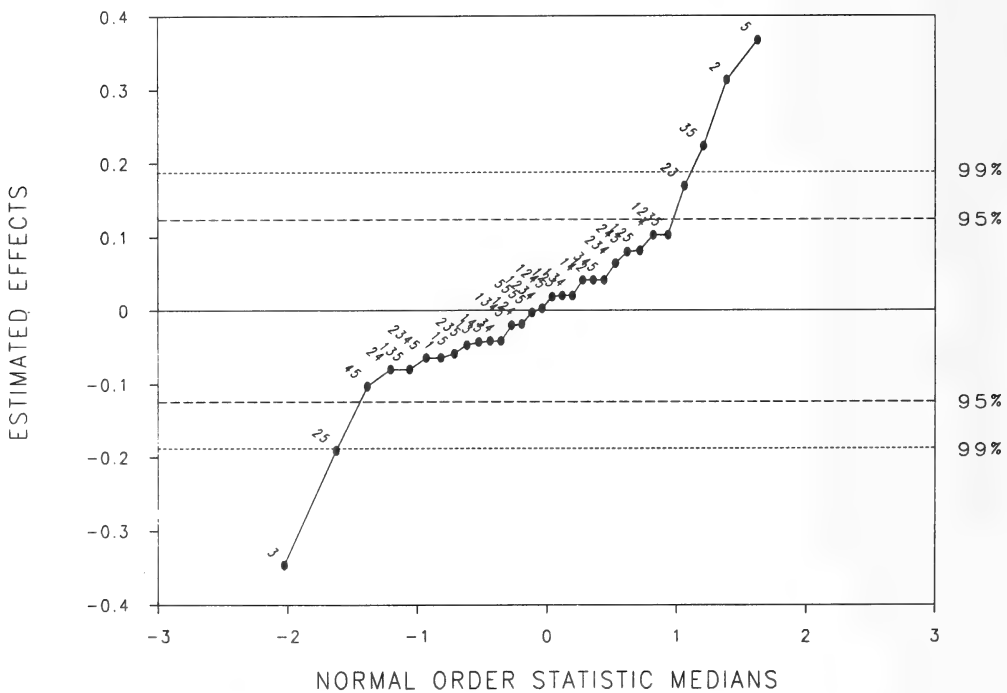

Fig. 1. Normal probability plot of estimated effects from a 2^5 factorial analysis of variance for the data in Table 2 for test configuration 1. The factors are coded as: 1 = Citrate concentration, 2 = Paper permeability, 3 = Packing density, 4 = Tobacco type (burley or flue-cured), 5 = Circumference (21 mm vs. 25 mm). Multiple factor labels represent corresponding interaction effects. The indicated 95% and 99% limits on the plot were computed using the 4- and 5-way interaction terms to estimate experimental error.

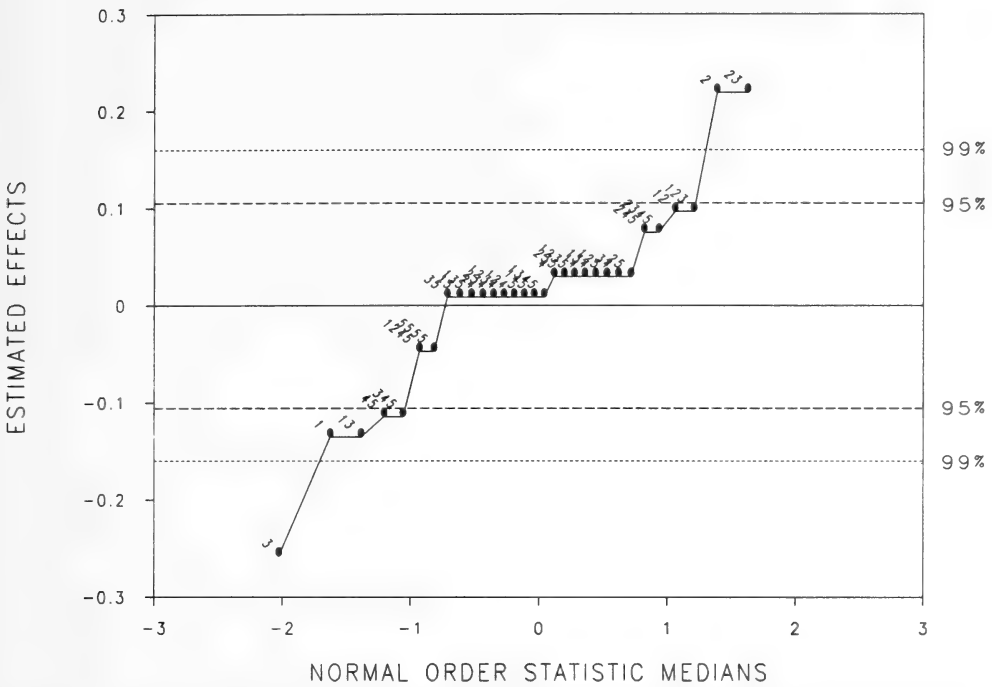

Fig. 2. Normal probability plot of estimated effects from a 2^5 factorial analysis of variance for the data in Table 2 for test configuration 2. The factors are coded as: 1 = Citrate concentration, 2 = Paper permeability, 3 = Packing density, 4 = Tobacco type (burley or flue-cured), 5 = Circumference (21 mm vs. 25 mm). Multiple factor labels represent corresponding interaction effects. The indicated 95% and 99% limits on the plot were computed using the 4- and 5-way interaction terms to estimate experimental error.

where X denotes the number of ignitions (out of 5 trials).

After transformation, the data were analyzed by standard ANOVA methods.[6] The results of the ANOVA are summarized numerically in Table 3. While this summary is adequate to convey the main conclusions of the analysis, the graphical summary of the same results given in Figures 1–4 was much more effective for communicating with the members of the Technical Study Group.

Figures 1–4 present normal probability plots of the estimated factorial effects from the 2^5 factorial analysis of variance for the data of Table 2. Under the null hypothesis that none of the design factors affects ignition propensity, the estimated effects would constitute (approximately*) a random sample from a normal distribution, and so the plotted points would be expected to cluster about a single straight line on the normal probability plot. The indicated 95% and 99% limits on Figures 1–4 were computed using the 4- and 5-way interaction terms in the ANOVA to estimate experimental error. These limits correspond, respectively, to 5% and 1% tests of the hypothesis that the factorial effects are zero.

In addition to providing an effective means of presentation, the normal probability plot analysis has the advantage that it automatically encourages consideration

*"Approximately" here relates to the degree of success achieved by the Freeman-Tukey transformation in achieving approximate normality for the distribution of the response variable Y. For exactly normal data, the estimated effects would follow an exact normal distribution. The figures suggest that normality was more nearly achieved in some cases (e.g. Figure 1) than others (e.g. Figure 2.)

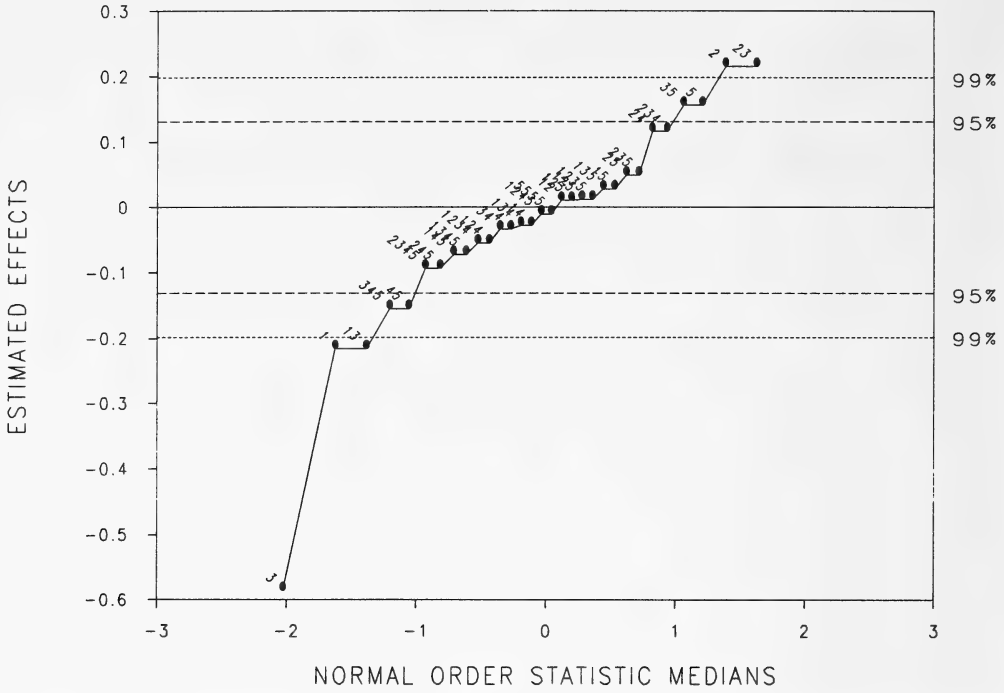

Fig. 3. Normal probability plot of estimated effects from a 2^5 factorial analysis of variance for the data in Table 2 for test configuration 3. The factors are coded as: 1 = Citrate concentration, 2 = Paper permeability, 3 = Packing density, 4 = Tobacco type (burley or flue-cured), 5 = Circumference (21 mm vs. 25 mm). Multiple factor labels represent corresponding interaction effects. The indicated 95% and 99% limits on the plot were computed using the 4- and 5-way interaction terms to estimate experimental error.

of the simultaneous inference aspects of the analysis.[7] This relates to the well-known phenomenon that if, say, 20 hypotheses are simultaneously tested at the 5% significance level, one of the hypotheses will be rejected, on the average, even if all 20 are true. In the present case, the normal probability plots compare the magnitudes of 31 estimated factorial effects. Setting aside the 4- and 5-way interaction effects that were used to estimate experimental error, there are 25 null hypotheses that could be tested (5 main effects, 10 two-factor interactions, and 10 three-factor interactions.) Even if none of the factors or interactions were active (i.e. all 25 null hypotheses were true), the expected number of rejections at the 5% significance level would be 1.25 (5% of 25) simply by chance. However, the probability plot analysis accounts for this phenomenon by drawing attention to only those effects that appear as *outliers* on the plot, whether or not they lie outside the 95% or 99% bands. For example, in Figure 4, the 1 × 3 × 5 and 3 × 4 × 5 interaction effects lie very close to the 99% bands, corresponding to p-values near 1%. However, the fact that those effects do not appear as outliers on the normal probability plot indicates that those p-values should not be interpreted as strong evidence of active effects.

To summarize the results of this experiment, Figures 1–4 and Table 3 show that two factors, namely, packing density and paper permeability, were consistently highly significant across all four test configurations. Two additional factors, circumference and citrate concentration, showed clear significance in two of the test configurations. The factor for tobacco

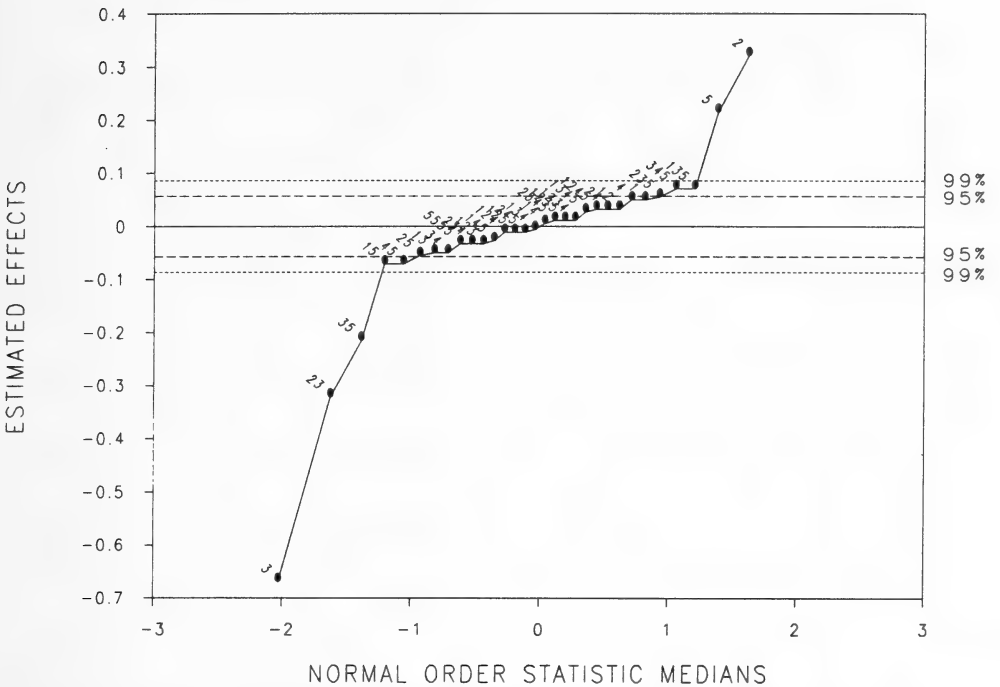

Fig. 4. Normal probability plot of estimated effects from a 2^5 factorial analysis of variance for the data in Table 2 for test configuration 4. The factors are coded as: 1 = Citrate concentration, 2 = Paper permeability, 3 = Packing density, 4 = Tobacco type (burley or flue-cured), 5 = Circumference (21 mm vs. 25 mm). Multiple factor labels represent corresponding interaction effects. The indicated 95% and 99% limits on the plot were computed using the 4- and 5-way interaction terms to estimate experimental error.

type (i.e., burley vs. flue-cured) did not show a significant effect on number of ignitions in any of the cases.

The interactions among these factors were frequently significant whenever the main effects were. Significant interactions indicate that the magnitude of the effect on ignition propensity for a given factor is not constant across the levels of the interacting factor. For example, on test configuration number 1, the significant interaction between Packing Density and Circumference indicates that the effect of Packing Density on ignition propensity is different in magnitude for the smaller circumference cigarettes than for the larger circumference cigarettes in the experiment. Detailed study of the data suggests that many of the significant interactions can be explained by the single fact that the maximum number of ignitions per test condition could not exceed five in this experiment, thus limiting the possible magnitudes of the estimated effects.

In practical terms, the relation between the cigarette design parameters studied and ignition propensity can be summarized as follows: The most significant factor was low packing density, achieved in this case by using expanded, large particle size tobacco. Low packing density apparently lowers ignitions because the available fuel per unit length is reduced. Another very influential factor is the use of low permeability paper, which reduces ventilation to the tobacco column. Cigarettes with a 21 mm circumference showed evidence of reducing ignitions in three of the four test configurations. Again, this may be due to the reduction in fuel (less tobacco and paper) per unit length. The effect of citrate, which is added to ciga-

rette paper to regulate the paper burn rate and to obtain ash of the desired appearance, did not show a consistent effect on ignition propensity. And tobacco type consistently showed no effect on ignitions.

4. Acknowledgements

The author is grateful to Ms. Susannah Schiller and Dr. Richard Gann, of the National Institute of Standards and Technology for their careful reading of, and comments on, this manuscript.

References Cited

1. **Krasny, J. F. and Gann, R. G.** 1986. Relative propensity of selected commercial cigarettes to ignite soft furnishings mockups. NBSIR 86-3421, [U.S.] National Bureau of Standards.
2. **Gann, R. G., Harris, R. H., Jr., Krasny, J. F., Levine, R. S., Mitler, H. E., and Ohlemiller, T. J.** 1988. The effect of cigarette characteristics on the ignition of soft furnishings. NBS Technical Note 1241, [U.S.] National Bureau of Standards.
3. **Lewontin, R. C. and Felsenstein, J.** 1965. The robustness of homogeneity tests in $2 \times N$ tables. *Biometrics*, *21*: 19–33.
4. **Plackett, R. L.** 1981. *The Analysis of Categorical Data*, 2nd edition. Macmillan, New York. Section 6.3.
5. **Freeman, M. F., and Tukey, J. W.** 1950. Transformations related to the angular and the square root. *Annals of Mathematical Statistics*, *21*: 607–611.
6. **Box, G. E. P., Hunter, W. G., and Hunter, J. S.** 1978. *Statistics for Experimenters*. John Wiley and Sons, New York. Chapter 10.
7. **Daniel, C.** 1959. Use of the half-normal plot in interpreting factorial two-level experiments. *Technometrics 1*: 311–341.

Journal of the Washington Academy of Sciences,
Volume 78, Number 4, Pages 333–338, December 1988

Environmental Statistics

N. Phillip Ross and Gilah Langner

ABSTRACT

Centralized environmental statistics are needed to provide credible answers to complex environmental issues. Environmental hazards arise from multiple sources, transcending geopolitical boundaries, and challenging our limited understanding of how ecosystems operate. Meanwhile, environmental data are scattered across a number of federal agencies. To deal coherently with the complex environmental issues, there must be sustained, planned, long-term data collection and analysis efforts.

Environmental issues are becoming an increasingly vital concern of our society. Polls taken early in last year's presidential campaign showed that high percentages of the American population consider the quality of the environment to be a national priority. More and more, Americans are recognizing the interdependence of their quality of life with the quality of the environment in which they live.

However, if one were to ask a representative group of American citizens whether (and in what ways) the state of the environment had improved or deteriorated in recent years, most of the answers received would likely fall in the "Don't Know" category. Common sense and intuition can no longer be relied on to provide an accurate indicator of the quality of any individual environmental resource, let alone of the state of the environment as a whole. One cannot step outside and sniff the air to get an accurate feel for air quality in major cities; indoors, one cannot rely on one's senses to determine whether a house is contaminated by radon, a radioactive gas that cannot be seen, smelled, or tasted.

Nor would the experts do much better than the layperson in answering the same question. Many old environmental problems have been addressed, but new ones have sprung up to take their place. On balance, it is hard to determine how well we are doing in keeping our environment safe and healthy and whether there has been a net improvement in recent years.

Need for Centralized Environmental Statistics

One major reason that it is so difficult to come up with credible answers is that we are not collecting and integrating the type of quality-controlled, scientifically-rigorous data that are necessary in order to develop a coherent picture of environmental trends. The Environmental Pro-

Correspondence should be sent to: Dr. N. Phillip Ross, Chief, Statistical Policy Branch, Office of Policy, Planning and Evaluation, PM232, U.S. EPA, 401 M. Street S.W., Washington, D.C. 20460

Dr. Ross is Chief of the Statistical Policy Branch Office of Policy, Planning and Evaluation, U.S. EPA, Washington, D.C.

Ms. Langner is President of Stretton Associates, a Washington-based consulting firm specializing in policy analysis and communications.

tection Agency (EPA) alone has an information collection budget of over 120 million hours and spends half a billion dollars annually on data collection. A variety of other federal agencies are actively involved in different aspects of environmental data collection as well. However, within the federal government, there is no dedicated bureau or agency with responsibility for the collection, integration, and analysis of the environmental data necessary for setting priorities and determining directions on a national scale.

Statistical agencies or bureaus are commonplaces in almost every major federal department. Such agencies include: the Bureau of Economic Analysis and the Bureau of the Census at the Department of Commerce; the Bureau of Justice Statistics at the Justice Department; the Bureau of Labor Statistics at the Labor Department; the National Center for Education Statistics in the Department of Education; the National Agricultural Statistics Service and the Economic Research Service in the Department of Agriculture; the Statistics of Income Division of the Internal Revenue Service; the National Center for Health Statistics in the Department of Health and Human Services; the Social Security Administration's Office of Research and Statistics; the Energy Information Administration in the Department of Energy; and the Division of Housing and Demographic Analysis in the Department of Housing and Urban Development.

As Paul Portney pointed out in a recent article,[1]

> While no one would argue that our current measures of population or economic activity are exact, it is impossible to imagine modern government operating in their absence. Indeed, these measures drive important federal grant and entitlement programs; they also help trigger, and then measure the success of, major tax and spending programs, monetary policies, and even foreign policy decisions. . . Simply put, we have not a *single* data series for the environment that goes back as far as even the most recently established of the economic and demographic series . . . nor one that is subject to the same quality control, careful measurement protocols, or subsequent thorough analyses.

The absence of an active, centralized focus for environmental statistics in this country is felt on an international scale as well. The United States is one of the few developed nations that lack a centralized governmental agency responsible for collecting and integrating environmental data and for producing a base of quantitative information to support an annual State of the Environment Report. The time has come for environmental statistics to emerge as the next major focus of federal statistical data collection.

Defining Environmental Statistics

What do we mean by environmental statistics? Beyond the simple tabulation of environmental data, the environmental statistics process we are speaking of is much broader. It concerns itself with evaluating the state of the biosphere (environmental media and the fauna and flora that inhabit media) and its changes over time. Environmental statistics thus involves identifying information needs, designing appropriate data collection activities (such as ambient monitoring and population surveys), ensuring the quality of the data, and conducting statistical analysis and interpretation of data in order to produce meaningful indicators of the state of the environment.

According to the Statistical Office of the United Nations, environmental statistics:

(a) cover natural phenomena, human activities that exert impacts on the environment, and the impacts them-

selves on the environment and on human living conditions;

(b) refer to the media of the natural environment, i.e., air, water, land/soil, and to the man-made environment which includes housing, working conditions, and other aspects of human settlements;

(c) synthesize data from different subject areas and statistical sources to facilitate integrated socio-economic and environmental planning and policies.[2]

In examining the current roster of environmental issues, we can identify three primary reasons why an enhanced capability to produce environmental statistics will be an indispensable aid to environmental decision-making in the coming years.

(*1*) *Increased Complexity of Environmental Problems and Solutions*

Environmental problems of the 1980's have taken on a subtlety and complexity that did not characterize the air and water pollution problems of previous decades. To be sure, there is still no lack of acute emergencies—chemical fires, toxic waste dumps lacking security, and trucks carrying hazardous materials overturning on the highway. However, many of our most pressing environmental problems involve us in new and uncertain terrain.

For example, setting standards for chemicals is an increasingly complex process, involving the assessment of the effects of long-term exposure to pollutants at levels in the parts per billion and trillion. There are vastly greater numbers of chemicals in production than were ever contemplated when much of our original environmental legislation was put in place. It is becoming clear that human activities are influencing and changing the environment—the balance of ecological systems; the availability of resources, the climate, the ozone levels. As each day goes by we are confronted with more and more potentially serious problems.

The questions facing environmental decision-makers are: which of these problems are most important; which ones should receive attention and resources; which ones require immediate action and which can wait; how do we educate and motivate the population to change its mode of interacting with the environment; how do we convey information about real health risks, without scaring the population and without encouraging complacency?

As environmental problems become less visible to our senses, the need arises for sophisticated approaches to the detection and monitoring of pollutants on human health and the environment. The physical and engineering sciences have rapidly progressed to meet the requirement of sophisticated detection and monitoring. However, the data collection and statistical methods needed to appropriately assess the impacts of newly discovered environmental problems have not kept pace. The costs of collecting adequate environmental data to determine health effects and to intelligently manage natural resources are becoming astronomical. Statistical approaches to data collection and interpretation are the only solution.

A major difficulty in looking to the future is that we have so little information available from the past. If we started today to coordinate our environmental data collection activities across the federal government, it would still be ten or more years before we could start to examine trends and to use those trends in the decision process.

Certain sources of environmental data already exist that have the potential to provide insights into a number of environmental processes. Unfortunately, these sources often represent "encountered data" that cannot be considered a random sample from the original population.[3] A variety of situations can give rise to encountered data—instances where

the only feasible sampling procedure gives unequal chances to the population units or where the only data available are historical data sets from diverse sources. While a great deal of work has been done on how to use encountered data in the context of animal sightings and monitoring of marine resources, this work must now be extended to a broader array of environmental modeling and monitoring situations.

Finally, many of the environmental issues now facing us transcend geopolitical boundaries—acid rain, the greenhouse effect, sea-level rise, ozone depletion, and the continued loss of species on the planet, to name a few. From an international perspective, it is time for the United States to join its Western neighbors in developing a comprehensive long-term base of environmental information. Although the United States has taken the lead in developing research programs in some of these areas, we have not yet committed ourselves to the type of statistically designed monitoring programs that are needed to support hard decisions on subjects involving national and global environmental effects.

(2) *Need for Integration*

A second compelling reason for giving more attention to environmental statistics arises out of a perceptible change in the prevailing image of the environment. From viewing the world as a warehouse of resources available for our benefit, we are starting to recognize the exhaustible nature of certain resources, such as oil and wood, and even certain living species. What once appeared as "infinite" is now clearly finite, and in some cases is quickly being depleted. There is growing talk of the interrelationships among organisms that sustain life, and even of the life-sustaining nature of the planet itself.

The logical corollary of such an approach is the need to develop an integrated view of environmental media. We can no longer deal separately with each environmental medium—air, water, and soil—and ignore the "environmental merry-go-round" effect of shifting pollutants from one medium to another. Environmental data bases need to reflect this recognition and allow us to examine and control for "cross-over" effects.

Here, part of the problem is that information and responsibilities for various segments of the environment are diffused throughout the Federal Government. While EPA deals with problems resulting from pollution of the environmental media, the stewardship of our environmental resources is in other hands. Thus, the Department of the Interior maintains primary responsibility for public lands, minerals, national parks, and endangered species; the Department of Agriculture handles forestry, soil, and conservation; and the Department of Commerce is involved in oceanic and atmospheric monitoring and research.

As a result, environmental media are very often addressed separately from the environmental resources they support. There remains a serious need for high quality data suitable for conducting long-term evaluations of the state of the environment, considered as a whole. Our ability to evaluate environmental progress in the past and set priorities for the future is compromised by a lack of appropriate, integrated trend indicators.

Within EPA itself, there are limitations in what we can do because many of the Agency's data bases are primarily oriented towards furthering EPA's compliance monitoring responsibilities. Compliance data do not necessarily lend themselves to analyses of long-term trends in the environment. Take, for example, the issue of waste reduction. Although EPA maintains numerous data bases with facility reports on environmental discharges, at the present time there is no single data base that can be used to measure waste reduction efforts. Even if it were possible to "cross-walk" from the facilities in one data base to those in another (which it generally is not), the data

bases are designed to measure different things in different units. Measurement of concentrations of hazardous chemicals in waste streams cannot provide useful data on waste generation.

On a wider ecological level, we need to collect new environmental data in conjunction with our environmental and ecological models. Using these models, we are beginning to develop an understanding of how ecosystems work. We need better ways of recognizing what constitutes a healthy ecosystem and better monitoring skills that will provide early warning signs of damage or injury to an ecosystem.

(3) *The Public's Right and Need to Know*

A third need for improved environmental statistics is the public's right to know more about the environment. This goes well beyond the legislated "right to know" provisions in Title III of the Superfund Amendments and Reauthorization Act of 1986. EPA has a clear responsibility to make a wide range of environmental statistics available to the public both to fulfill the mandate of democratic government and to establish the credibility of its decision-making on environmental issues.

The public's need for environmental information is significant in numerous sectors of society. Accurate and up-to-date environmental information is important for public decision-making at the state, county, and local levels. It is equally important for eliciting responsible natural resource management decisions from industry. Increasingly, we are relying on industry to voluntarily cooperate in conserving environmental resources and to make farsighted management decisions. These decisions must be based on an adequate information base. A 1984 report by the World Wildlife Fund,[4] for example, identifies 11 corporate decisions that require the use of resource information, including strategic direction, market research, resource acquisition, production capacity, plant siting, plant design, environmental compliance, production and materials purchasing, research and development, bank lending, and investment recommendations.

Members of the public also need a reliable, quality-controlled, credible source of environmental information to help them evaluate the plethora of health risks that appear almost daily in the news, and to help them make judicious personal decisions, whether that involves testing a home for radon or installing water treatment devices. A related need is for clear and usable explanations of environmental statistics. Given the complexity of our environment, the advanced scientific tools being used, and the probability concepts in which much of our risk information is couched, the development and publication of environmental statistics alone will not necessarily mean that the public has "access" to the information. Communication of the significance and context of the information is essential to satisfying the public's right and need to know.

The increasing complexity of environmental problems, the need for integrated environmental approaches, and the need to provide more environmental information to the public, all point to the importance of giving serious attention to environmental statistics. One way of solving the problems and deficiencies outlined in this article may be to create a federal Environmental Statistics Agency or Bureau. Recent months have seen a renewed interest in this idea, with proposals and recommendations coming from a variety of sources. At the Environmental Protection Agency, a Science Advisory Board panel chaired by Alvin Alm recently recommended the creation of a new ecological research institute as part of a renewed EPA commitment to long-term environmental research. The institute would have responsibilities for ecological research, environmental monitoring, and statistics. Whatever the approach, it is apparent that

an enhanced environmental statistics capability will be vitally important to the conduct of environmental protection in this country and in the international sphere in the coming decade.

References Cited

1. **Paul R. Portney,** "Needed: a Bureau of Environmental Statistics," in *Resources,* Resources for the Future, Winter 1988.
2. **Bartelmus, Peter,** "Environmental Statistics: Systems, Frameworks and International Approaches" in Society of American Foresters/et al, Intl Renewable Resource Inventories for Monitoring Conf, Corvallis, OR, Aug. 15–19, 1983, pp. 524(5).
3. **Hennemuth, R. C., G. P. Patil, and N. P. Ross,** "Encountered Data Analysis and Interpretation in Ecological and Environmental Work: Opening Remarks." Presented at the annual ASA meeting, San Francisco, 1987.
4. *Corporate Use of Information Regarding Natural Resources and Environmental Quality,* prepared by Train, Russell E., World Wildlife Fund for the Council on Environmental Quality, 1984.

Journal of the Washington Academy of Sciences,
Volume 78, Number 4, Pages 339–353, December 1988

Some Uses of a Modified Makeham Model to Evaluate Medical Practice

R. Clifton Bailey*

Health Standards and Quality Bureau
Health Care Financing Administration

ABSTRACT

The modified Makeham survival model describes the time course of a medical intervention or illness in many cases. A precise understanding of the time course may be used to evaluate the follow-up time for special studies, to know when the available follow-up is not adequate, and to balance follow-up time against the number of cases. Also a knowledge of the time course may be used to evaluate long-term and short-term risk factors. Long-term and short-term risk factors are studied separately using the widely available Cox proportional hazards model. This is compared with a fully integrated model in which the modified Makeham is used with concomitant variables in each of three structural parameters. The examples rely on data collected by the Health Care Financing Administration to evaluate the effectiveness of medical interventions on the course of illness.

*The opinions expressed in this paper are those of the author and do not necessarily reflect the opinions or policies of the Health Care Financing Administration.

Correspondence should be sent to: R. Clifton Bailey, Health Standards and Quality Bureau, Health Care Financing Administration, 2-D-2 Meadows East, 6325 Security Blvd., Baltimore, MD 21207.

Introduction

The value of data in making decisions and understanding medical practice is a long standing tradition in medicine. This article describes some statistical tools which can be used to advance our understanding of medical practice. The main focus is on a modified Makeham survival model. This basic model describes the time course of a medical intervention or illness in many cases. A precise understanding of the time course may be used to evaluate the follow-up time for special studies and to know when the available follow-up is not adequate. Examples of the computations demonstrate the balancing of follow-up time and number of cases. Additional examples demonstrate how a knowledge of the time course may be used to isolate long-term and short-term risk factors. Two approaches are described. First, the long-term and short-term risk factors are studied separately using the widely used Cox proportional hazards model. This approach is compared with a fully integrated modified Makeham model with concomitant variables in each of three structural parameters. The examples in this paper rely on data collected as part of an effort by the Health Care Financing Administration to evaluate the effectiveness of medical interventions on the course of illness.

Recognition of Statistical Methods in Extracting Information of Value from Data

Before beginning a technical description of the Makeham model promised above, it is worth setting the stage with a few reminders about the role of statistics in extracting information from data.

In his presidential address before the First Indian Statistical Conference, 1938, R. A. Fisher[1] reminded his colleagues that "in the original sense of the word, 'Statistics' was the science of Statecraft." He points out that the task of providing public information as a function of official statistics ". . . enables the public, *if it will*, to size up its own problems. The Socratic dictum 'know thyself' is applicable even more to peoples than to individuals." Also he notes that with the development of a theory of estimation and an understanding of the magnitude and the nature of the sampling errors, "The whole tone of the subject has been altered. The Statistician is no longer an alchemist expected to produce gold from any worthless material offered him. He is more like a chemist capable of assaying exactly how much of value it contains, and capable also of extracting this amount, and no more. In these circumstances it would be foolish to commend a statistician because his results are precise, or to reprove because they are not. If he is competent in his craft, the value of the result follows solely from the value of the material given him. It contains so much information and no more. His job is only to produce what it contains." "To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."

There are many notable examples of the need to extract useful information from data. For example, Walter Shewhart[2] recognized the value of information generated by industrial processes and developed methods to chart measurements sequentially—the control chart. An example closer to the subject of this paper is found in the medical arena. Florence Nightingale[3] noted inconsistency in recording the number of deaths at military hospitals during the Crimean War. At home she also found that English hospital records followed no common nomenclature or standard. She worked with Dr. Farr, of the Registrar-General's Office to prepare standard lists for classes and orders of diseases and model Hospital Statistical Forms to "enable us to ascertain the relative mortality in different hospitals, as well as of different diseases and

injuries at the same and at different ages, the relative frequency of different diseases and injuries among the classes which enter hospitals in different countries, and in different districts of the same countries." Furthermore, use of the proposed forms "could enable the mortality in hospitals, and also the mortality from particular diseases and injuries, and operations to be ascertained with accuracy; and these facts, together with the duration of cases, would enable the value of particular methods of treatment and of special operations to be brought to statistical proof." For her efforts as a "passionate statistician," Florence Nightingale was made Honorary Member of the American Statistical Association.[4]

Outcomes in Medical Statistics

In evaluating medical practices and interventions, an evaluation of chances of death seems to be an ultimate concern. The history of the life table and of medical statistics to a large degree centers around techniques for statistical summary of mortality.[5] Actually the techniques developed for mortality statistics have broad application. For example, in engineering, failure time is often used as a measure of outcome when we want to know long we can expect a car, a light bulb, or a personal computer to last. In all of these problems, classification of the failure point and items under test are crucial to the utility of the studies being conducted. Many problems in applying statistical methods to medical data revolve around classifying and recording conditions present, procedures and interventions used, and the outcomes. The outcomes of interest in health care can be classified as mortality, morbidity, disability, and expenditure. In this paper, the focus is on mortality. However, methods for mortality analysis have a role whenever we measure an outcome by the time or duration for an occurrence such as a readmission, relapse or failure of a medical intervention or procedure.

The Modified Makeham Model for Survival Analysis

A modification of the Makeham model was proposed by Bailey *et al.* to evaluate kidney graft survival.[6,7] The modified Makeham model has a decreasing hazard or risk function of the form

$$r(t) = \alpha \exp(-\gamma t) + \delta$$

where α is the initial excess risk, δ is a long-term risk and γ determines the rate for the decay of the initial excess risk. The risk function for the Makeham model as it is commonly found in the actuarial literature (Jordan, 1967)[8] has the form

$$r(t) = \alpha\rho^t + \delta.$$

In actuarial applications to human mortality data, ρ is greater than 1 and the risk function is increasing with time, t. This corresponds to an increased risk of death with older age over a lifetime.

When a constant δ is added to a Gompertz force of mortality to obtain the form shown above the model is referred to as Makeham's modification of the Gompertz even though Kurtz (1930)[9] attributes the idea for the use of the additive constant δ to Gompertz himself. In our modified form of the Makeham, ρ is less than 1, since $\rho = \exp(-\gamma)$. The decreasing risk form of the Makeham works well when there is a high risk intervention followed by a period of recovery during which the excess risk of the intervention diminishes with time.

The risk function is also known as the hazard function or the force of mortality. The risk curve shows us at what rate failures occur relative to the number of survivors at any time, t. The corresponding expression for the proportion surviving at time t is

$$S(t) = \exp\left\{-\int_0^t r(\tau)\, d\tau\right\}.$$

or

$$S(t) = \exp\{-\{\delta t + (\alpha/\gamma)[1 - \exp(-\gamma t)]\}\}.$$

The modified Makeham model also includes concomitant variables in each of its positive parameters as follows:

$$\alpha = \exp(\alpha_0 + \alpha_1 x_1 + \cdots + \alpha_i x_i + \cdots + \alpha_k x_k)$$

$$\beta = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i + \cdots + \beta_k x_k)$$

$$\gamma = \exp(\gamma_0 + \gamma_1 x_1 + \cdots + \gamma_i x_i + \cdots + \gamma_k x_k).$$

This survival model has great utility for evaluation of survival data following some high risk event such as surgery. The following description of this and related tools will be applied to examples using Medicare data to evaluate the effectiveness of medical interventions.

Time Course of the Makeham Model: Its Role in Statistical Studies

The time course of the Makeham model is important for two important statistical activities: the design of follow-up studies, the evaluation of factors affecting long-term and short-term risk.

The following sections will attempt to isolate the long-term from the short-term components of the modified Makeham model. In evaluating the design for follow-up studies, we focus our design efforts on getting the information to evaluate the long-term risk components. This approach is in the spirit conveyed in presidential address cited above[1] when Professor Fisher recalls that his teacher Professor Whitehead of Cambridge used to say: "The essence of applied mathematics is to know what to ignore." With sufficient lapse of time we can ignore the short-term components. Then an evaluation of the follow-up time required for a study is reduced to a simpler problem in which the risk function is a constant once the short-term risk has become negligible. As you will see, we don't really ignore the short term components. Instead we use our best estimates of these components to recast a complicated problem in terms of a simpler problem.

The second aspect of isolating the long-term from the short-term is to evaluate complex data on medical intervention. An estimate of the time for the short-term hazard to become negligible relative to its initial value allows us to use the popular Cox proportional hazards model to better understand the role of important risk factors. As before the estimates are based on estimates for the parameters of a modified Makeham model with no concomitant variables. The full Makeham model with concomitant variables provides the more satisfying approach when we have a high risk intervention followed by a period of recovery or return to stable or constant risk. The full Makeham model provides in one frame work a means to jointly evaluate the proper balance of various risk factors. The result is a more complex, comprehensive summary of complex data.

Design of Survival Studies When the Makeham Model is the Appropriate Survival Model

To determine an appropriate follow-up time for a study, there must be sufficient time to estimate the long-term risk component δ when the Makeham model is the appropriate survival model. Basically information on δ is not available until the short-term risk, $\alpha \exp(-\gamma t)$, has become negligible. A reasonable approach is to

find the short-term time, T_s, such that the hazard

$$r(t) = \alpha \exp(-\gamma t) + \delta$$

is equal to a value close to the long-term risk itself. For this example, we use the value $r(t) = 1.1\,\delta$. That is, the short-term effect is diminished to within 10% of δ. This means that

$$\alpha \exp(-\gamma t) + \delta = 1.1\,\delta$$

or equivalently,

$$\alpha \exp(-\gamma T_S) = 0.1\,\delta.$$

The solution for the time at which this risk is achieved gives

$$T_S = -(1/\gamma)\ln(0.1\,\delta/\alpha).$$

Note that T_S is negative when α is less than $0.1\,\delta$. In this case simply use $T_S = 0$. Now, as an approximation, we have the constant risk form

$$r(\Delta T) = \delta$$

and the information on δ begins accumulating at time, T_S according to the expression for the information

$$I(\delta, \Delta T) = 1 - \exp(-\delta\,\Delta T)$$

where ΔT is the time lapse after time T_S. When ΔT is zero, the above equation implies there is no information about δ. As ΔT approaches infinity, or symbolically

$$\Delta T \longrightarrow \infty,$$

the information on δ approaches 1 or 100%. Not only is $I(\delta, \Delta T)$ the information on δ, but also it is related to the survival function for the constant hazard function where

$$I(d, \Delta T) = 1 - S_T(\Delta T)$$

where $S_T(\Delta T)$ is the survival curve for those cases not failed at T_S. This means that information on the long-term risk accumulates as failures occur after time T_S. To get an idea of how this works, see Figure 1. For purposes of illustration, consider a value of $\delta = 0.2$ per year. Then to get 50% of the potential information on δ would require a study to extend

$$\Delta T = -(1/\delta)\ln(0.5)$$
$$= 0.693/\delta$$

years beyond T_S or 3.5 years when $\delta = 0.2$ per year. A period of one year beyond T_S would provide 18% of the information on δ. The idea is to design a study based on getting the more difficult, long-term information for the parameter δ. In taking this approach, we rely on the very simple, pragmatic notion, that it is more difficult to obtain reliable long-term information. By focusing the problem on the long-term risk parameter, δ, we simplify a complex problem and insure an adequate design for a full evaluation of long- and short-term risk factors. On this basis we evaluate sample sizes for follow-up study and illustrate the balance between the number of cases and the follow-up time.

Sample Size for a Follow-up Study

To plan a follow-up study, we must evaluate a sample size or number of cases to be included in our study. To do this we reduce the complex modified Makeham model to the simpler model with a constant risk function by first estimating the time for our model to be reduced to this simpler form. This adaptation permits us to borrow techniques already established for the simpler problem. Essentially we must find the sample size required to estimate δ. For example, this can be computed from an approximate formula in Gross and Clark (1975)[10] for the confidence limits on δ. Use the formula for the $100(1 - \alpha)\%$ confidence limits on δ

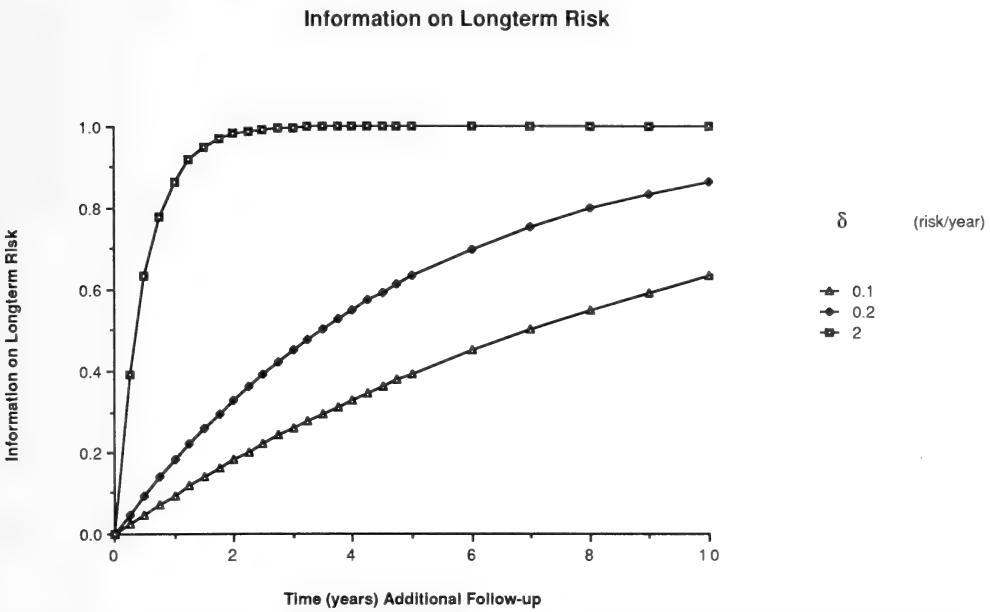

Fig. 1. Information accumulates on the long-term risk, δ, as the follow-up time increases. The time shown in the figure must be added to the time required for the short-term risk to become negligible.

$$\hat{\delta}\left[1 - \frac{z_{(1-\alpha/2)}}{\sqrt{m}}\right] < \delta < \hat{\delta}\left[1 + \frac{z_{(1-\alpha/2)}}{\sqrt{m}}\right]$$

to find the number of cases m needed to obtain an estimate within $100k\%$ of δ. The result

$$m = \left[\frac{z_{(1-\alpha/2)}}{k}\right]^2$$

is the number of cases, assuming infinite follow-up time and full information on δ. Then we adjust the sample size to account for the expected follow-up time. If m' is the actual sample size at time T_S, the effective sample size for a total follow-up time of $\Delta T + T_S$ is

$$m = m'(1 - \exp(-\delta\,\Delta T)).$$

To adjust the sample size m for this follow-up time, first compute m then

$$m' = m/(1 - \exp(-\delta\,\Delta T)).$$

This formula allows us to evaluate options to use more cases or a longer follow-up time in survival studies. An additional adjustment to the sample size accounts for the expected number of cases lost before T_S. The adjusted sample size is

$$m'' = m'/S(T_S)$$

where

$$S(T_S) = \exp\{-\{\delta T_S + (\alpha/\gamma)[1 - \exp(-\gamma T_S)]\}\}$$

is the Makeham survival curve.

For example, one year of follow up from T_S with 95% confidence would yield a sample size of

$$m = (1.96/0.1)^2 = 384$$

would provide an estimate of δ within 10% of $\delta(k = 0.1)$.

With $\alpha = 10.6, \gamma = 39.1$ and $\delta = 0.183$, (parameter estimates from the time course of medicare heart attack data),

$$T_S = -(1/\gamma)\ln(0.1\delta/\alpha)$$
$$= -(1/39.1)\ln((0.1)(0.183/10.6))$$
$$= 0.16 \text{ years},$$

Using the m above, we find that one year of follow-up from the 0.16 years or a total follow-up of 1.16 years, gives

$$m' = m/0.167 = 2{,}299$$

and

$$m'' = m'/0.741 = 3{,}103 \text{ cases.}$$

The calculations for two years of follow-up beyond 0.16 years for a total follow-up of 2.16 years gives

$$m' = m/0.306 = 1{,}255$$

and

$$m'' = m'/0.741 = 1{,}694 \text{ cases.}$$

In contrast, the time course of heart failure with $\alpha = 2.37$, $\gamma = 10.77$ and $\delta = 0.377$, has $T_S = 0.38$ years.

Then for 1 year of follow-up beyond 0.38 years or 1.38 years

$$m' = 384/0.314 = 1223$$

and

$$m'' = 1223/0.698 = 1752 \text{ cases,}$$

and for two years of follow-up beyond 0.38 years for a total of 2.38 years gives

$$m' = 384/0.530 = 725$$

and

$$m'' = 725/0.698 = 1039 \text{ cases.}$$

These calculations clearly illustrate the trade offs between follow-up time and sample size that can be used in designing studies. These computations are important even in retrospective studies based on administrative data because they provide insight into the number of years of back records that will be needed for a study. A careful evaluation of the required sample size for a study must include other factors, such as the expected recruitment rate and an allowance for cases lost to follow up. A fuller discussion of these issues can be found in Meinert.[11]

Application of The Makeham Model in the Evaluation of Data for the Effectiveness and Use of Medical Interventions—Long-Term Versus Short-Term Risk Factors

As part of a project developed in late 1985 and early 1986, the Health Care Financing Administration (HCFA) and the Peer Review Organizations (PROs) undertook a project to develop data on the patterns of use and the effectiveness of medical interventions. This project is part of a comprehensive effort undertaken by HCFA[13] to assess and stimulate improvement in the quality of medical care rendered to Medicare beneficiaries. The effort is in four segments:

1. Monitoring of trends over time in use and outcome of medical interventions.
2. Analysis of variation in use and outcome over geographic areas and providers of care.
3. Detailed investigation of the patterns of medical practice that underlie the time trends and variations in outcomes among localities.
4. Feedback and education to exchange and improve information.

The examples which follow are derived from the third segment in which detailed investigations are being conducted for cardiovascular problems. For this inves-

Table 1.—Maximum likelihood Estimates for the Structural Parameters of the Makeham Model

	Heart Attack	Heart Failure
Number of cases	3152	3274
α (per year)	10.6	2.37
γ (per year)	39.1	10.77
δ (per year)	0.183	0.377
T_L (days)	22	78

tigation, a sample of just over 3100 cases for each condition is available for analysis. As a first step in understanding the short- and long-term course of patient survival following a medical intervention, it is useful to fit the Makeham model with decreasing hazard to the data. The result of this fit can be used to graph the hazard over time, estimate the time course of the risk following intervention, and engage in a more thorough examination of the role of key risk factors over the time.

To demonstrate these ideas, maximum likelihood estimates were obtained for the Makeham model for two data sets. One has 3152 heart attack cases and the other 3274 cases with heart failure. The results of this fit are shown in Table 1.

The risk curves are shown in Figure 2 and the corresponding survival probabilities are shown in Figure 3.

Where do the Survival Curves Cross?

One important property of the Makeham model is that survival curves can cross

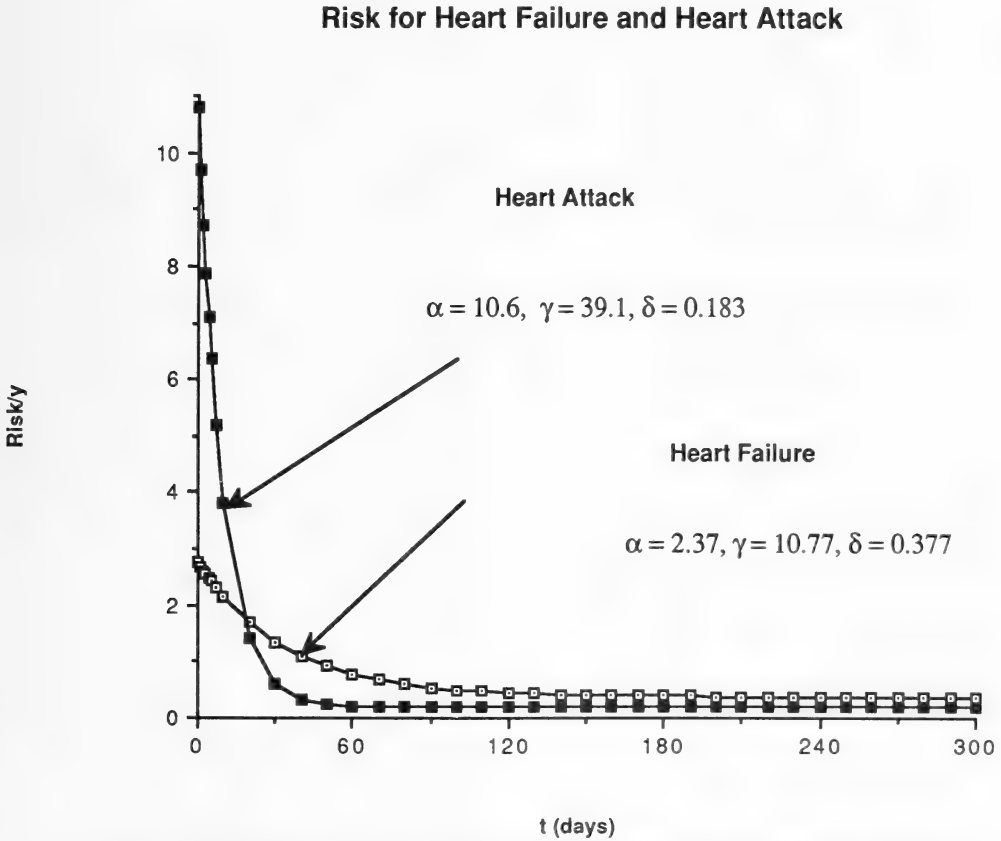

Fig. 2. The modified Makeham risk functions for the heart attack and heat failure data. Note the crossing of the hazard (risk) functions at 0.04832 year or 17.6 days.

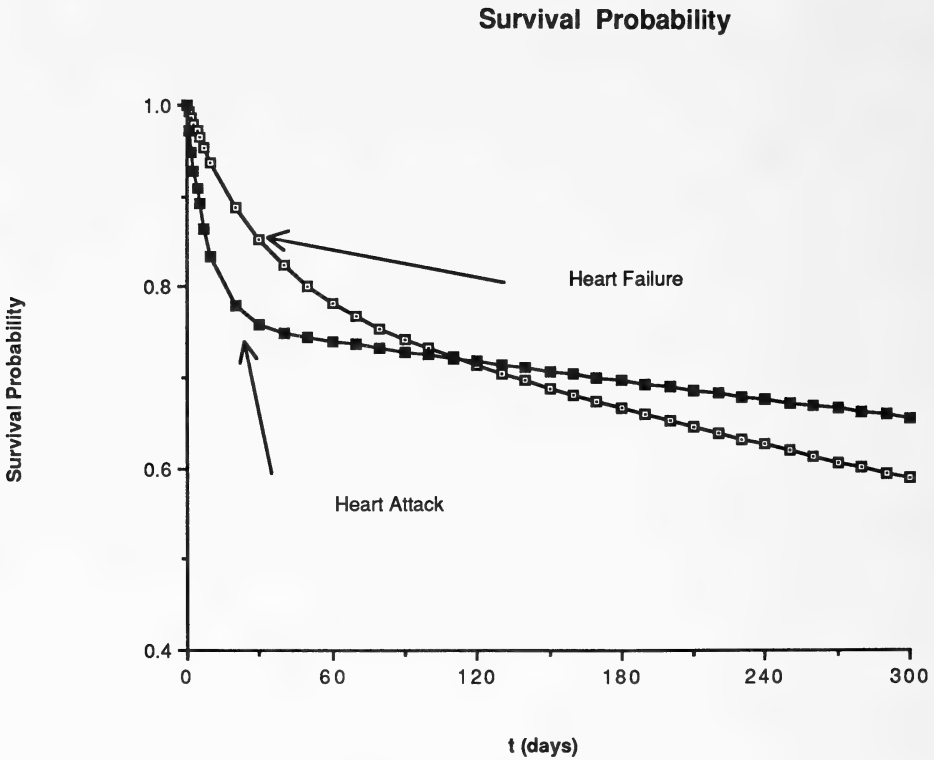

Fig. 3. The Survival Probability for the Modified Makeham Model for the Heart Attack and Heart Failure Data. Note the crossing of the survival functions at a time later than that shown for the risk functions shown in Figure 2.

when different sets of parameters are used. For example, when we compare two survival curves with parameters α, γ, δ and α', γ', δ' graphical solution of this problem will be apparent in many cases (see Figure 3). An algebraic solution of the nonlinear equation for the cross over time is not generally available. However, if the graphical display of several survival curves does not show a cross over, we know that if a cross over occurs, it will occur for a large value of time. In this case we can equate the expressions for survival curves and assume γt_c and $\gamma' t_c$ are large. Then approximate cross over time is

$$t_c = \frac{(\alpha'/\gamma') - (\alpha/\gamma)}{\delta - \delta'}.$$

Once the approximate cross over time is known, more precise estimates of the cross over time can be easily found from the graph or by iterative search. Note that a negative cross over time means the curves don't cross. For the survival curves shown in Figure 3, the cross over time from this approximation is 96 days. Also observe that the crossing of the risk functions shown in Figure 2 occurs before the crossing of the survival functions.

The Heart Attack Data for Separating the Short- and Long-Term Risk Factors: A Methodological Example

In this section, we explore the heart attack data described above. There are many possible variables to consider and

we wish to evaluate the risk factors that play initially and long-term.

In considering these variables, we used the time T_L where

$$T_L = -(1/\gamma)\ln(0.1)$$

to make separate evaluations of the short- and long-term risk factors. This formulation uses the time T_L for the initial excess hazard to diminish to 10% of the initial value of the excess risk, α, so

$$\alpha \exp(-\gamma T_L) = 0.1\alpha.$$

This formula for separating long and short term risks uses a criterion which differs from that for T_S used for sample size calculations. In our examples, the values T_S used in the sample size calculations are longer than the corresponding values of T_L. The use of a longer time for the sample design problem is more conservative. In the estimation problem, there is the additional consideration that a longer period leaves fewer cases for the evaluation of the long-term risk factors. Obviously either approach for separating long-term and short-term components of risk could be used. With either approach, the first step is to obtain parameter estimates for the simple form of the Makeham with no concomitant variables in the model. The values shown in Table 1 provide a time course which can be use to partially isolate the long-term from the initial or short-term risk factors.

Once the time T_L has been determined, two applications of the Cox proportional hazards model were made to evaluate the long-term and the short-term risk factors. The SAS[12] procedure, PROC PHGLM, provides a handy tool for evaluating a number of concomitant variables in the Cox proportional hazards model. One handy feature of the SAS implementation is the option for stepwise model building in which variables are tried systematically in succession to assess their contribution to the model. The initial risk factors were evaluated using the data with no modification. Significant risk factors are found using the stepwise feature of the SAS PROC PHGLM. The analysis is then repeated on the same data with the survival time shifted from zero to T_L. In the case of the heart attack data, $T_L = 22$ days. Consequently, the value 22 days is subtracted from each survival time in the initial data set. Negative values are assigned a missing value and deleted from the long-term survival analysis. The results of these analyses are shown in Tables 2 and 3.

Table 2 shows the parameter estimates for 24 risk factors which were significant ($P < 0.05$) predictors of patient survival in the Cox proportional hazards model. When the survival time was modified by subtracting $T_L = 22$ days from each survival time, 702 of the 3152 survival times became less than zero and had to be excluded from the long-term analysis shown in Table 3. Consequently, the number of cases used in the long-term model was reduced from the 3152 cases used in the initial risk evaluation to 2450 cases for the long-term risk evaluation. At the same time the number of censored survival times is reduced from 1311 to 609. Censored cases are those with a death or actual failure not yet observed during the period of observation. Note that of the 24 variables found to be significant predictors in the initial risk model, only 14 remained significant in the long-term model.

We consider this approach a handy kludge for dealing with a complex problem. However, with the results of the modified Makeham shown in Table 4, we estimated simultaneously the components of risk for all 24 variables shown in Table 2. In Table 4 an asterisk (*) was inserted by the parameter estimates for each of the delta components which were excluded from the long-term model shown in Table 3. There are the 10 variables in Table 2 which are excluded from Table 3. If the 10 delta components with asterisks in Table 4 are judged against their standard errors, these components are not statistically different from zero.

Furthermore the parameter estimates

Table 2.—Parameter Estimates for the Cox Proportional Hazards Model for Unmodified Survival Time for 3152 Cases of Heart Attack

The parameters estimates were obtained from 3152 cases of heart failure of whom 1311 died during the period observed. The MODEL CHI-SQUARE = 801.90 WITH 24 D.F. (−2 LOG L.R.) P = 0.0 The output is from SAS® PROC PHGLM.

PARAMETER ESTIMATES

STEPWISE PROPORTIONAL HAZARDS GENERAL LINEAR MODEL PROCEDURE

3152 OBSERVATIONS
1311 UNCENSORED OBSERVATIONS
0 OBSERVATIONS DELETED DUE TO MISSING VALUES

−2 LOG LIKELIHOOD FOR MODEL CONTAINING NO VARIABLES = 20381.20

MODEL CHI-SQUARE = 1141.74 WITH 24 D.F.
MAX ABSOLUTE DERIVATIVE = 0.1537D-09. −2 LOG L = 19579.30.
MODEL CHI-SQUARE = 801.90 WITH 24 D.F. (−2 LOG L.R.) P = 0.0

FINAL PARAMETER ESTIMATES

VARIABLE	CODE*	BETA	STD. ERROR	CHI-SQUARE	P
e^{AGE}	ERAGE	0.051	0.006	69.23	0.0000
Leukocytosis	HVWBC	0.012	0.003	14.54	0.0001
Low Value Potassium	LVK	0.172	0.072	5.64	0.0176
High Value PH	HVPH	1.727	0.635	7.39	0.0066**
Readmitted <30da	F122	0.291	0.115	6.40	0.0114**
Disoriented	F293000	0.403	0.160	6.36	0.0116**
Ischemia	F411800	−0.176	0.065	7.44	0.0064**
MI Age not determined	F412000	−0.157	0.065	5.80	0.0160**
Glucose	V425	0.000989	0.000229	18.53	0.0000**
Dissociation	F426890	1.004	0.171	34.32	0.000
Congestive Heart Failure	F428000	0.418	0.064	42.17	0.0000
BUN	V430	0.00526	0.00153	11.87	0.0006
PO_2	V461	−0.00723	0.00183	15.56	0.0001**
Coma/Stupor	F780005	0.906	0.128	50.27	0.0000
Pseudonomias	V785000	0.00185	0.00054	11.47	0.0007
Murmur	F785201	0.381	0.102	14.05	0.0002**
Systolic Blood Pressure	V785501	−0.00997	0.00148	45.16	0.0000
Respirations	V786010	0.0120	0.0034	12.18	0.0005
Stroke/TIA History	F826	0.367	0.088	17.36	0.0000
Coronary Heart Failure History	F832	0.261	0.075	11.99	0.0005
Myocardial Infarction History	F876	0.143	0.064	4.96	0.0259
Cancer	CA	0.632	0.147	18.44	0.0000
Chronic Renal Disease	RN	0.323	0.145	4.99	0.0254**
Diabetic	DB	−0.219	0.103	4.52	0.0336**

*Codes which begin with the letter F are indicator variables and codes that begin with V are values.
**Indicates variables not significant long-term as shown in Table 3.

Table 3.—Parameter Estimates for the Cox Proportional Hazards Model for a Modified Survival Time for 2450 Cases of Heart Attack

For this analysis the survival time was modified by subtracting 22 days. Negatives values were deleted from the analysis. The parameters estimates were obtained from the remaining cases of heart failure using SAS® PROC PHGLM. The estimate of a 22 day period for the long-term risk to dominate was based on the Makeham fit of the time course of the survival data. The estimates shown are for the variables shown in Table 1 which were identified as significant in the stepwise proportional hazards general linear model procedure with the default settings ($P < 0.05$).

609 UNCENSORED OBSERVATIONS
702 OBSERVATIONS DELETED DUE TO MISSING VALUES

FINAL PARAMETER ESTIMATES

Variables	CODES*	BETA	STD. ERROR	CHI-SQUARE	P
ᶜAGE	ERAGE	0.0691	0.0093	55.42	0.0000
Leukocytosis	HVWBC	0.0122	0.0058	4.46	0.0347
Potassium	LVK	0.277	0.112	6.09	0.0136
Dissociation A-V	F426890	0.896	0.338	7.02	0.0080
Congestive Heart Failure	F428000	0.555	0.090	37.64	0.000
BUN	V430	0.0155	0.0022	48.33	0.0000
Coma/Stupor	F780005	0.691	0.311	4.92	0.0266
Pseudonomias	V785000	0.00321	0.00081	15.80	0.0001
Murmur	F785201	0.378	0.148	6.50	0.0108
Respirations	V786010	0.0228	0.0053	18.63	0.0000
Stroke/TIA History	F826	0.552	0.123	20.24	0.0000
Coronary Heart Failure History	F832	0.374	0.109	11.74	0.0006
Myocardial Infarction History	F876	0.254	0.091	7.85	0.0051
Cancer	CA	0.939	0.199	22.31	0.0000

*Codes which begin with the letter F are indicator variables and codes that begin with V are values.

for the modified Makeham model provide the raw material to evaluate specific individual situations. With a specific case at hand, the component parameters can be used to estimate the structural parameters α, γ, and δ. With these estimates, both the predicted survival curve and the predicted risk function can be evaluated and plotted to provide a comprehensive forecast for the situation. When the model is used with a treatment or procedure that can be administered as appropriate, the results of the model prediction summarize the experience of thousands of observations to succinctly predict the course of survival for each option for the specific individual case. This means that a simple answer that one approach is preferred over another must be replaced by the more elaborate evaluation specific to the case. And the choices are more complex in that the survival curves may cross just as they did in the comparision of the heart attack and the heart failure data. In such cases, the choice is a trade-off between the short-term outcome and the long-term outcome. The result is an elaborate quantitative assessment of the expected outcome that is custom fit to cover a complex mix of individual situations. Clearly such analyses provide a resource for summarizing experience that awaits further exploration and exploitation.

The appeal of the proportional hazards model rests largely on the fact that the conclusions derived from comparisons based on this model remain simple, even though complex adjustments are made. The comparisons are simple because the proportional hazards model does not re-

Table 4.—Estimates of the Makeham Parameters for the Heart Attack Data Set

Maximum likelihood estimates are shown along with approximate standard error for the estimate (S.E.). Variables are identified with the code shown in Table 2 and a suffix designates alpha (α), gamma (γ) or delta (δ) components.

MAXIMUM OF THE LOG LIKELIHOOD = −8209.242

PARAMETER	α_0	γ_0	δ_0	ERAGE_α	ERAGE_γ	ERAGE_δ
ESTIMATE	−0.260	34.0	9.79	3.41E-02	−1.64E-02	6.26E-02
S.E.	7.77	9.23	13.1	1.54E-02	2.35E-02	1.17E-02
PARAMETER	HVWBC_α	HVWBC_γ	HVWBC_δ	LVK_α	LVK_γ	LVK_δ
ESTIMATE	1.03E-02	−4.89E-03	1.00E-02	−2.93E-03	−0.240	0.266
S.E.	4.60E-03	5.89E-03	7.08E-03	0.135	0.140	0.129
PARAMETER	HVPH_α	HVPH_γ	HVPH_δ	F122_α	F122_γ	F122_δ
ESTIMATE	0.712	−3.79	−2.00*	0.449	0.198	0.210*
S.E.	1.03	1.23	1.75	0.223	0.231	0.200
PARAMETER	F293000_α	F29300_γ	F29300_δ	F411800_α	F411800_γ	F411800_δ
ESTIMATE	0.874	0.201	0.127*	−0.292	−0.113	−0.117*
S.E.	0.267	0.274	0.343	0.135	0.140	0.110
PARAMETER	F412000_α	F412000_γ	F412000_δ	V425_α	V425_γ	V425_δ
ESTIMATE	−0.442	−0.305	−0.141*	1.69E-03	5.25E-04	1.53E-04*
S.E.	0.138	0.154	0.116	4.12E-04	4.48E-04	4.61E-04
PARAMETER	F426890_α	F426890_γ	F426890_δ	F428000_α	F428000_γ	F428000_δ
ESTIMATE	1.03	0.155	1.11	0.180	−0.264	0.461
S.E.	0.286	0.402	0.397	0.122	0.132	0.117
PARAMETER	V430_α	V430_γ	V430_δ	V461_α	V461_γ	V461_δ
ESTIMATE	−3.72E-04	−3.43E-03	1.62E-02	−6.26E-03	4.07E-03	−1.37E-03*
S.E.	2.96E-03	3.98E-03	3.09E-03	3.23E-03	3.68E-03	3.77E-03
PARAMETER	F780005_α	F780005_γ	F780005_δ	V785000_α	V785000_γ	V785000_δ
ESTIMATE	0.764	−0.495	−0.560	3.15E-04	−9.72E-04	3.44E-03
S.E.	0.175	0.221	0.722	1.03E-03	1.16E-03	9.96E-04
PARAMETER	F785201_α	F785201_γ	F785201_δ	V785501_α	V785501_γ	V785501_δ
ESTIMATE	3.21E-02	−0.388	0.387	−2.47E-02	−1.11E-02	5.35E-03*
S.E.	0.223	0.260	0.199	2.37E-03	2.59E-03	3.55E-03
PARAMETER	V786010_α	V786010_γ	V786010_δ	F826_α	F826_γ	F826_δ
ESTIMATE	−3.28E-03	−5.64E-03	3.11E-02	0.205	−5.54E-02	0.581
S.E.	6.21E-03	6.65E-03	6.53E-03	0.195	0.240	0.152
PARAMETER	F832_α	F832_γ	F832_δ	F876_α	F876_γ	F876_δ
ESTIMATE	0.147	0.157	0.484	−3.71E-02	−2.63E-02	0.338
S.E.	1.174	0.212	0.130	0.136	0.152	0.107
PARAMETER	CA_α	CA_γ	CA_δ	RN_α	RN_γ	RN_δ
ESTIMATE	−0.231	−0.837	0.803	−0.277	−1.22	−0.468*
S.E.	0.296	0.352	0.280	0.268	0.395	0.464
PARAMETER	DB_α	DB_γ	DB_δ			
ESTIMATE	−0.376	6.54E-02	7.89E-02*			
S.E.	0.218	0.216	0.161			

*Estimated delta components which correspond to values not found significant in the Cox proportional hazards model as shown in Table 3.

sult in survival curves that cross. The modeling approach illustrated by the modified Makeham places the focus on estimation of complex outcomes rather than the artificial reduction of complex outcomes to simplistic conclusions that came from formulating issues in terms of simplistic statistical hypotheses.

Conclusion

The examples of the use of the modified Makeham model to evaluate medical interventions are promising. The computational cost associated with fitting this complex model to large data sets may be a limitation. If computational resources become a problem, then an appropriate statistical answer is to use a sample. The computational limitations are more likely to come from the need to include many complex adjustments in our models. To deal with these problems in an economical fashion will be a challenge. In this paper, it has been shown that simpler models, such as the Cox proportional hazards model, can be used with the simple structural form of the Makeham model to gain useful insights. A related example by Olshen *et al.*[14] describes the Cox model and other powerful statistical methodology such as classification trees. We anticipate the need to build models in which it is necessary to examine many variables in many combinations. To do this efficiently, we must being various statistical and computing resources into play. For example, in medicine it is often the case that many variables may carry very similar information to the problem. When this is the case, there are numerous alternative models that are essentially indistinguishable. In such cases, external information can be very useful in the final selection of a model for a particular purpose. The external information may be in the form of costs, convenience, or reliability or risk to the patient of a clinical measure. The external information may include a realization that a variable has been miscoded or that the information is not coded consistently. The biggest challenge to using complex models for evaluating medical procedures and practices from the information contained in large data bases comes from the problems inherent in such large collections of data. On the other hand, the challenge is to have the courage to make the best of what is available. The very probing of the data to glean valuable insights regarding current medical practice will stimulate new questions and the very use of the data in this productive fashion will encourage those responsible for providing and maintaining these data to do a better job. Without use, the information is lost and the process of learning from experience is far more parochial than it need be. Many things are being tried. We need to study the really good practices and the really bad practices to learn what works and when it works. We need only look at the conventional gathering of medical knowledge to realize that clinical trials are conducted on very select populations, each practitioner has at most limited experience, information in journals is often based on studies at a single institution, and the current thinking instilled in medical school graduates changes slowly with experience. With models which focus on effective use of data to jointly evaluate short- and long-term risk factors, there is a challenge to fully investigate major data resources in order that we may better understand medical practice.

Acknowledgement

The author wishes to express his gratitude to Henry Krakauer and Miles Davis of the Health Care Financing Administration for their encouragement and comments on this paper.

References Cited

1. **Bennett, J. H.** 1974. *Collected Papers of R. A. Fisher, Volume IV 1937–1947,* The University

of Adelaide, Australia. Paper 159, pages 160–163.

2. **Shewhart, Walter.** 1931. *Economic Control of Quality of Manufactured Product,* New York, D. Van Nostrand Co., Inc.
3. **Cook, Edward.** 1914. *The Life of Florence Nightingale, in two volumes,* Macmillan and Co., London, Vol. **1,** Chapter II, "The Passionate Statistician (1859–1861)," pages 428–438.
4. **Kendall, Maurice and R. L. Plackett.** 1977. *Studies in the History of Statistics and Probability, Volume II.* Macmillan Publishing Company, New York. Chapter 19, Florence Nightingale as a statistician, paper by E. W. Kopf, Journal of the American Statistical Association, **15,** 388–404 (1916).
5. **Pearson, E. S. and M. G. Kendall.** 1970. *Studies in the History of Statistics and Probability,* Hafner, Darien, Conn. Chapter 7, Medical statistics from Graunt to Farr, papers by Major Greenwood reprinted from Biometrika, **45,** 101–27 (1943), **32,** 203–25 (1942); **33,** 1–24 (1943).
6. **Bailey, R. C. and Homer, L. D.** 1977. Computations for a best match strategy for kidney transplantation, *Transplantation,* **23:** 329–336.
7. **Bailey, R. C., Homer, L. D. and Summe, J. P.** 1977. A proposal for the analysis of kidney graft survival, *Transplantation,* **24:** 309–315.
8. **Jordan, Jr., Chester Wallace.** 1967. *Life Contingencies,* 2nd Edition, The Society of Actuaries.
9. **Kurtz, E. B.** 1930. *Life Expectancy of Physical Property Based on Mortality Laws.* The Ronald Press Company, New York.
10. **Gross, Alan J. and Clark, Virginia A.** 1975, Survival Distributions: Reliability Applications in the Biomedical Sciences, John Wiley & Sons, New York, page 61.
11. **Meinert, Curtis L.** 1986. *Clinical Trials, Design, Conduct, and Analysis,* Oxford University Press, New York.
12. *SUGI Supplemental Library User's Guide, Version 5 Edition.* 1986. SAS Institute, Inc. Cary, NC.
13. **Roper, William L., Winkenwerder, William, Hackbarth, Glenn M. and Krakauer, Henry.** Effectiveness in health care: An initiative to evaluate and improve medical practice. The New England Journal of Medicine, vol. **319,** No. 18. Nov. 1988.
14. **Olshen, Richard A., Gilpin, Elizabeth A., Henning, Hartmut, LeWinter, Martin L., Collins, Daniel and Ross, Jr., John.** Twelve-month prognosis following myocardial infaction: Classification trees, logistic regression, and stepwise linear discrimination. 1985. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Volume I,* Wadsworth, Monterey, California and the Institute of Mathematical Statistics, Hayward, California. pp. 245–267.

DELEGATES TO THE WASHINGTON ACADEMY OF SCIENCES, REPRESENTING THE LOCAL AFFILIATED SOCIETIES

Philosophical Society of Washington ..... Barbara F. Howell
Anthropological Society of Washington ..... Edward J. Lehman
Biological Society of Washington ..... Austin B. Williams
Chemical Society of Washington ..... Jo-Anne A. Jackson
Entomological Society of Washington ..... Manya B. Stoetzel
National Geographic Society ..... Gilbert Grosvenor
Geological Society of Washington ..... James V. O'Connor
Medical Society of the District of Columbia ..... Charles E. Townsend
Columbia Historical Society ..... Paul H. Oehser
Botanical Society of Washington ..... Conrad B. Link
Society of American Foresters, Washington Section ..... Mark Rey
Washington Society of Engineers ..... George Abraham
Institute of Electrical and Electronics Engineers, Washington Section ..... George Abraham
American Society of Mechanical Engineers, Washington Section ..... Michael Chi
Helminthological Society of Washington ..... Robert S. Isenstein
American Society for Microbiology, Washington Branch ..... Vacant
Society of American Military Engineers, Washington Post ..... Charles A. Burroughs
American Society of Civil Engineers, National Capital Section ..... Carl Gaum
Society for Experimental Biology and Medicine, DC Section ..... Cyrus R. Creveling
American Society for Metals, Washington Chapter ..... James R. Ward
American Association of Dental Research, Washington Section ..... Eloise Ullman
American Institute of Aeronautics and Astronautics, National Capital Section ..... Paul Keller
American Meteorological Society, DC Chapter ..... A. James Wagner
Insecticide Society of Washington ..... Albert B. DeMilo
Acoustical Society of America, Washington Chapter ..... Richard K. Cook
American Nuclear Society, Washington Section ..... Paul Theiss
Institute of Food Technologists, Washington Section ..... Melvin R. Johnston
American Ceramic Society, Baltimore-Washington Section ..... Joseph H. Simmons
Electrochemical Society ..... Alayne A. Adams
Washington History of Science Club ..... Albert Gluckman
American Association of Physics Teachers, Chesapeake Section ..... Peggy A. Dixon
Optical Society of America, National Capital Section ..... William R. Graver
American Society of Plant Physiologists, Washington Area Section ..... Walter Shropshire, Jr.
Washington Operations Research/Management Science Council ..... Doug Samuelson
Instrument Society of America, Washington Section ..... Carl Zeller
American Institute of Mining, Metallurgical and Petroleum Engineers, Washington Section ..... Ronald Munson
National Capital Astronomers ..... Robert H. McCracken
Mathematics Association of America, MD-DC-VA Section ..... Alfred B. Willcox
D.C. Institute of Chemists ..... Miloslav Rechcigl, Jr.
D.C. Psychological Association ..... Bert T. King
Washington Paint Technical Group ..... Robert F. Brady
American Phytopathological Society, Potomac Division ..... Roger H. Lawson
Society for General Systems Research, Metropolitan Washington Chapter ..... Ronald W. Manderscheid
Human Factors Society, Potomac Chapter ..... Stanley Deutsch
American Fisheries Society, Potomac Chapter ..... Robert J. Sousa
Association for Science, Technology and Innovation ..... Ralph I. Cole
Eastern Sociological Society ..... Ronald W. Manderscheid
Institute of Electrical and Electronics Engineers, Northern Virginia Section ..... Ralph I. Cole
Association for Computing Machinery, Washington Chapter ..... James J. Pottmyer
Washington Statistical Society ..... R. Clifton Bailey

Delegates continue in office until new selections are made by the representative societies.

Washington Academy of Sciences
1101 N. Highland St.
Arlington, Va. 22201
Return Requested with Form 3579

2nd Class Postage Paid
at Arlington, Va.
and additional mailing offices.