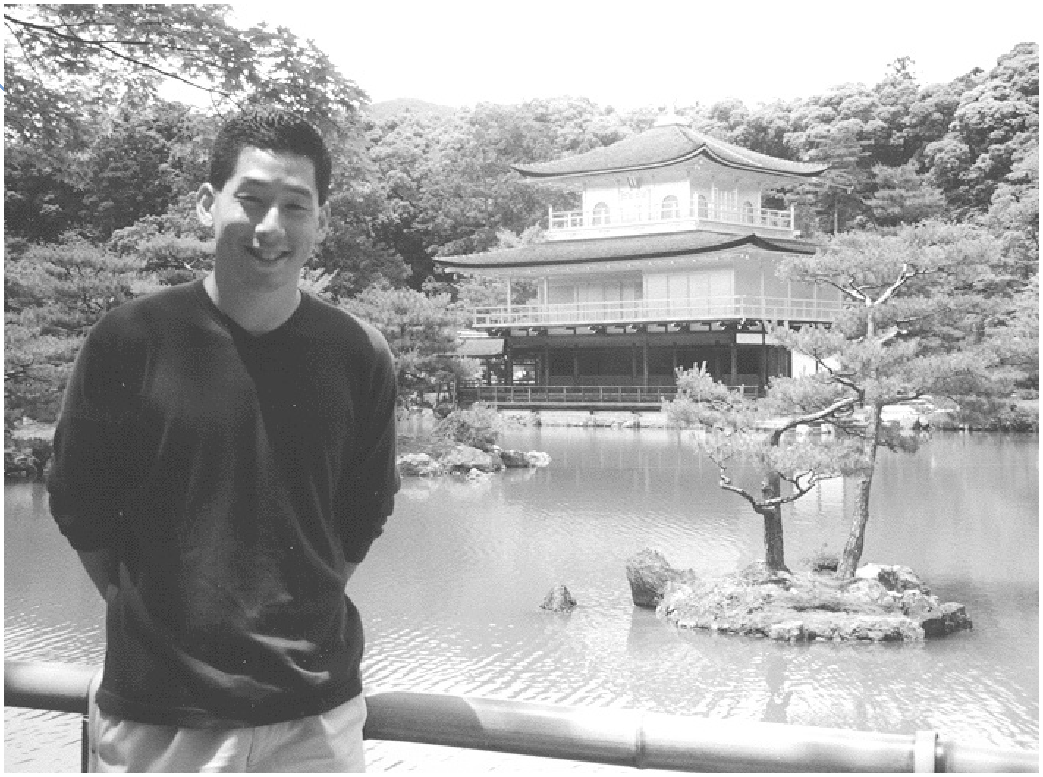
STATS

THE MAGAZINE FOR STUDENTS OF STATISTICS

SPRING 2001 • ISSUE # 31



INSIDE:

Reflections from the NSF/Monshubo Summer Program in Japan

AMERICAN STATISTICAL ASSOCIATION FOUNDED 1839

PRIZE COMPETITION

for the BEST STUDENT PAPER applying **STATISTICS** to **DEFENSE** ISSUES

The Committee on Statisticians in Defense and National Security of the American Statistical Association is pleased to announce the establishment of an annual prize for the best student paper applying statistics to defense issues.

The prize competition is open to any undergraduate or graduate student enrolled in an American institution of higher education. The paper must have been written in the preceding academic year (for this year's prize, July 1, 2000 to June 30, 2001). The paper must be nominated and submitted by a faculty member at the institution. Papers are limited to 10,000 words or 40 pages, including graphics. Student theses meeting these length requirements are acceptable.

Papers will be judged on the quality of the statistical work, the quality of the written presentation, and the significance of the contribution to understanding of defense issues.

For 2001, the prize consists of a plaque and $500.

For 2001, nominations and three copies of the paper should be submitted to Professor Dave Olwell, Department of Operations Research, Code OR/OL, Naval Postgraduate School, Monterey, California, by 1 July 2001. Questions should also be addressed to Professor Olwell (dholwell@nps.navy.mil). The prize announcement will be made at the annual JSM meetings.

AMERICAN·STATISTICAL·ASSOCIATION·FOUNDED·1839

STATS

The Magazine For Students Of Statistics

Spring 2001 • Number 31

STATS: The Magazine for Students of Statistics (ISSN 1053-8607) is published three times a year, in the winter, spring, and fall, by the American Statistical Association, 1429 Duke St., Alexandria, Virginia 22314-3415 USA; (703) 684-1221; fax: (703) 684-2036; Web site: *www.amstat.org*

STATS is published for beginning statisticians, including undergraduate and graduate students who have a special interest in statistics, and is distributed to student members of ASA as part of the annual dues. Subscription rates for others: $13.00 a year to members; $20.00 a year to nonmembers.

Ideas for feature articles and material for departments should be sent to the Editor, Jerome P. Keating. Addresses of the Editor and Editorial Board are listed above.

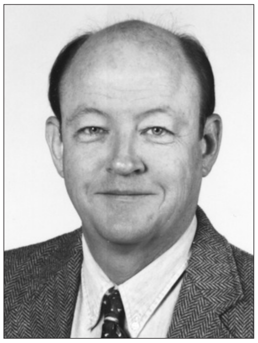
Requests for membership information, advertising rates and deadlines, subscriptions, and general correspondence should be addressed to *STATS* at the ASA office.

Produced by Oakland Street Publishing • Printed in U.S.A.

Features

Departments

Editor's Column



Jerome P. Keating

This is my last column, as Editor of *STATS*. The American Statistical Association establishes a committee to select a new Editor every three years. The new editor should be selected by April 15, 2001. He or she will then establish a new Editorial Board. A new and fresh perspective to *STATS: The Magazine for Students of Statistics* will begin again. I look forward to reading future issues of STATS and the continuing development of this magazine. We, the Editor and Associate Editors, have invested considerable time and effort in the magazine and its success is paramount. As Editor, I want to take this opportunity to thank the many people, who made *STATS* a success over the last three years.

The Associate Editors have worked tirelessly on behalf of the magazine. They have contributed greatly to the changes in STATS over the last three years. They contribute columns to each issue, review submitted manuscripts, and often solicit manuscripts from prospective authors. Bob Stephenson has been a member of the Editorial Board for twelve years and has produced the AP STATS column for the last three years. This column has been a great addition to the magazine in that the needs of new audiences of statistics students and teachers, especially at the pre-collegiate level, have been addressed. Allan Rossman has contributed many lively inquiries to the Outliers' column. I must confess that I look forward to receiving his contributions to each new issue and I read his column first. I still enjoy working the assignments. Jackie Dietz has edited the Student Voices' column and has mentored several students in the publishing process.

Beth Chance and Rudy Guerra have been major solicitors of new material for *STATS*. They have carefully reviewed other submissions, and provided sage commentary on the direction of *STATS*. Several of the personal profiles that appeared under the column, "A Day in the Life of a Statistician," were responses to inquiries from Beth Chance. The last member of the editorial board is Bob Mason, who is a candidate for the President-Elect of the American Statistical Association. Bob has been such a strong advocate for *STATS* and its mission of continuing outreach to each new generation of students. Bob has been involved with *STATS* since its inception more than a decade ago. We wish him well in his attempt for the position of President-Elect.

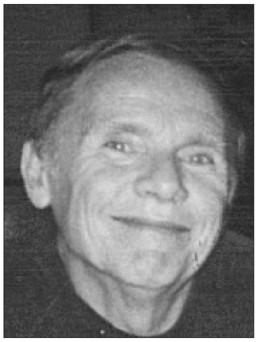
I also thank the ASA office and in particular, Mary Fleming and Megan Murphy for their support of *STATS*. You can also see their influence in the new cover and internal style for STATS. Mr. Peter Lindeman of Oakland Street Publishing has greatly reduced the process of turning manuscripts into magazines. He has been extremely responsive to the rapidly changing nature of the magazine. So there are many people who are responsible for the magazine.

In this issue, we have an eloquent tribute to the life of Lucien Le Cam by David Brillinger and Grace Yang. They trace the statisticien extraordinaire's many contributions that span almost half a century. Tom Moore and Vicki Bentley-Condit provide a primer on Permutation Tests with an application to infant handling by female baboons. These primers have been a feature that we introduced to meet requests by both our student and faculty readership. Jimmy Doi contributes another Student Voice column on the NSF/Monbusho summer programs in Japan, Korea, and Taiwan. Follow Jimmy's eight-week program and get a glimpse of our field of research in a foreign setting and his exposure to a completely new language and culture. Perhaps you will follow Jimmy and consider this wonderful opportunity for statistics graduate students. Thomas Gwise contributes a Student Project column and takes a brief look at the intersection of statistics and gamma spectroscopy by discussing the properties of gamma counting. Read this brief introduction to gamma spectroscopy to see why Tom says that this little corner of the physics world is almost all probability and statistics.

In Roxy Peck's AP STATS column, she provides us with another update on the AP STATS exam. Last May, a staggering number of 34,500 students took the AP STATS exam. The phenomenal growth in the popularity of the exam highlights the fact that statistics is rapidly being incorporated into the mainstream of the pre-college curriculum.

Jerome P. Keating

Lucien Le Cam, statisticien

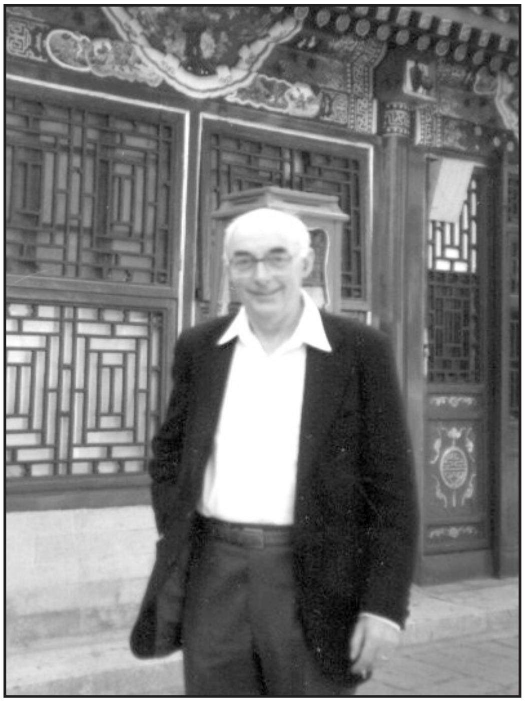


David Brillinger

Grace Yang

The mathematical statistician's mathematical statistician, the student's protector and one of the gentlemen of statistics, Lucien Marie Le Cam died April 25, 2000 in the San Francisco Bay Area, his home for fifty years. He was 75 years old.

Professor Le Cam was born into a Breton farm family on November 18, 1924. He graduated from a Catholic boarding school in the early years of World War II. He attended a seminary for exactly one day. (Details of that event may be found in Yang (1999).) After that he attended mathematics courses at a lycée in Clermont-Ferrand and stayed there for two years before registering as a student at the University of Paris in 1944. He obtained the Licence des Sciences from that university in 1945. Luckily, the War did not affect his education too much, for he "managed" to fail the physical examination thereby avoiding the draft by the Germans during their occupation of France.



Visiting the Lama Temple in Beijing, May 1985

For the following five years he worked at Electricité de France researching how to operate dams effectively and how to estimate risk probabilities of droughts and floods.

In 1950 he came to the University of California in Berkeley at the invitation of Jerzy Neyman as an Instructor for a year. Neyman urged him to stay on to work for a Ph.D. So, Lucien became a graduate student his second year at Berkeley after being an Instructor the first year! For those years he showed much amusement in telling the story of failing the Ph.D. Qualifying Exam in Mathematics on his first attempt. (That story is also related in Yang (1999).) He received a Ph.D. in 1952 writing a thesis titled "On Some Asymptotic Properties of Maximum Likelihood and Related Bayes' Estimates" under the supervision of Jerzy Neyman.

He was appointed Assistant Professor of Mathematics at Berkeley in 1953 eventually becoming Professor of Statistics in 1960. During the year 1972-3 he was Director of the Centre de Recherches Mathématiques in Montreal, Canada; otherwise he remained at Berkeley, becoming Professor Emeritus in 1991.

He was Chair of the Berkeley Department of Statistics from 1961–5, Acting Director of the Statistical Laboratory and Co-Chair of the Biostatistics Group at other times. He co-edited with Jerzy Neyman and Elizabeth Scott the celebrated Proceedings of the Berkeley Symposia. He ran the Neyman Seminar in the years after Jerzy Neyman's death in 1981 with panache and civility. Following Neyman's tradition, Lucien financed the weekly high tea for the seminar attendees and the drinks at the Faculty Club

David Brillinger is Professor of Statistics at the University of California, Berkeley where he was Lucien Le Cam's colleague for thirty years. Grace Yang is Professor of Statistics at the University of Maryland, College Park. She obtained her doctorate under Professor Le Cam's supervision in 1966 and has coauthored papers and a book with him.

Le Cam and his wife in the Summer Palace, Beijing, May 1985, with members of the Academica Sinica, left to right, W.-S. Xu, K.-T. Fang, S.-R. Wang.

afterwards. The Berkeley years are further chronicled in Lehmann (1997).

Lucien obtained important honors in statistics: Honorary Degree of Doctor of Science from the Free University of Brussels in 1997, Fellow of the New York Academy of Sciences, Fellow of the American Academy of Arts and Sciences, Hotelling Lecturer and President of the Institute of Mathematical Statistics. In 1989 there was a symposium at Berkeley honoring his 65th birthday and in 1994 another at Yale honoring his 70th. The 70th birthday celebration was preceded by a well-attended week-long workshop and symposium on Lucien's research and there was a Festschrift, that is a collection of papers written in his honor, Pollard et al (1997).

The Le Cam name lives on in Le Cam's Lemmas, Lecamian rainfall models, Le Cam's distance between experiments, Le Cam's metric dimension, Le Cam's one-step estimators (see below), Le Cam's students; and his spirit remains in the many concepts he introduced including contiguity, tightness, LAN (Local Asymptotic Normality), discretization of estimates, insufficiency, deficiency, and chaining.

Throughout the years his approach to research was strongly colored by the Bourbaki, a French series of mathematics books. These works were often viewed as excessively abstract, but Lucien defends that approach in Albers et al (1990) remarking:

> "The abstract side has the virtue that you can simplify things, and when you simplify things, they appear more reasonable, clearer and so forth. But if you want to apply [statistics] you have to get down to the nitty-gritty and do things that might not be so simple."

Turning to some specifics of his research work, scientific researchers often come to statisticians because they are concerned with the efficiency of some natural estimate they are using or of some experiment they are designing. Professor Le Cam studied the efficiency of estimates and experiments and constructed simpler ones. He was concerned with systematically reducing complex statistical problems to simpler ones, often involving the normal distribution.

His specialties included statistical decision theory, large sample theory and approximations. Le Cam's works are summarized in his Magnum Opus, *Asymptotic Methods in Statistics*, (1986a). To quote some of his own words, Le Cam (1984):

> "Asymptotic statistical theory is a body of limit or, better yet, approximation theorems used by statisticians to elude the intractability of all but the very simplest practical statistical problems and to obtain usable results."

To illustrate the spirit of this idea and his work consider his one-step estimator. It is very applicable and is often easy to compute. The general setup is the following:

Suppose $\{X_1, X_2, \ldots, X_n\}$ is a random sample from some probability density function, $f(x; \theta)$, and that one wants to estimate $\theta \in \Theta$. Assume that Θ is an open subset of a k-dimensional Euclidean space. Now in many practical problems, the likelihood function $\Pi_j f(x_j; \theta)$ is too complicated to use, for instance, for computing the maximum likelihood estimates (MLEs) for θ. Le Cam's estimation procedure is to look first for a $(\sqrt{n})$ consistent estimator (which is often easy to find), say, θ_n^*, and then make a correction to it to obtain the final estimate $\hat{\theta}_n$, of the form,

$$\hat{\theta}_n = \theta_n^* + n^{-1/2}(M_n^{-1}Y_n).$$

for some Y_n, M_n (details below). The elegant LAN theory says that if the so-called LAN conditions are satisfied, then $\hat{\theta}_n$ is asymptotically optimal. This is a remarkable procedure in that it calls for only a "one-step" correction, unlike the usual Newton-Raphson procedure. The same optimality would hold for the MLE as well, but it would be under more stringent conditions. The correction term $M_n^{-1}Y_n$ introduced above is computed based on "local maximization" as follows. By consistency of θ_n^*, for every $\varepsilon > 0$, there exist b and n_ε such that for $n \geq n_\varepsilon$

$$P[|\theta_n^* - \theta_0| \leq bn^{-1/2}] \geq 1 - \varepsilon$$

whatever the true θ_0. Thus for large n, the preliminary estimate will bring one (with a high probability) to local neighborhoods $\Theta_n = \{\theta: |\theta - \theta_0| \leq bn^{-1/2}\}$ of radius $bn^{-1/2}$ of the true θ_0. Estimation of θ now becomes that of estimating the local parameter $\tau = \theta - \theta_0$.

To estimate τ, one uses the log likelihood ratio

$$\Lambda(\theta,\theta_0)=\sum_{j=1}^{n}\log\frac{f(x_j;\theta)}{f(x_j;\theta_0)}$$

for θ in the neighborhood of radius $b/\sqrt{n}$ of θ_n^*. Replacing θ_0 and θ in the function Λ by θ_n^* and $\theta_n^* + \tau n^{-1/2}$ respectively yields

$$\Lambda(\theta_n^*+\tau n^{-1/2},\theta_n^*)\quad \tau\in\{t:|t|\leq b\}.$$

Next one locally approximates this function of τ by a linear-quadratic form as in

$$\Lambda(\theta_n^*+\tau n^{-1/2},\theta_n^*)\approx \tau' Y_n-\frac{1}{2}\tau' M_n\tau,$$

with Y_n the random vector and M_n the positive definite random matrix mentioned above. Maximizing this linear-quadratic approximation with respect to τ yields an estimate $\hat{\tau} = M_n^{-1}Y_n$, which is the correction term appearing in the final estimate $\hat{\theta}_n$.

Y_n and M_n can be constructed in various ways and they are not necessarily the first two derivatives of the function Λ. For instance, let

$$Z_{n,ji}=\frac{f(x_j,\theta_n^*+n^{-1/2}u_i)}{f(x_j,\theta_n^*)}-1 \quad \text{for } i=1,\ldots,k$$

where $\{u_i;\ i = 1, \ldots, k\}$ is the natural k-dimensional basis. Then Y_n could have its ith component

$$Y_{n,i}=\sum_{j=1}^{n}Z_{n,ji}\quad \text{for } i=1,\ldots,k$$

and M_n its components

$$M_{n,il}=\sum_{j=1}^{n}Z_{n,ji}Z_{n,jl},\quad \text{for } i,l=1,\ldots,k.$$

The inverse matrix M_n^{-1} serves as an estimate of the covariance matrix of $\hat{\theta}_n$. In the LAN theory $\hat{\theta}_n$ is an asymptotically most concentrated estimate by the Hájek Convolution Theorem.

Now, for an elementary example, suppose that the real-valued X has a double exponential density

$$f(x)=\frac{1}{2}e^{-|x-\theta|}$$

or a Cauchy density

$$f(x)=\frac{1}{\pi[1+(x-\theta)^2]}.$$

Both of these densities satisfy LAN conditions (see, e.g., pages 109–111, Le Cam and Yang (1990)).

The reader is invited to try to estimate θ using Le Cam's method. Here θ is a one-dimensional parameter. One can take $u_1 = 1$ in the computation of $Z_{n,j1}$. Since Le Cam's procedure is asymptotically optimal, it is important to check, for finite samples, if the resulting estimate $\hat{\theta}_n$ works. This can be done by checking if $f(x, \hat{\theta}_n)$ fits the original data $x_1, \ldots, x_n$. It should be pointed out that there is a certain amount of flexibility in using Le Cam's procedure. If $\hat{\theta}_n$ does not fit the data well for the choice of u_1 (which is 1 in our case), one can try other values of u_1, say 1.5, .8, or others for improving the fit. Perturbing the value of u_1 would not affect the asymptotic optimality of the resulting estimate.

It is perhaps worth recording that one of us (DRB) teased Lucien a number of times asking whether he would have the will power to carry out but one iteration. His reply was always, well, the theory indicates that only one step is necessary.

Professor Le Cam's theoretical work was sometimes stimulated by practitioner's questions, sometimes by teaching courses and often by his own applied work. The practical bent is well evidenced in the following list of his principles, Le Cam (1990):

Basic Principle 0. Do not trust any principle.

Principle 1. Have clear in your mind what it is you want to estimate.

Principle 2. Try to ascertain in some way what precision you need (or can get) and what you are going to do with an estimate when you get it.

Principle 3. Before venturing an estimate, check that the rationale which led you to it is compatible with the data you have.

Principle 4. If satisfied that everything is in order, try first a crude but reliable procedure to locate the general area in which your parameters lie.

Principle 5. Having localized yourself by (4), refine the estimate using some of your theoretical assumptions, being careful all the while not to undo what you did in (4).

Principle 6. Never trust an estimate which is thrown out of whack if you suppress a single observation.

Principle 7. If you need to use asymptotic arguments, do not forget to let your number of observations tend to infinity.

Principle 8. J. Bertrand said it this way: "Give me four parameters and I shall describe an elephant; with five, it will wave its trunk."

These principles are certainly meant to be applied to any of his LAN estimates.

Lucien had special concerns for the education and welfare of students, both undergraduate and graduate. None were ever turned away from his office and he was invariably in the Department Coffee Room chatting from noon til 1:00. On one occasion he returned to the Coffee Room upset after seeing a Dean. Just what had happened wasn't clear but a student was involved. Lucien erupted with: "Rules are meant to help, not to hinder." Indeed.

Professor Le Cam was the driving force behind the setting up of the Line and Michel Loève Fellowship awarded to the graduate student at Berkeley with the greatest promise in probability.

He tried hard to involve people in the content of his papers. For example Le Cam (1986b) starts with the words:

> "In the beginning there was de Moivre, Laplace, and many Bernoullis, and they begat limit theorems, and the wise men saw that it was good and they called it by the name of Gauss."

Lucien had a photographic memory and was a student of languages: Ancient Greek, Latin, and Chinese. Re the last, here's an anecdote. He had a memorable visit to mainland China in 1985 and returned with various amusing stories. One concerned the words for rat and teacher. In Chinese, the word teacher is pronounced as "Lao Shi" and rat "Lao Shu". One day he saw a dead rat outside the window of his hotel room. Since the maintenance people could hardly speak or understand English, to let the hotel know about the rat, he decided to attract the attention of the waitresses at the hotel restaurant. He introduced himself to them in Chinese "I am a Lao Shu." The waitresses laughed their heads off and then he told them that "There is a dead Lao Shi in my room". The waitresses followed him happily to his room to see the dead teacher!

Lucien had some pet peeves: persons who proved theorems without assumptions, Deans as mentioned above, subjective Bayesians, military officers and squirrels who ate the tops of his redwood trees. There was a period when he liked to tack onto his papers, even ones in French, the footnote: "The paper is submitted in partial fulfillment of the promotion requirement of the University of California, Berkeley." See e.g. Le Cam (1970).

He had a clear presence on the Berkeley Campus. His office door was always open. Most everyone working anywhere near the large building, Evans Hall, where the Statistics Department is housed in seemed to know and be very fond of him.

In the last years of his life one continually saw him sitting at a computer terminal word processing revisions of his books, e.g. Le Cam and Yang (1990). The Department computer staff were much impressed with how much he had learned about LaTeX. A few hours before being taken to the hospital, he was busy making LaTeX corrections of the second edition. The volume was published posthumously in August.

At his death Lucien Le Cam remained a proud citizen of France. He had continued to write papers in French throughout his career and to keep in close contact with the French mathematicians.

With his death there will be a consequent reassessment of his work. Many students will learn about it and him for the first time and will be intrigued. It is clear that he has affected both graduate and undergraduate education by the route of today's theory becoming tomorrow's practice, and this will continue.

We are all the richer for having obtained the fruits of Professor Le Cam's research accomplishments and many for receiving his personal advice and help, but we are the poorer for having lost his physical presence.

Professor Le Cam's publication list may be found at *www.stat.berkeley.edu/~lecam* as may memorials.

References

Albers, D. J., Alexanderson, G. L. and Reid, C. (1990). Editors of *More Mathematical People: Contemporary Conversations*. Harcourt Brace Jovanovich, Boston.

Beran, R. J. and Yang, G. L. (2000). Obituary of Lucien Le Cam. *Bulletin of the Institute of Mathematical Statistics* 29, 464–6.

Bourbaki, N. (various years). Éléments de Mathématiques. Hermann, Paris.

Le Cam, L. M. (1970). Remarques sur le théorème limite central dans les espaces localement convexes. Pp. 233–249 in *Colloques internationaux de Centre National de la Recherche Scientifique* No. 186.

Le Cam, L. M. (1984). Review of books by Ibragimov and Has'minski and by Pfanzagl. *Bulletin of the American Mathematical Society* 11, 392–400.

Le Cam, L. M. (1986a). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York.

Le Cam, L. M. (1986b). The Central Limit Theorem around 1935. *Statistical Science* 1, 78–96.

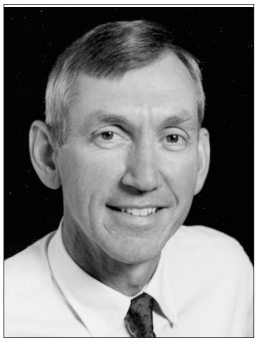
Le Cam, L. M. (1990). Maximum likelihood: an introduction. *International Statistical Review* 58, 153–171.

Le Cam, L. M. (1992). Letter from Lucien Le Cam. Pp. xv-xviii in Pollard et al (1997).

Le Cam, L. M. and Yang, G. L. (1990). *Asymptotics in Statistics: Some Basic Concepts*. Springer, New York.

Lehman, E. L. (1997). Le Cam at Berkeley. pp. 297–

Using Permutation Tests to Study Infant Handling by Female Baboons



Thomas L. Moore

Vicki Bentley-Condit

Introduction

Anthropologist Vicki Bentley-Condit of Grinnell College studied interactions between female and infant yellow baboons (*Papio cynocephalus cynocephalus*) at the Tana River National Primate Reserve, Kenya. She collected the data for her study by observing baboons in twenty-minute focal samples over an 11-month period in 1991-92 for a troop including 23 female baboons, 11 of which were mothers with infants (no mother with more than one offspring). Bentley-Condit observed and recorded interactions between females and infants, excluding interactions between a mother and her own offspring. One objective of her study was to see if female rank (described in the next paragraph) impacted the pattern or success rate of these infant-handling interactions.

Separately from the infant-handler interactions, Bentley-Condit computed a "dominance hierarchy score" for each female using a calculation based on aggressive and submissive interactions between pairs of the females in the troop. These scores use a standard method and exhibit natural "break points" that made it possible to translate the scores into High, Mid, and Low ranks, which we will code respectively as 1, 2, and 3 throughout the paper.

Regarding the objective of investigating the relationship of female rank and infant handling, Professor Bentley-Condit established the following research hypothesis:

Research Hypothesis: Females will tend to handle the infants of females who are ranked the same as or lower than themselves.

Interactions between females and infants were categorized as:

(1) *Passive*: movement to within 1m of the mother-infant pair with no attempt to handle,

(2) *Unsuccessful*: movement to within 1m of the mother-infant pair with an attempted (but not successful) handle, or

(3) *Successful*: a successful handle.

Each female has either one or zero offspring, and interactions between a mother and her own infant were not of interest and so were excluded from the data set.

Even though differentiating among these categories of interactions is ultimately of interest to the study, we will concentrate initially on Table 1, which gives for each female-infant pair the number of interactions *of any category*. For example, the value of 13 in cell (2,1) represents 13 interactions between Handler KM and Infant HZ over the

Thomas L. Moore has taught statistics and mathematics at Grinnell College since 1980. He has been active in statistics education within both ASA and the Mathematical Assocation of America and has spent academic leaves at the Eastman Kodak Company and Mt. Holyoke College.

Vicki K. Bentley-Condit has taught in the anthropology department at Grinnell College since 1995. She has been studying both feral and captive baboon behavior since 1988 and is an active member of the American Society of Primatologists.

Table 1: Data matrix giving the total number of interactions by handler (columns) and infants (rows) over the course of the observational period.

HANDLERS ranks		KM	KN	NQ	PO	HQ	LL	NY	PS	SK	ST	WK	AL	CO	DD	LS	LY	MH	ML	MM	PA	PH	PT	RS
		1	**1**	**1**	**1**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
INFANTS/ Mothers **ranks**																								
KG/KM	**1**	0	0	4	2	1	0	0	0	3	1	0	0	0	0	0	0	0	0	0	0	0	2	1
HZ/HQ	**2**	13	23	7	5	0	2	1	1	5	6	18	1	6	3	0	1	4	1	0	9	0	10	1
LC/LL	**2**	4	0	1	4	3	0	2	1	1	5	3	1	0	0	1	0	2	1	1	1	0	1	6
NK/NY	**2**	12	4	10	5	9	1	0	2	3	11	7	8	6	3	1	0	2	1	1	5	3	3	3
PZ/PS	**2**	1	3	4	1	0	0	0	0	0	0	2	0	2	0	0	0	3	0	1	1	0	3	0
CY/CO	**3**	2	2	7	3	1	1	2	0	3	12	16	3	0	2	0	0	2	0	0	1	0	0	2
LZ/LS	**3**	1	0	3	2	1	1	0	0	2	0	5	2	2	2	0	1	9	2	0	0	0	3	2
MQ/ML	**3**	0	1	5	2	2	4	2	2	2	4	5	7	5	2	1	1	7	0	4	4	1	0	2
MW/MH	**3**	3	0	7	4	2	3	0	5	2	8	13	7	14	2	0	0	0	4	0	8	0	13	6
MX/MM	**3**	2	3	4	5	0	0	0	0	0	5	2	9	3	1	0	0	2	0	0	1	2	2	3
PK/PH	**3**	2	0	6	4	3	4	1	0	0	15	10	8	5	1	0	3	1	1	6	3	0	7	5

Boldface numbers give ranks of handlers or infants, with 1 being High ranking, 2 being Mid ranking, and 3 being Low ranking. Each handler and infant have a two-letter ID, infant ID's are separated by their mother's ID with a /. Horizontal and vertical lines separate the rank categories.

observational period of the study. For now we will ignore the fact that this breaks down into 2 passive interactions, 4 unsuccessful interactions, and 7 successful interactions. Near the end of the paper, we will return to the distinction between types of interactions.

A Descriptive Look at the Data

A first step in investigating the research hypothesis is to look at a contingency table of infant rank by female handler rank (Figure 1). By 'infant rank' we refer to the rank of the infant's mother.

Notice that High ranked handlers are more likely to handle High and Mid ranked infants than are Mid or Low ranked handlers and that Mid ranked handlers are more likely to handle Mid ranked infants than are Low ranked handlers. Both observations support the research hypothesis.

We can also see the same pattern in a pseudo-Z-score breakdown (Figure 2). Figure 2 gives $(O - E)/\sqrt{E}$ for each cell of the table in Figure 1 (A), where O = the observed count and E = the expected cell frequency using the conventional expected value calculation for a chi-square test of independence.

This table corroborates what we learned from Figure 1. The progression 1.32, 0.16, and -1.17 in Figure 2B shows that high-ranked handlers are more likely to handle high-ranked infants than mid-ranked handlers, which in turn are more likely to handle high-ranked infants than low-ranked handlers. Similarly, the pattern goes in the same direction for the handling of mid-ranked infants. Finally, low-ranked infants have a higher than expected frequency of handles by low-ranked handlers and a much lower than expected frequency of handles by high-ranked handlers. This summary of the data also supports the research hypothesis.

Statistical Significance

With the observations of the previous section in hand, Bentley-Condit approached Moore about the question of statistical significance: the descriptive analysis uncovers patterns that support the research hypothesis, but could chance variation explain these patterns?

Figure 1: Interactions by Infant Rank and Handler Rank (A); column percentages (B)

Handler's rank		Hi	Mid	Low	Hi	Mid	Low
Infant	Hi	6	5	3	3.5%	2.2%	1.1%
Rank	Mi	97	83	96	56.7%	36.7%	33.9%
	Lo	68	138	184	39.8%	61.1%	65.0%
			(A)			(B)	

Figure 2

	Hi	Mid	Low
Hi	3.52	4.65	5.83
Mi	69.41	91.73	114.86
Lo	98.07	129.62	162.31

(A: Expecteds)

	Hi	Mid	Low
Hi	1.32	0.16	-1.17
Mi	3.31	-0.91	-1.76
Lo	-3.04	0.74	1.70

(B: Pseudo-Z-scores)

On the left (A) are expected values from Figure 1, under the assumption that handler rank and infant rank are independent. On the right (B), each entry gives (O – E)/√E for the cell, where O is the observed frequency and E is the expected. These "pseudo-Z-scores" can be interpreted as we usually interpret standard normal scores.

The notion of statistical significance carries here the "caveat emptor" of any inferential procedure applied to observational data: the randomness required by the test is assumed and not designed; that is, randomness will be embedded in the null hypothesis of the significance tests we will apply to these data.

We could consider testing the research hypothesis using a chi-square test of independence for the 3-by-3 contingency table. This yields a test statistic of

$$C=\sum\frac{\left(O-E\right)^2}{E}=30.76$$

and, with 4 degrees of freedom, a *P*-value of 0.000 (3.43 x 10^{-6}, to be precise). This suggests we reject the null hypothesis that handler rank is independent of infant rank and strengthens the descriptive conclusion reached above.

We should, however, be cautious in applying this chi-square test to these data. The primary assumption in the chi-square test of independence is that these 680 observations are independent of one another. Clearly this is impossible given the repetition of interactions between particular females and infants; for example KN interacts with HZ a total of 23 times and these 23 interactions necessarily end up in the same cell of the 3-by-3 table.

Permutation Tests

To take account of this complexity in the data we use a permutation test. Permutation tests (also known as randomization tests) were first proposed by R.A. Fisher in the 1930's for the analysis of randomized experiments. Efron and Tibshirani (1998) describe permutation tests as "... a computer-intensive statistical technique that predates computers." They are conceptually simple, but before the days of high-speed computing were of limited practical value. Fortunately, Fisher saw that standard normal-based methods approximated the permutation tests in most classical situations, so permutation tests were, for Fisher, a theoretical and conceptual tool, more than a practical tool (1998, page 202). One permutation test that did find its way into standard practice was Fisher's famous exact test for two-by-two tables, whose computation is a simple application of the hypergeometric distribution. This test provides an alternative to the chi-square test, useful for situations where the chi-square approximation is questionable because of small samples.

Examples

We will now illustrate the notion of permutation tests with examples, starting simple, and working our way back to the baboon data. Permutation tests have wide applicability beyond categorical data problems leading to two-way tables; consult books by Efron and Tibshirani (1998), Good (2000), and Manly (1991) to learn more about the range of applications.

Example 1—A lady tasting tea. We begin with a simple example of Fisher's exact test. In his famous book *The Design of Experiments*, Fisher (1937, pages 13–29) says that "A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or tea infusion was first added to the cup." Fisher constructs an experiment to test her claim: We will prepare 8 cups of tea, at random assigning 4 of them to be "milk first" and (by implication) the other 4 to be "tea first." The woman proceeds to taste from each of the 8 cups and to label each as "milk first" or "tea first." If she correctly identifies a sufficiently high number of the 8, she will have established her claim.

Here the null hypothesis is H_0: "The lady's claim is false" versus an alternative of H_a: "The lady's claim is true." What would make us believe the lady's claim? Suppose the experiment resulted in the data set given in Figure 3A and the resulting 2-by-2 table of Figure 3B.

If the null hypothesis were true, the lady would have no ability to distinguish cups based upon her ability to tell if milk or tea is added to the cup first. Yet, she has correctly identified 75% of

Figure 3: A hypothetical outcome of the lady tasting tea.

Cup ID	"Guess"	First Poured
1	M	M
2	M	M
3	M	M
4	M	T
5	T	M
6	T	T
7	T	T
8	T	T

Raw Data (A)

		Lady's "Guess"		
		Milk	Tea	(total)
	Milk	3	1	(4)
Poured First	Tea	1	3	(4)
	(total)	(4)	(4)	

2-by-2 table (B)

In figure A, an M in column 2 represents an instance where she said the milk was put into the cup first; a T represents where she said tea was put in first. Similarly, an M or T in column 3 represents the truth about which was added first to the cup. The lady has correctly identified 3 of the 4 cups with the milk first and 3 of the 4 cups with the tea first.

the 8 cups. Do the data provide evidence against H_0 in favor of H_a? The logic of Fisher's exact test is this: Devoid of her claimed ability, the lady makes her decisions about the 8 cups through other means: guessing, cup order, temperature differences, etc. Thus, if H_0 is true, then because of the random assignment of treatments (Milk First or Tea First) to cups, any permutation of the labels in column 3 of Figure 3A is equally likely, so that any 2-by-2 table with row and column totals as given in Figure 3B is equally likely. If random permutations of column 3 have small probability of resulting in a table as skewed toward H_a, or more skewed, than is our data, then we would have evidence against H_0 and toward H_a. This probability is, of course, the P-value of the test. Because both margins of the 2-by-2 table are constrained by design, we can base the P-value calculation on X, the count in the upper left cell of the 2-by-2 table; X = 3 for our data set. The P-value is then $P(X \geq 3)$ assuming H_0 to be true.

Under H_0, running the experiment is equivalent to randomly permuting 4 M's and 4 T's to column 3 of Figure 3A. Each such assignment leads to a 2-by-2 table. There are

$$\binom{8}{4} = 70$$

such random permutations. Of these we must count how many lead to tables as extreme (in the direction of H_a) or more extreme than Figure 3B. There are

$$\binom{4}{3}\binom{4}{1} = 16$$

that lead to the same 2-by-2 table as Figure 3B and there is just one more permutation, namely when the 4 M's are placed before the 4 T's in column 3, leading to 8 correct guesses and the 2-by-2 table.

$$\begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$$

The P-value is thus seen to be

$$\frac{16+1}{70} = .243$$

a value that does not support rejection of the null hypothesis. In this example, in order for the lady to substantiate her claim she would have to be perfect. That is, if the experimental data had been

$$\begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$$

then our P-value would have been

$$\frac{1}{70} = .0143$$

a result that gives reason to reject H_0.

Example 2: An illustrative example with fake baboons.

Our second example is again simple, but is more similar to the structure of Table 1 than was the previous example. Besides illustrating the mechanics of a permutation test, this example will suggest other important properties of permutation tests:

- They provide a method for establishing significance in contingency table data where standard inferential procedures, such as chi-square, cannot be used because assumptions are not met.
- They provide more flexibility in the choice of test statistic than a chi-square analysis and

Figure 4: A hypothetical, smaller example of baboon data similar in structure to the real data.

					Handler, rank			
			A	B	C	D	E	F
			[1	1	2	2	3	3]
Infant	a	[1]	0	3	2	1	0	1
Rank	c	[2]	5	5	0	5	0	0
	e	[3]	5	4	6	6	0	9

			Handler's rank	
		[1	2	3]
Infant's	[1]	3	3	1
Rank	[2]	10	5	0
	[3]	9	12	9

Here, there are 3 infants (a, c, and e), one of each rank. And there are 6 females, two of each rank. A is a's mother, C is c's mother, and E is e's mother. To the right is the 3-by-3 contingency table, with infants along rows and females along columns.

may provide a more powerful test for a particular research hypothesis; and

- Except for certain, small data sets, the calculation of the null distribution for one's test statistic will be hard to calculate analytically, but will be fairly easily approximated with modern statistical computing software.

Consider the data matrix in Figure 4. This data set is similar in structure to Table 1, but is smaller and is fake. There are 3 infants (a, c, and e) and 6 females (A, B, C, D, E, and F). Three of the females are mothers of precisely one infant and the notation is chosen to be helpful: A is a's mother, C is c's mother, and E is e's mother. To the right of the data is the 3-by-3 contingency table of infant rank (rows) by handler rank (columns). For example, the 10 count in cell (2,1) results from the 5 interactions between female A and infant c plus the 5 interactions between female B and infant c.

A chi-square test of independence yields a test statistic of C = 8.117 and, with 4 degrees of freedom, a P-value of .087, a result that is marginally significant. The validity of this analysis requires a sampling assumption and while the analysis is valid under several sampling models (i.e., multinomial, product-multinomial, or Poisson; see Fienberg [1983, pp. 15–16]), all such models require that, in this example, the 52 observations be independent of one another. Observations are clearly not independent here; for example, the 9 handlings of infant e by handler F would necessarily have to be in the same cell of the contingency table, as would be the case for all other frequencies in the original data matrix. That is, frequencies from the data set necessarily enter the contingency table in *clusters*, a complication not admitted by the chi-square analysis, but amenable to a permutation test.

To illustrate a permutation analysis of these data we first consider the null hypothesis and an alternative statement of the null.

H_0: Handler rank and infant rank are independent.

H_0 (alternative statement): The female handlers interacted with infants as given in the data set. These interactions involved a variety of complex causes, but none of this complexity had anything to do with ranks. That is, ranks can be viewed as meaningless labels attached to infants and females.

The null hypothesis posits a randomness to infant and handler ranks, which we employ in constructing the null sampling distribution of a test statistic. Before computing this distribution, we choose a test statistic that reflects the level of agreement between the data and the null hypothesis. One such test statistic is the usual chi-square test statistic, C. We will discuss other test statistics below. The sampling distribution of the test statistic under the null is defined by the following process:

(1) Assign ranks at random to infants and females using the rank distributions of the data set. That is, assign ranks at random so that infants are assigned, in this case, 1 High, 1 Mid, and 1 Low and so that females are assigned 2 High's, 2 Mid's, and 2 Low's. This assignment leads to the original data table but with permuted ranks.

(2) Re-form the 3-by-3 table.

(3) Compute the value of test statistic for this table.

Figure 5 gives two possible permutations that could be assigned at random. Each permutation of the ranks to females and infants induces a 3-by-3 contingency table. The top data table and 3-by-3 table corresponds to a permutation of 3, 1, 2 on the infants and 3, 1, 1, 3, 2, 2 on the females. This permutation is shown in the usual way in the figure and we take note of the agreement between infant's and mother's ranks.[1]

The first permutation leads to a chi-square test statistic of C = 8.410, greater than the value observed in the data set, while the second

Figure 5: Two possible random permutations of ranks.

(1)		A	B	C	D	E	F		[1	2	3]
		[3	1	1	3	2	2]	[1]	8	0	7
a	[3]	0	3	2	1	0	1	[2]	10	9	11
c	[1]	5	5	3	2	0	0	[3]	5	1	1
e	[2]	5	4	6	6	6	3		C = 8.410(S = 16)		

(2)		A	B	C	D	E	F		[1	2	3]
		[3	2	1	1	2	3]	[1]	5	5	5
a	[3]	0	3	2	1	0	1	[2]	12	10	8
c	[1]	5	5	3	2	0	0	[3]	3	3	1
e	[2]	5	4	6	6	6	3		C = .960 (S = 16)		

Values of C, the chi-square test statistic are given. In parentheses we give the values of a different test statistic, S, which we discuss below.

permutation leads to a value of C = .960, smaller than that observed. The sampling distribution of C, under the null is determined by computing C for each of the equi-probable (under H_0) permutations of ranks. In this case, the data set is small enough that one can completely enumerate the permutations. (There are 36 total: 6 ways to assign 1, 2, and 3 to infants and, for each of these, 6 ways to assign ranks 1, 2, and 3 to the non-maternal handlers. The ranks of the maternal handlers are fixed and not randomized.) Table 2 lists these permutations along with their C values.

Of these 36 values, 12 are greater than or equal to the observed value of 8.117, so the *P*-value is 12/36 = .3333. This result is much less significant than the earlier chi-square analysis (*P*-value=.087) would have led us to believe.

While the smallness of this example allows a complete enumeration of all possible permutations and, thus, an exact calculation of the null distribution and *P*-value, such an exact calculation is usually intractable. In these situations one must, instead, approximate the null distribution and *P*-value using a computer simulation. For this example, one could generate at random, say, 10,000 permutations from the possible 36, each time computing the 3-by-3 table and value of C. In this way one would get an empirical distribution of C under H_0 and an empirical estimate of the *P*-value by observing what proportion of the 10,000 values of C are greater than or equal to 8.117. With high-speed computing, generating 10,000 iterates would be fast and one could estimate the *P*-value to within .01 with 95% confidence, a result accurate enough for practical purposes.

Other Test Statistics

Thus far, we have seen two advantages to the permutation test over a chi-square test. First, the permutation test makes few assumptions on the data. The disparity between the two *P*-values above suggests that the independence assumption matters for these data. The second advantage provided by

Table 2: The 36 permutations of ranks along with the corresponding test statistic values.

123112233	123112332	123122133	123122331	123132132	123132231	132113223	132113322	132123123
8.117	7.135	8.410	5.962	2.426	0.960	7.135	8.117	2.426
132123321	132133122	132133221	213211233	213211332	213221133	213221331	213231132	213231231
0.960	8.410	5.962	8.410	5.962	8.117	7.135	0.960	2.426
231213213	231213312	231223113	231223311	231233112	231233211	312311223	312311322	312321123
2.426	0.960	7.135	8.117	5.962	8.410	5.962	8.410**	0.960 **
312321321	312331122	312331221	321312213	321312312	321322113	321322311	321332112	321332211
2.426	8.117	7.135	0.960	2.426	5.962	8.410	7.135	8.117

The first 3 elements of each permutation gives ranks assigned to a, c, and e, respectively; the final 6 elements give the ranks assigned to A, B, C, D, E, and F. The two values of C marked with ** are the two examples shown in Figure 6.

the permutation test is that it handles data sets with complicated dependencies and repetitions, as our simple Example 2 illustrates.

The third advantage to the permutation test is that one can use test statistics other than the chi-square test statistic C used above. Different test statistics may naturally arise from the researcher's hypothesis.

For example, in data arising from the baboon study, the research hypothesis states that handlers tend to avoid handling infants of higher rank than themselves. This suggests the following test statistic, S, calculated from the 3-by-3 table:

$$\begin{matrix} a & b & c \\ d & e & f \\ g & h & i \end{matrix}$$

as: a – b – c + d + e – f + g + h + i. Notice that cells for which the female's rank is at least as high as the infant's rank agree with the research hypothesis and so contribute positively, while other cells contribute negatively. For example, the value of the test statistic S for the table:

$$\begin{matrix} 3 & 3 & 1 \\ 10 & 5 & 0 \\ 9 & 12 & 9 \end{matrix} \quad = \quad \begin{matrix} 3 & -3 & -1 \\ +10 & +5 & -0 \\ +9 & +12 & +9 \end{matrix} = 44.$$

Larger values of S correspond to contingency tables further from H_0 in the direction of H_a. For example, the following table clearly is closer to H_0 than the previous one (since frequencies have shifted up the columns) and has, correspondingly, a smaller value of S:

$$\begin{matrix} 7 & 7 & 5 \\ 8 & 3 & 0 \\ 7 & 10 & 5 \end{matrix} \quad = \quad \begin{matrix} 7 & -7 & -5 \\ 8 & 3 & -0 \\ 7 & 10 & 5 \end{matrix} = 28.$$

We can generate the permutation distribution of S:

- Assign the ranks at random, first to infants, then to handlers, constraining rank distributions according to the original data,
- re-form the 3-by-3 table,
- compute S for the 3-by-3 table.

Again, using the data from Example 2, we can calculate the precise *P*-value by listing all 36 permutations and the corresponding values of S. Figure 5 gives S values (both 16) for the two examples given there. The sorted values of S are:

-10 -4 -4 -2 0 2 4 6 6 6 6 6 10 10 14 16 16 16
20 20 20 22 22 22 26 26 26 30 30 30 36 36 36 40 40 44.

This value of S = 44 from Figure 4 gives a *P*-value of 1/36 = .028, a more significant result. Figure 5 suggests how we have gained significance. The top example of Figure 5 shows a value of C = 8.410, but a value of S = 16. This table is somewhat extreme in terms of the C statistic because it suggests a lack of independence between rows and columns, but not in the direction of the research hypothesis; low handlers are more inclined to handle low infants in this table. C measures deviations from independence in a multi-directional way, while S focuses on a direction consistent with the research hypothesis.

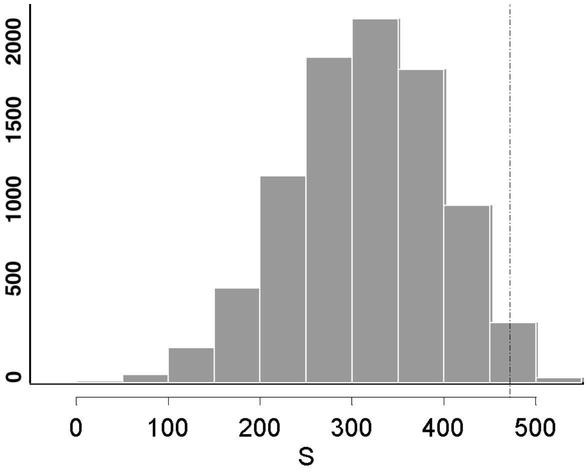


Figure 6: Empirical sampling distribution of the test statistic, S, based upon 10,000 iterates. The dashed line marks the value of S = 472 from the data set. Only 166 of the 10,000 values of S are 472 or larger, which gives us an estimated *P*-value of .0166.

Analysis of Table 1

We return now to the problem of ascertaining statistical significance for our original data matrix, Table 1. We concentrate here on the test statistic, S, which measures deviations in the direction of the research hypothesis. Because we now have 11 infants and 23 handlers, enumeration of the null distribution is infeasible. Thus we computed an empirical approximation to the null distribution of S using S-Plus 2000 on a Compaq Deskpro computer. The value of S for the 3-by-3 table in Figure 1 (original data) is 472. Figure 6 shows the empirical distribution of S using 10,000 iterates; notice that 472 is clearly in the right-hand tail, suggesting fairly strong evidence against the null in favor of the research (alternative) hypothesis. Indeed, only 166 of the 10,000 values of S in the empirical distribution are 472 or larger, which represents an estimated *P*-value of .0166. The small P-value gives substantial evidence against the null hypothesis in favor of the research hypothesis that handler rank is positively correlated with infant rank. At the same time, it is interesting to

Table 3: Empirical *P*-values for interactions broken down by All types, Passive, Unsuccessful, and Successful.

Type of interaction	*P*-value	95% Margin of Error
All types	.0166	.0025
Passive	.0118	.0021
Unsuccessful	.0123	.0022
Successful	.3885	.0096

P-values are based upon 10,000 iterates computed by S-Plus 2000 on a Compaq Deskpro computer and margins of error are computed using the common formula for the confidence interval for a sample proportion. We have evidence for our research hypothesis with all categories except the successful interactions.

note that the permutation test *P*-value is less convincing than we obtained initially using the conventional chi-square test. In that case, the test statistic was C = 30.76 and with 4 degrees of freedom the *P*-value was 3.43×10^{-6}, which is much more convincing than the permutation analysis, but it is also inappropriate, given the assumptions required of the chi-square test.

We now return to the question of different types of interactions between handlers and infants. Table 3 gives the permutation results for different types of handler-infant interactions. Evidence for the research hypothesis is strong for both passive and unsuccessful interactions (*P*-values of .0118 and .0123, respectively), but not for successful interactions. Passive interactions turn out to be difficult to interpret, since it is not clear that passive interactions are real attempts to interact with infants. It was of interest to the researcher that infant and handler ranks are positively associated for unsuccessful interactions but not for successful ones.

Summary

We have seen permutation tests provide a viable alternative for contingency table analysis where a conventional chi-square analysis would be invalid because basic assumptions are violated. Permutation tests not only provide an alternative, but also provide the flexibility of more powerful test statistics. Readers interested in learning more about permutation tests and of their range of applicability should consult Efron and Tibshirani (1998), Good (2000), and Manly (1991).

References

Bentley-Condit, Vicki K., Thomas L. Moore, and E. O. Smith (2000), "Infant Handling by Tana River Adult Female Yellow Baboons (*Papio Cynocephalus Cynocephalus*)," (submitted to *American Journal of Primatology*)

Efron, Bradley and Robert J. Tibshirani (1998), *An Introduction to the Bootstrap,* Boca Raton: Chapman & Hall/CRC.

Fienberg, Stephen E. (1983), *The Analysis of Cross-Classified Categorical Data (2nd edition),* Cambridge, MA: MIT Press.

Fisher, R. A. (1937), *The Design of Experiments (2nd edition),* Edinburgh: Oliver and Boyd.

Good, Phillip (2000), *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses (2nd edition),* New York: Springer-Verlag.

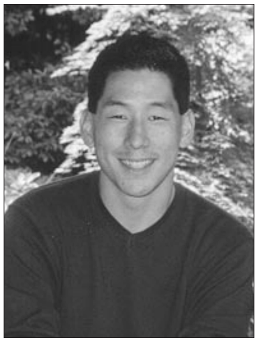
Manly, Bryan F. J. (1991), *Randomization and Monte Carlo Methods in Biology,* London: Chapman and Hall.

Notes

1. We constrain permutations so that both handler and infant ranks retain the same distributions as in the original data. One could instead constrain just the handler ranks and allow the permutation on handlers to induce ranks on infants through their mothers. Constraining both distributions is analogous to constraining both margins in Fisher's exact test and seems more natural to the authors.

Student Voices

Reflections from the NSF/ Monbusho Summer Program in Japan



Jimmy Doi

H*ajimemashite! Dozo yoroshiku onegai shimasu!* ("How do you do? It is a pleasure to meet you!")

Last summer, I had the privilege of participating in a research program in Fukuoka, Japan, as part of the National Science Foundation (NSF) Summer Programs in Japan, Korea, and Taiwan. During this eight-week program, participants have the opportunity to get a glimpse of their fields of research in a foreign setting, to build contacts with researchers for possible future collaboration, and to be exposed to a completely new language and culture. This program is extremely worthwhile both on an academic and personal level, and it is my hope that by the end of this article I will be able to compel at least some statistics graduate students to consider this wonderful opportunity.

The NSF Summer Programs in Japan, Korea, and Taiwan consist of four components: the Summer Institute in Japan, the *Monbusho* Summer Program in Japan, the Summer Institute in Korea, and the Summer Institute in Taiwan. I took part in the second program (officially titled "Research Experience Fellowships for Young Foreign Researchers"), which was established in 1995 by the Japanese Ministry of Education, Science, Sports, and Culture (*Monbusho*). The two NSF programs in Japan are similar, although there is a difference in language training (the Summer Institute has a more intensive program) and in host institution placement. For the Summer Institute, assignments are mainly in the form of internships at government and corporate laboratories in the Tokyo or Tsukuba areas. For the *Monbusho* Program, participants are usually assigned to national universities or inter-university research institutes. The Summer Institute participants are all from the US, whereas the *Monbusho* Program participants are from various parts of the world (US, France, Germany, and the UK). This year there were about 80 *Monbusho* participants representing disciplines such as anthropology, biology, chemistry, engineering, psychology, and statistics (*tokeigaku* in Japanese).

At this point, I would like to dispel a popular misconception about the Summer Programs in Japan.

MYTH: You must know Japanese to participate in the program.

FALSE! Although there were many participants who had studied the language, the majority had not taken a formal course and spoke little or no Japanese. However, most Japanese researchers are quite fluent in English, so language should not prove to be a major obstacle in terms of making progress in research. A lack of language skills may prove to be a challenge as you go through the day-to-day life in Japan – but that's all part of the experience! Of course, the more of the language you know, the better off you'll be, but fluency is not mandatory, and all participants seemed to do just fine. Since my formal studies in Japanese began at the grade school level, and since I was brought up in a Japanese household, I was pretty comfortable in my new environment, and I had the good fortune of being able to communicate with most people I met (although it wasn't always easy!).

Author note: Jimmy Doi is pursuing his Ph.D. in statistics under the advisement of Dr. Roger Berger at North Carolina State University. He earned his Master of Statistics at NC State and a BA in mathematics at California State University, Northridge. His research interests include hypothesis testing applications to biostatistics, bioequivalence, and statistics education. While at NC State he has worked as a teaching assistant, editorial assistant for the Journal of Statistics Education, and instructor for undergraduate statistics courses. After completing the doctorate, he hopes to secure a faculty position in a university-level mathematics/statistics program. His email address is: jadoi@unity.ncsu.edu

During the first week of the *Monbusho* Program, we participated in language and cultural training sessions at the Graduate University for Advanced Studies (*Sokendai*). After separating into class levels corresponding to our respective language abilities, we spent three days learning about the Japanese culture and language. For those new to Japanese, there was certainly no expectation for participants to master the language in only three short days. However, the instructors tried to make the best use of this limited time by teaching helpful vocabulary and expressions to get through daily life – for example, the all-important and essential phrase "*Sumimasen, o te arai wa doko desuka?*" – "Excuse me, where is the *restroom*?"

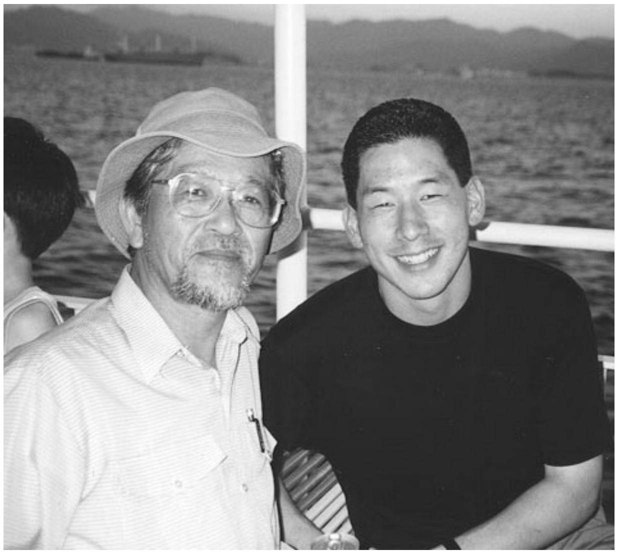


Figure 1. Host advisor Dr. Takashi Yanagawa, Professor of Statistical Sciences, Kyushu University.

Aside from the language classes, we also participated in cultural activities that included a visit to the famous and picturesque area of Kamakura and a cultural exposition at *Sokendai* showcasing *origami*, *shodo* (calligraphy), and *chado* (tea ceremony). For most participants, the highlight of this first week was the homestay experience. For a period of two days, we stayed with a host family, and we had the opportunity to experience life in a Japanese household. In addition to interacting with family members and learning about the daily Japanese lifestyle, most families gave participants tours of the local sights. My host family treated me to a visit to a refreshing *onsen* (natural hot spring) and the famous Ramen Museum in Shin-Yokohama. (Yes, a *ramen museum* – it's far more interesting than it sounds!) Although my homestay experience was brief, the time we spent together was wonderful, and I enjoyed getting to know each of the family members. I am certain we will continue to keep in touch for a long time to come.

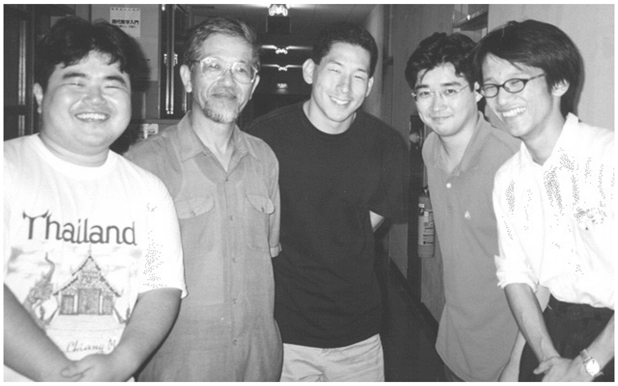


Figure 2. Graduate students and host advisor in the halls of Kyushu University. L to R: K. Yonemoto, Dr. T. Yanagawa, J. Doi, S. Imoto, and Y. Takita

After returning from our homestay experiences and completing our first introductory week, we prepared for departure to our respective host research institutions. Assignments were particularly widespread this year, as they spanned from the northern area of Hokkaido to the southern islands of Okinawa. I flew to the southern region of Japan known as Kyushu and landed in the prefecture of Fukuoka. I spent the following seven weeks at my host institution Kyushu University, one of several universities in Japan active in statistics research. Dr. Takashi Yanagawa, one of the well-known statistics faculty members at Kyushu University, kindly agreed to serve as my host advisor during this program.

Dr. Yanagawa is known as one of the leading biostatisticians in the country, but his research interests are quite broad and include areas such as multivariate discrete data analysis and nonlinear/chaotic time series. The research topic we selected was more related to biostatistics. Along with one of Dr. Yanagawa's graduate students, Masahiro Makishita, our research subject was based upon ongoing work by Sakata, Yanagawa, and Fukuichi (2000) entitled "The $\hat{q}$ Value and its Application to the Determination of the No-Observed-Adverse-Effect Levels in Dichotomous Response." To provide background for this research, let X and Y be independent binomial random variables. The sample size for X is n_1, and the success probability is p_1. The sample size for Y is n_2, and the success probability is p_2. Consider cases such as toxicology studies where the usual set of hypotheses are H: $p_1 = p_2$ versus K: $p_1 < p_2$. In this setting, the problem of controlling the type-II error rate is of main concern. In an effort to help control the type-II

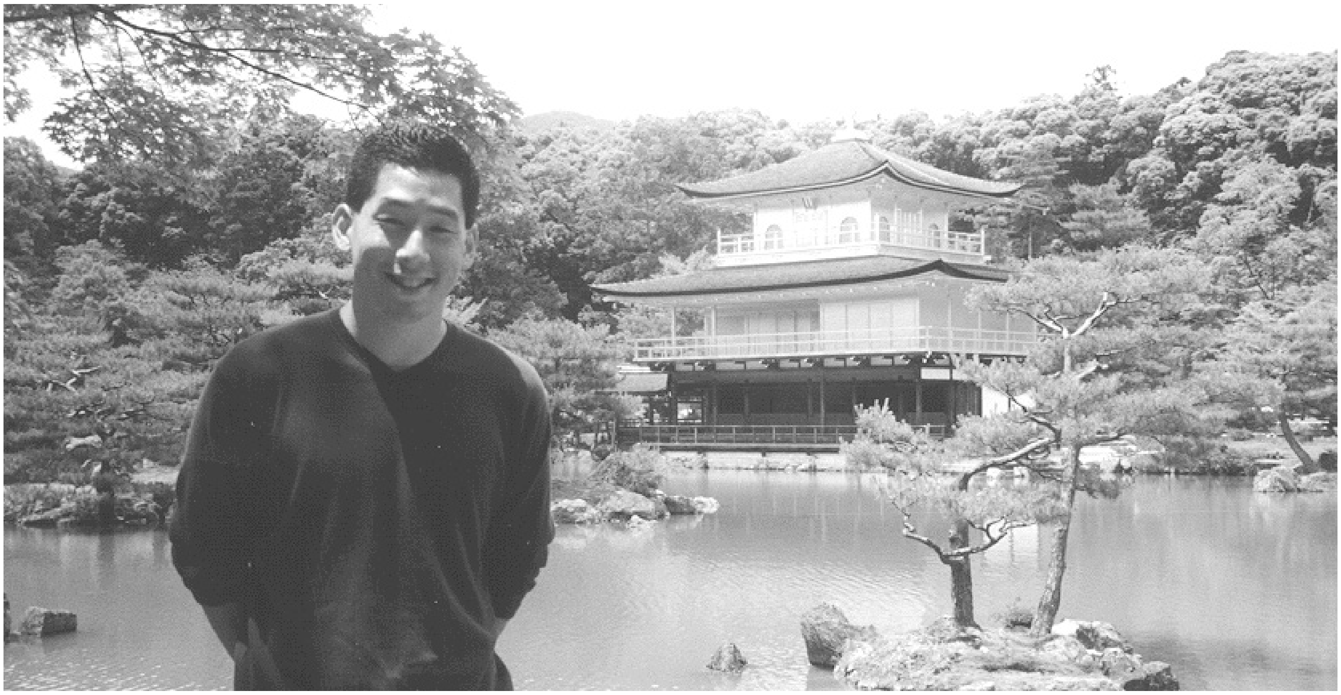


Figure 3. The famous golden pavilion Kinkakuji in Kyoto.

error rate, Dr. Yanagawa and his colleagues have proposed a statistic called the $\hat{q}$ value, an analogue of the p value. During the research stage of my program, we examined the properties of the $\hat{q}$ value and examined size considerations of the test based upon $\hat{q}$. Due to the presence of a nuisance parameter, we tried to incorporate the confidence interval p value method proposed by Berger and Boos (1994) and to extend the idea to address the smallest change point problem for binomial proportions.

Although most of the seven weeks were dedicated to conducting research at Kyushu University, I had the opportunity to visit a number of institutions during this time and to interact with other statisticians. One of the places I visited was the Institute of Statistical Mathematics (ISM) in Tokyo. Consisting of four major departments (Fundamental Statistical Theory, Statistical Methodology, Prediction and Control, and Interdisciplinary Statistics), the ISM has been one of the main leaders of statistics research in Japan for many years. After I gave a brief presentation on my preliminary work at NC State University, I had the chance to meet several ISM faculty members and graduate students and to learn about their broad range of research interests and projects.

Another one of my visits was to Hiroshima University, where I had the opportunity to meet faculty and graduate students from the statistics and the biostatistics research groups. The areas of research in both groups were quite diverse, and some were specifically aimed at the statistical analysis of radiation effects data stemming from the 1945 atomic bomb. Such research has also been a focus at the site of my next visit, the Radiation Effects Research Foundation (RERF). The RERF is a bi-national research institution managed by both Japan and the US and has continued health follow-up studies of atomic bomb survivors started over 50 years ago. The RERF consists of many departments, including the departments of clinical studies, epidemiology, genetics, and radiobiology. I had the chance to visit their statistics department and meet some of the lead statisticians. As I spoke with these researchers and learned about their many ongoing projects, it was immediately clear the RERF offers a rich source of many fascinating statistics research problems. Many who I encountered expressed an interest in recruiting students from places like the US. There is a need for more statisticians to work on these projects, especially people with a strong applied background. For more information about the RERF and possible postdoctoral and internship opportunities, visit the RERF web site (listed below).

In comparing the statistics environment in the US to that in Japan, something I found interesting was the fact that there was not a single department of statistics within a university throughout Japan. This was quite surprising given the active level of statistics research and the important contributions by people like Taguchi and Akaike over the past 50 years. However, with the current growth of statistics research, their widespread applications, and the expansion of the statistics community nationwide, researchers are optimistic about the establishment of a department of statistics or biostatistics in the near future.

Although research is a main component of the *Monbusho* Program, it is certainly not *all-work-and-no-play*! Aside from conducting research, participants are encouraged to take personal time, not only to visit other research institutions, but also to do some sightseeing to get to know this beautiful country, which many of the participants are visiting for the first (and possibly last) time. Many of this year's participants took trips to famous and historic sites such as Kyoto and Nara. Also, several participants were able to plan a joint hiking expedition up Mount Fuji, which I heard was a great success! During my stay at Kyushu University, the graduate students and faculty members extended generous hospitality to me. They regularly took time out of their busy schedules and invited me on trips to local scenic attractions, the downtown area of Tenjin, and *many* delectable dining excursions (which surely led to some weight gain!). As I traveled through the country on my own, one of the most memorable and moving experiences was my visit to the Hiroshima Peace Memorial Museum. If you ever have the chance to visit Japan, *do not* miss the opportunity to see this exhibit. It will surely have an impact upon you.

My participation in the *Monbusho* Summer Program was, without a doubt, one of the best experiences of my life, both on an academic and personal level. Through my interaction with Dr. Yanagawa and our research activities, I have been able to develop new ideas and a stronger foundation for work related to my dissertation. Through the many contacts I have been able to establish at Kyushu University, Hiroshima University, the ISM, and the RERF, I believe there is potential for future collaborative work and possible postdoctoral or visiting position opportunities. I am grateful to the National Science Foundation and *Monbusho* for their outstanding support and the opportunity to be a part of this unique program. Without hesitation, I would recommend this program to anyone — even if you have *never* studied Japanese. It provides the special opportunity to conduct research in a foreign setting, and it will probably be one of the last chances for a graduate student to have such an experience before entering the job market. I would especially encourage graduate students in statistics to participate. As I have looked over the backgrounds of *Monbusho* participants from previous years, I believe I was one of the first participants whose area of research is statistics. As I encountered researchers across the nation, many expressed their hope that programs like the *Monbusho* Summer Program will attract more and more statistics students from the US to conduct research in Japan.

The annual deadline for submission of applications is December 1st. Below, I have included some online resources that provide more information about the NSF Summer Programs. Finally, please feel free to contact me through email if you have *any* questions about the program, as I would be more than happy to share my experiences. This program is indeed a *once-in-a-lifetime* opportunity and, if you are able to participate, it will *surely* be one of the best experiences of your life.

References

Berger, R. L., and Boos, D. D. (1994), "*P* Values Maximized Over a Confidence Set for the Nuisance Parameter," *Journal of the American Statistical Association*, 89, 1012–1016.

Sakata, T., Yanagawa, T., and Fukuchi, J. (2000), "The $\hat{q}$ Value and its Application to the Determination of the No-Observed-Adverse-Effect Levels in Dichotomous Response," in preparation.

Online resources

NSF Social, Behavioral, and Economic Sciences – International Programs
http://www.nsf.gov/sbe/int/

NSF Tokyo Home Page
http://www.twics.com/~nsftokyo/

Kyushu University
http://www.kyushu-u.ac.jp/english/index-e.htm

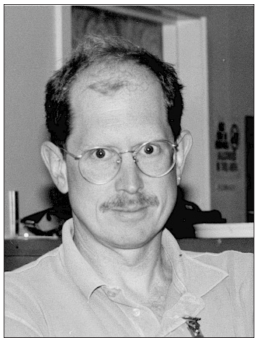
Radiation Effects Research Foundation (RERF) Home Page
http://www.rerf.or.jp

The Institute of Statistical Mathematics
http://www.ism.ac.jp/index-e.html

Home Page of Jimmy Doi
http://www.stat.ncsu.edu/~jadoi

Student Project

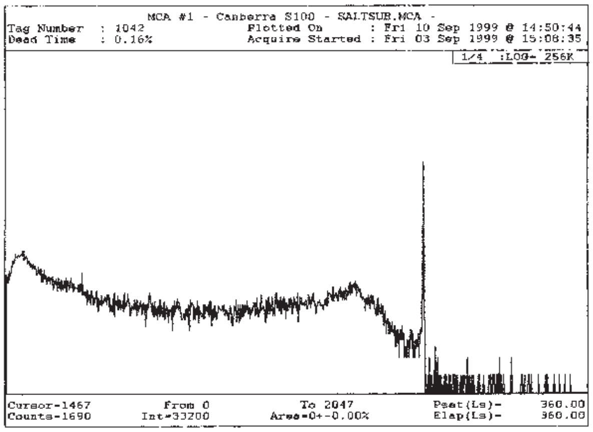
Radioactive Salt and Gamma Spectroscopy



Thomas Gwise

There are ways to identify radioactive elements by looking at the radiation they emit. Through investigating some table salt substitute, I will guide you through one of these processes. Salt substitute, a food seasoning that uses potassium chloride in place of sodium chloride, was chosen as a subject because it is fairly common and 0.0117% of all potassium is the naturally occurring radioactive isotope of potassium, potassium-40 (K-40). Figure 1 shows the energy peak from radioactive K-40 in a sample of table-salt substitute. The graph is output of a gamma ray spectroscopy program.

Gamma spectroscopy is a computationally intense technique for identifying elements by the radiation they emit. It is employed in a diverse group of fields, from oil exploration deep under the oceans to astronomical investigations of the heavens. As a student of statistics, I find gamma spectroscopy particularly interesting because it is so dependent on statistical methods. Every time a gamma spectrum is analyzed a series of statistical applications, from differencing processes to control charting and goodness of fit tests, are performed. By the end of this article, one will see that commercial gamma spectroscopy programs are in essence specialized statistical packages locked into one data format.



```
                 MCA #1 - Canberra S100 - SALTSUB.MCA -
Tag Number   : 1042              Plotted On      : Fri 10 Sep 1999 @ 14:50:44
Dead Time    : 0.16%             Acquire Started : Fri 03 Sep 1999 @ 15:08:35
                                                              1/4  :LOG- 256K

Cursor-1467         From 0          To 2047         Psat(Ls)-       360.00
Counts-1690      Int-33200      Area-0+-0.00%       Elap(Ls)-       360.00
```

Figure 1.

Gamma Spectroscopy

The best way to start this tour is with a clear description of gamma spectroscopy. Gamma spectroscopy, as the name implies is the investigation of gamma spectra. What are gamma spectra? We will review some basic physics to make the rest of the explanation understandable to those unfamiliar with the terminology. Gamma rays are a subset of the electromagnetic spectrum. Recall that the electromagnetic spectrum consists of all forms of electromagnetic radiation. Visible light and radio-waves are examples of electromagnetic radiation. The whole electromagnetic spectrum is made up of the following types of radiation in order of ascending energy: radio waves, microwaves, infrared radiation, visible light, ultraviolet light, x-rays and gamma rays. The different kinds of electromagnetic radiation are really all the same basic stuff. Electromagnetic radiation can be thought of as little packets of energy called photons. The energy levels of photons determine their common names. For example, a photon with an energy level of 200 kilo electron volts (kilo eV or keV) will be called a gamma ray. Another way in which we refer to subsets of the spectrum is to name the different energy levels according to their wavelengths. The equation $E = hc / \lambda$, where h = Plank's constant and c = speed of light, relates photon energy to wavelength. Convenience normally determines how one identifies waves in the spectrum. Referring to 100 kilo hertz (3000 meter) radio-waves as 4.1E-10 eV waves would be cumbersome.

Back to gamma spectroscopy. A gamma spectrum can be thought of as a type of organized photograph. In a photograph the light photons interacting with the film over a controlled time

Thomas Gwise wrote this article as part of a special project while working on his MS in Statistics at the California State University at Hayward under the guidance of Dr. Bruce Trumbo. Mr. Gwise is a Health Physicist at The Stanford Linear Accelerator Center and does gamma spectroscopy as part of his job.

Table 1

Common Name	Wavelength (m) (Approximately)	Energy (eV) (Approximately)
Gamma Rays	< 1 x 10^{-10}	> 1.2 x 10^{4}
X-Rays	1 x 10^{-15} to 1 x 10^{-8}	124 to 1.2 x 10^{9}
Ultraviolet	1 x 10^{-8} to 4 x 10^{-7}	3.1 to 124
Visible Light	4 x 10^{-7} to 7 x 10^{-7}	1.77 to 3.1
Infrared	1 x 10^{-6} to 1 x 10^{-4}	1.24 x 10^{-2} to 1.24
Microwaves	1 x 10^{-3} to 3 x10^{-2}	1.24 x 10^{-5} to 1. 24 x 10^{-3}
Radio Waves	1 x 10^{-2} to 3 x 10^{3}	4.1 x 10^{-10} to 1.24 x 10^{-4}
Electric Waves	> 3 x 10^{3}	< 4.1 x 10^{-10}

period causes an image to be recorded. A gamma spectrum is the result of recording gamma photons interacting with a crystal over a controlled time period. The big conceptual difference is that in gamma spectroscopy we organize the photons according to energy level rather than form a picture. To collect a spectrum we use a detector, a series of amplifiers and a computer. Photons interact with the detector creating electronic pulses proportional to their respective energy levels. These pulses go through a series of amplifiers and then to a computer that tallies them in bins according to their pulse size. At the end of a prescribed collection time, each bin will contain a number corresponding to the number of photons counted for a given energy level. The results are usually represented graphically. An example of such a graph is seen on the opening page. The graph resembles a histogram with the energy level as the x-axis and the number of photons detected in each energy gradation as the y-axis. The spectroscopy software used here does not label the axes, but the display does show that for the selected channel (number 1,467) 1,690 events have been recorded. Generally, the elements of the graph are some gamma peaks and what is called a

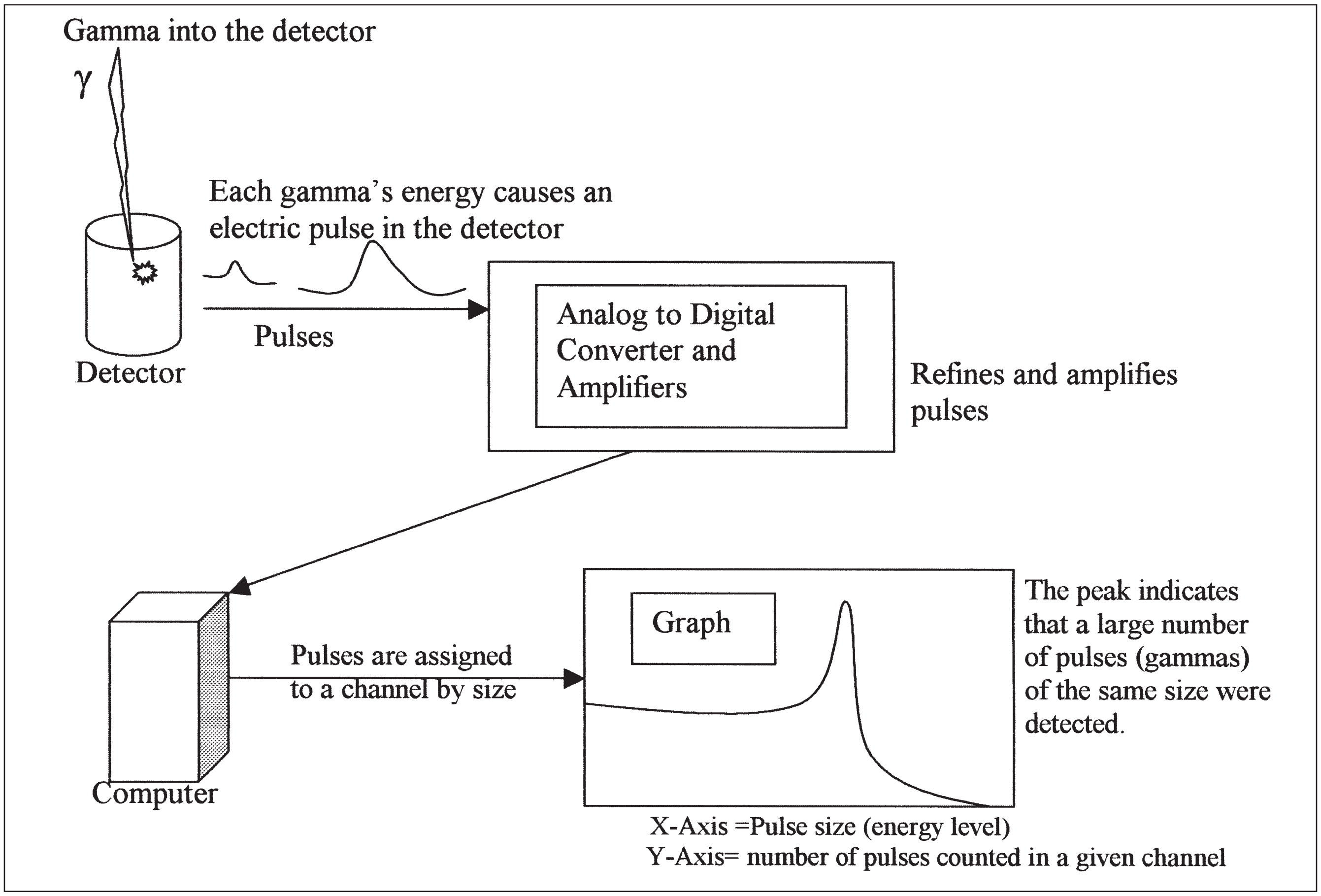

Gamma into the detector
γ
Each gamma's energy causes an electric pulse in the detector
Pulses
Detector
Analog to Digital Converter and Amplifiers
Refines and amplifies pulses
Computer
Pulses are assigned to a channel by size
Graph
The peak indicates that a large number of pulses (gammas) of the same size were detected.
X-Axis =Pulse size (energy level)
Y-Axis= number of pulses counted in a given channel

Figure 2.

continuum. The continuum, without being too technical, is the rest of the graph made up of background radiation and residual energy from photons not completely absorbed in the detector. Existence of a peak in the graph indicates the detection of a large number of gamma rays of the same energy level. The flow chart, Figure 2, should help to visualize the process.

Gamma spectroscopy is useful because it can detect and identify gamma rays. Many radioactive isotopes of elements, or radionuclides, have unique gamma ray signatures. From a gamma spectrum it is often possible to tell the type and amount of radioactive element present in a sample. This fact makes gamma spectroscopy a useful tool in many fields. A geologist may want to know the quantity of thorium in a certain area. A medical technologist may need to know if he has been exposed to the radioactive iodine recently administered to a patient. Astronomers use a similar technique to investigate far off stars. In fact, different variations of this technique turn up in many surprising places.

Statistics Employed

Now that we know the nature and some uses of gamma spectroscopy, we will discuss some of the statistics involved. Since the object of the process is to match energy levels with peaks, we need to calibrate the computer's pulse counting bins, or channels, with respect to energy. Recall that pulses are sorted into bins according to their size and pulse size is proportional to energy. In our machine, the ratio of channels to keV is approximately one channel per keV. A gamma calibration source emitting gamma rays with energies of 59.5, 88, 122, 166, 279, 662, 898, 1173, 1332 and 1836 keV was placed on a germanium detector and a count, the term used for acquiring data, was taken for a period of 1000 seconds. Five more counts were taken. Each spectrum was inspected to determine the channels at which each peak had a maximum.

The channel numbers and the energy level vectors were entered into the analysis program Minitab and linear regression was then used to find a function relating energy to channel number. One must remember that the calibration problem is slightly more complicated than a standard linear regression problem.

I will restate the situation. In sample analysis, we have data stored in bins. We would like to assign an energy level to each channel. Using a known source of gamma rays and linear regression, we can obtain a representation of channel as a function of energy level. That is not the end. We need to identify the sample gamma rays based on channel number. To do this we invert the regression equation. It is important to note that this is not the same as using channel as the predictor and energy as the response in the initial regression. Our regression equation is:

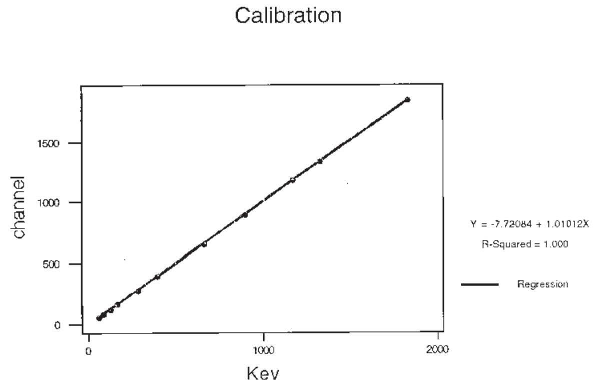


Figure 3.

$$\text{Channel} = -7.72084 + 1.01012^{*}\text{Energy}.$$

On the graph above, Y values represent channel numbers and X values represent energy levels in Kev. Solving this equation for energy, we get

$$\text{Energy} = .98998^{*}\text{Channel} + 7.643488.$$

When analyzing specimens of unknown radioactive material, we use the second equation above to assign energy levels to peaks found in the spectrum. Once the value for energy is found, a confidence interval for that value can be calculated. To see the details of these computations and for further discussion of regression as it is used in calibration see *Regression Analysis, Concepts and Applications* by Graybill and Iyer (pp. 427–431).

When viewing the graph of our original regression, it may seem that the line looks "too perfect." One should note that the amplifiers used for gamma spectroscopy are designed to have this linear property.

The plots in Figure 4, labeled "Calibration Residuals" and "Calibration Residuals vs. Fits" are for the regression above where "Channel" is the response variable. Inspection of these residual plots indicates that our standard regression assumptions of normal data and constant variance appear to hold.

Other parameters requiring calibrations are peak shape and efficiency. The fact that energy absorption properties in the detector will cause the peaks to be wider at the bottom for higher energy gamma rays necessitates a calibration of peak shape with respect to energy. Efficiency calibration is the standard process of relating a measurement to a known value. Because many radiation sources are isotropic, that is, they are emitted in equal number in all directions, only a fraction of their output can

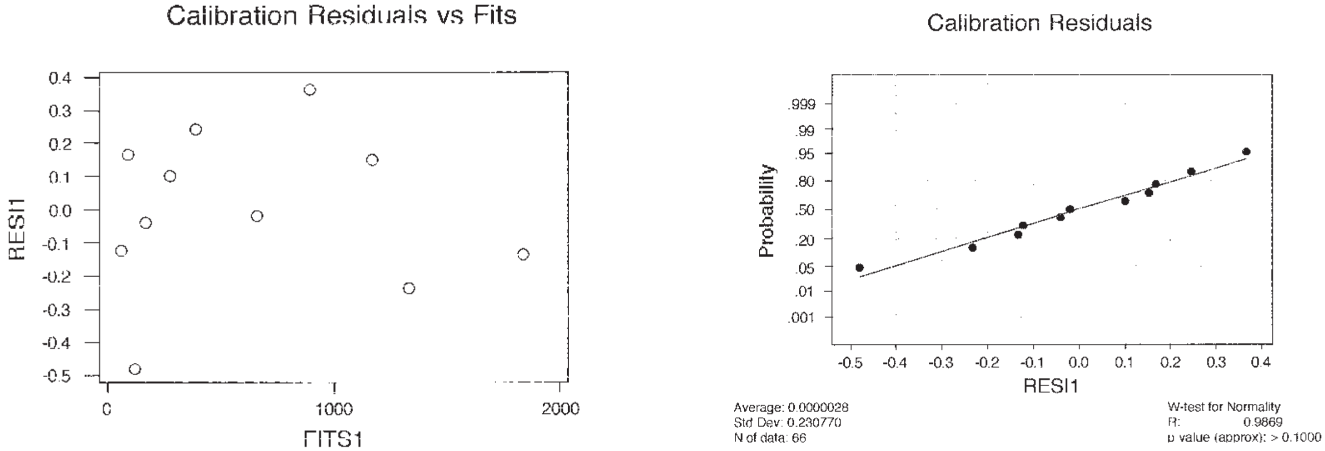


Figure 4.

be incident on the detector. These two calibrations are more complicated than energy calibration and sometimes employ nonlinear regression procedures.

Now that we have our orientation with respect to energy, we can look at the rest of the analysis process. Many of the procedures used in gamma spectroscopy are complicated and a thorough discussion here would be very lengthy. In the interest of brevity, I will list, with a short description, some of these computations, many of which are statistical in nature. Most spectroscopy software packages run through a set of procedures similar to the following.

Peak Search

The second difference of the spectrum is obtained through a process similar to the differencing done in time series analysis. This process is analogous to finding the second derivative. Using the second difference, we are able to locate relative maxima, the peaks, in the spectrum.

Fitting

Least squares procedures identical to those used in linear regression are performed to fit continuous functions to the discrete data.

Integration

The area under each peak is proportional to the number of gamma rays counted for that energy. Numerical integration can be used to determine the true number of photons present.

Calibration

The number of counts in a peak can be related though the efficiency calibration to the units we desire. In other words, depending on our sample, we can get a number of photons emitted per unit area or volume or mass, etc.

Goodness of Fit

Chi-square fit tests are automatically performed on the functions to alert the analyst to poorly fitting functions.

Propagation of Errors

The uncertainty in the count is determined and combined with the uncertainty of the calibrations to report a confidence interval or "error" with the measurement results.

Quality Control

Many commercial spectroscopy programs automatically keep quality control charts of key parameters.

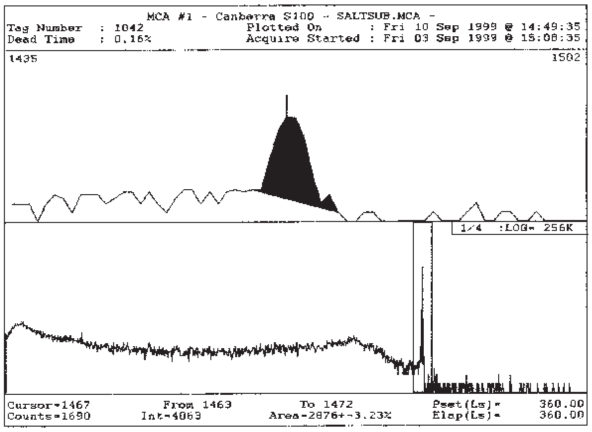
Minimum Detectable Calculations

Minimum detectable concentrations or minimum detectable activities (MDA's) are determined for each sample count through a procedure based on the t-test showing what could loosely be thought of as the "power" of the analysis.

Data Collection

Aside from the statistical procedures, another aspect of gamma spectroscopy that may interest statisticians is the probabilistic nature of the data collection. The process of amassing data into a spectrum is a Poisson process. Recall that a Poisson process is a counting process with independent increments, and that the number of events counted in any interval of a given length has a Poisson distribution. Gammas arriving at a detector can be shown to meet these conditions. Further explanation of the Poisson process can be found in *Introduction to Probability Models*.

Another very interesting fact is that for a given gamma energy, the energy absorbed in the crystal will not be exactly the same, but will vary slightly. This is caused by certain characteristics of the detector crystal. The slight differences in detected energy level will cause peaks to have the shape of



```
                MCA #1 - Canberra S100 - SALTSUB.MCA -
Tag Number   : 1042             Plotted On       : Fri 10 Sep 1999 @ 14:49:35
Dead Time    : 0.16%            Acquire Started  : Fri 03 Sep 1999 @ 15:08:35
1435                                                                    1502
                                                   1/4  :LOG= 256K
Cursor=1467       From 1463        To 1472          Pset(Ls)=     360.00
Counts=1690     Int=4063        Area=2876+-3.23%    Elap(Ls)=     360.00
```

Figure 5.

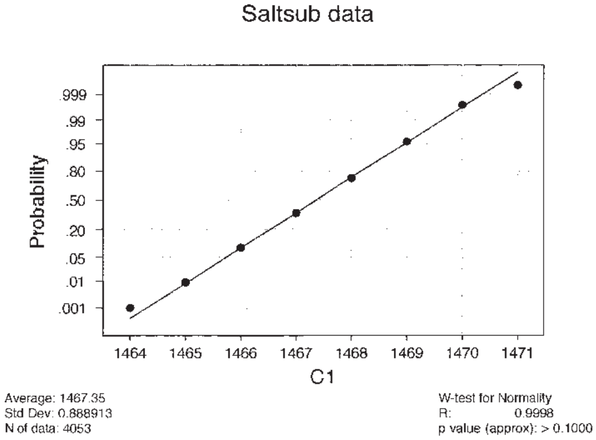


Figure 6.

the normal distribution. That means for a given gamma energy, the probability of collecting a certain number of counts over time is distributed as Poisson and the exact location on the x axis of the counts representing that gamma ray energy are distributed normally with the "true" energy level as the mean.

The plot of the salt substitute spectrum in Figure 5, with a close-up of the region between channels 1435 to 1502 in the upper frame, shows the Gaussian shape of the K-40 peak. To more closely investigate the shape of the K-40 peak, the data in the peak region were analyzed using Minitab. Keep in mind that we are investigating the energy spectrum, which is continuous, but our data are discrete. This fact will be important when choosing a test of normality. First, we look at a histogram of the data.

The shape of the histogram appears to be close to normal. Notice that the average of the peak data, when entered into the regression equation gives a result of 1460.3 keV. The energy level for K-40 given by The Table of Isotopes (Firestone et al., 1999) is 1460.8 keV. Our calibration is fairly accurate considering that the peak bounds, on which the mean will depend, were determined by eye rather than using an analytical method. Figure 6 is a normal plot of the data. The plot indicates that there is only a slight deviation from normality in the tails of the peak. This deviation is explained in the theory of charge collection in detectors, and the transition from the peak to the continuum portion of the spectrum. Because the spectrum is continuous, but the data are discrete and forced into a small finite set of bins, the Anderson-Darling test of normality, based on the empirical distribution function and the default normality test in Minitab, indicates that the data are far from normal with a p-value of 0.000. The Ryan Joiner test of normality, which is based on a regression fit of the normal plot, gives an *R*-value of .9998 and a *p*-value reported as greater than .1. This seems to be in better agreement with the plot and supports our belief that the data is distributed normally.

To read more about the normality tests mentioned here, see *Goodness-of-Fit Techniques*, edited by D'Agostino and Stephens. Also, another reference is a paper entitled *Normal Probability Plots and Tests for Normality* by Thomas A. Ryan and Brian L. Joiner (1976) which is available on the Minitab website (*http://www.minitab.com/resources/whitepapers/index.htm*).

Measurement Results

To round out the tour, we will compare our measurement results with some standard reference information. The amount of K-40 we detected in

Table 2. Character Histogram; Histogram of C1, N = 4053 Each * represents 35 obs.

Midpoint	Count	
1464	6	*
1465	61	**
1466	540	****************
1467	1690	***
1468	1423	***
1469	313	*********
1470	18	*
1471	2	*

our analysis agrees with the data found in the periodic table and the chart of the nuclides. I have outlined the information and calculated an estimate for the amount of radioactivity that is expected in our sample of the salt substitute.

- .0117% of natural potassium is radioactive potassium 40 (periodic table).
- Potassium is 50% of the salt substitute (from product label).
- Our sample of salt substitute weighs 578 grams.
- The specific activity of K-40 is 1E-5 Curies/gram (Ci/gram). (*Handbook of Health Physics*).
- 1 Curie = 3.7E10 disintegrations per second. (*Handbook of Health Physics*).
- Efficiency of our detector for this sample = 0.01066 counts per gamma (prior calibration)
- The counting time for this sample was 360 seconds.
- The number of counts in the 1460.8 keV peak was 4053.
- The number of 1460.8 keV gammas per K-40 disintegration =0.1067 (*Handbook of Health Physics*).

.000117 * 578 grams*(50%) = .033813 grams of K-40 [percentage of K-40 in natural K]
.033813 gram*1E-5 Ci/gram [grams K-40* specific activity]
=3.38E-7 Ci [K-40 radioactivity in the sample]
= 338,130 pico Ci. (pCi) [unit conversion]
338,130 pCi / 578grams = 585 pCi/gram [K-40 radioactivity per gram of specimen]

Based on standard reference material, the amount of K-40 radioactivity we expect to see in a gram of salt substitute is 585 pCi.

Our measurement of the sample showed:
4053 counts / 360 seconds = 11.2583 counts/second
(11.2583 counts/second) / (0.01066 counts/gamma) = 1056.129 gammas/second
(1056.129 gammas/second) / (0.1067 gammas/disintegration) = 9898.11 disintegrations/second
(9898.11 disintegrations/second) / {3.7E10 (disintegrations/ second)/Ci}
=2.68E-7 Ci
(2.68E-7 Ci) / 578 grams = 4.63E-10Ci/gram
=463 pCi/gram

The result of our measurement is 463 pCi/gram. That translates to about 16 potassium 40 atoms self-destructing every second per gram of our sample. A commercially produced software package calculated a value of 437 pCi/gram ± 5.6%. The confidence interval is calculated by considering the variances for all measurement and calibrations. The results of our measurement are within reasonable agreement with the references when one considers the allowed variability on food labels. So yes, salt substitute contains small amounts of radioactivity, but so do all foods high in potassium, such as bananas.

Conclusion

With the aid of some salt substitute, we have taken a brief look at the intersection of statistics and gamma spectroscopy. We discussed the properties of gamma counting which allow us to model it as a Poisson process. The shapes of the gamma peaks being investigated are Gaussian. In fact, almost all of the tools used to analyze the data are statistical methods. Even the analysis report should please the statistician. Accompanying all measurements are well-founded confidence intervals and in the MDA value, an indication of the "power" of the analysis. This brief introduction to gamma spectroscopy shows why I say that this little corner of the physics world is almost all probability and statistics.

References

D'Agostino and Stephens, (eds.) (1986) *Goodness-of-Fit Techniques*, Marcel Dekker, Inc.

Firestone, R.B., Shirley, V.S., Baglin, C.M., Chu, S.Y.F, and J. Zipkin (1999) Table of Isotopes, 8th Edition, Wiley, with CD-ROM.

Gilmore, G. and Hemingway, J. (1995) *Practical Gamma Ray Spectrometry*, Wiley.

Graybill, F. A. and Iyer, I. K. (1994) *Regression Analysis, Concepts and Applications*, Duxbury Press.

Knoll, G. F. (1989) *Radiation Detection and Measurement*, second edition, Wiley.

Mariscotti, M. A. (1967) "A Method for the Automatic Identification of Peaks in the Presence of Background and its Application to Spectrum Analysis" *Nuclear Instruments and Methods* 50 309–320

Ross, S. M. (1997) *Introduction to Probability Models*, Academic Press

Routti, J. T. and Prussin, S.G. (1969) "Photopeak Method for the Computer Analysis of Gamma-Ray Spectra from Solid State Detectors" *Nuclear Instruments and Methods* 72 125–142

Ryan, T. and Joiner, B.L., (1976), Normal Probability Plots and Tests for Normality, MINITAB Homepage, URL: *http://www.minitab.com/resources/whitepapers/normprob.htm.*

Shleien, B., et al. (eds.) (1998) *Handbook of Health Physics and Radiological Health,* Williams & Wilkins.

AP Statistics

A Report on the 2000 Exam

Roxy Peck

Introduction

The College Board Advanced Placement Program consists of 33 college-level courses and exams in 19 disciplines designed for highly motivated high school students. Five of these courses are in the mathematical sciences–Calculus AB, Calculus BC, Computer Science A, Computer Science AB, and, most recently, Statistics. This past year, the AP Statistics program's fourth, was another year of remarkable growth. The number of students taking the exam has more than quadrupled from about 7500 in 1997 to more than 34,500 in 2000. Students from 2,242 high schools took the exam, and students requested that scores be sent to 1,381 different colleges and universities. Males and females were nearly equally represented among the exam takers, with women accounting for 49% of the exams.

The AP Statistics course is a forward looking course that is firmly grounded in data analysis, and the AP exam reflects this focus. The exam consists of a multiple choice section and a free response section. The free response section is made up of 5 open-ended questions and one longer investigative task that requires integration and synthesis of multiple concepts. The exam itself is challenging, and requires that students not only employ appropriate statistical methods, but also that they demonstrate statistical thinking and clear communication.

The free response section of the exam was graded at a week-long reading held this year at the University of Nebraska at Lincoln. One hundred and fifty-five university and high school statistics teachers participated in the reading. The free response questions were scored using holistic rubrics that allowed for scores ranging from 0 to 4 on each question.

Roxy Peck is Associate Dean of the College of Science and Mathematics and a professor of Statistics at Cal Poly San Luis Obispo. She is a fellow of the American Statistical Association and the chief faculty consultant for the Advanced Placement Statistics Program.

Exam Results

Table 1 summarizes the scores on each section of the exam, as well as the overall scores. There was a perceived decline in student performance again this year, but this is not too surprising given the rapid growth and the number of schools and teachers participating in the program for the first time this year. Student performance on the multiple-choice section was lower than last year, and lower scores were also observed on a set of 15 multiple choice questions that have been used on previous exams. These questions, called equating questions, are used as a basis for comparison of the current cohort of students with those from previous years. Although the mean score on the free response improved this year, this was primarily due to the fact that there were two questions that involved constructing and interpreting graphical displays and students found these to be fairly easy. In spite of the fact that the mean score was higher, performance on the other four free-response questions, and particularly on the investigative task, was disappointing. We hope to see student performance improving in the coming year.

The Free Response Questions

The free response questions from the 2000 exam covered topics in descriptive statistical methods, inferential statistical methods, probability, and experimental design. Question 1 assessed the student's ability to interpret simple graphical displays. It was a fairly straightforward question, and students tended to score well on it. To receive full credit for this problem, the student needed to describe both the effect of each of two pain relievers and how the effect was related to dose. A complete answer required both a recommended drug and dose, an explanation of why one drug was chosen over the other, and a justification of the selected dose.

Question 2 evaluated the student's knowledge of the basic assumptions necessary for inferential

Table 1: Summary Statistics for Raw AP Scores

	1997	1998	1999	2000
Composite Score (100)				
Mean	46.99	44.15	37.36	37.53
Standard Deviation	18.95	19.57	18.71	14.77
Free Response Section (50)				
Mean	16.62	17.71	15.82	18.34
Standard Deviation	9.67	10.49	10.14	6.89
Multiple choice Section (50)				
Mean	30.47	26.55	21.62	19.26
Standard Deviation	10.45	10.30	9.67	9.22

Raw composite scores are used to determine the score reported to the student (1, 2, 3, 4, or 5). Table 2 gives the distribution of reported scores for the four years the exam has been given. The percent of 5's and 4's is somewhat lower than in previous years, a consequence of the weaker student performance in 2000. Most colleges and universities that give credit for Advanced Placement award credit to students who score 3 or higher on the exam. Some universities only grant credit to those scoring a 4 or 5.

procedures to be valid. To correctly answer this question, the student needed to be able to state the assumptions required for a small sample confidence interval for a population mean and to evaluate whether these assumptions were reasonably met for the scenario described. While many students stated and checked the assumption of normality of the population distribution, to receive full credit for this problem it was also necessary to recognize that a random sample of footprints was not equivalent to a random sample of adults (the population of interest).

Question 3 determined whether a student could construct an appropriate graphical display given discrete numerical data for each of two groups and then use the graphical display to describe the similarities and differences between the distributions for the two groups. There were several different types of graphs that could be constructed using the given data, and any one that allowed for easy comparison of the distributions was acceptable. Students were expected to compare the two distributions with respect to at least two of center, shape, and spread. Students seemed to find this question to be relatively easy, and scores on this question were high.

Question 4 evaluated whether the student could carry out a test of hypotheses and state conclusions in context. To receive full credit on this question, the student needed to state hypotheses, identify an appropriate test procedure, check (not just state) any necessary assumptions, compute the value of the test statistic and the associated *P*-value (or rejection region), and then, based on the result of the test, give an appropriate conclusion in context. The student was also expected to recognize that the study was an observational study and that a cause-and-effect conclusion was not appropriate.

Question 5 assessed understanding of some of the basic principles of experimental design, including randomization, blocking, and the concept of blinding. Many students did not have an adequate understanding of blocking, and as a result had difficulty with this question.

Question 6 was the exam's investigative task. As such, its purpose was to evaluate the student's understanding in several course topic areas and to assess ability to integrate statistical ideas and apply them in a new context. This year's investigative task involved statistical inference, probability and the concept of independence, and bivariate graphical representation. It was a great problem, and the poor student performance on this question was the BIG disappointment of this year's exam.

Some General Comments on Exam Performance

The following comments are offered in the hope that this report will help teachers better understand the expectations of the Advanced Placement Statistics course and assist them in preparing students for the AP Statistics exam.

As was the case in previous years, student performance tended to be stronger on the mechanical and computational aspects of problems than on parts that required interpretation or conceptual understanding. Communication of results was often weak, and many students failed to answer questions in context. (Explanations and conclusions in context were always necessary for a complete answer.) Surprisingly few students were able to give a correct interpretation of a confidence interval.

Assumptions required for inference continued to be a problem area. More students stated assumptions when carrying out a hypothesis test, but they didn't always understand that assumptions must also be checked. Very few

Table 2: Reported Score Distribution

Reported Score	1997	1998	1999	2000
5	15.7	13.7	11.1	9.7
4	22.1	21.4	20.3	21.6
3	24.4	24.6	25.8	22.4
2	19.7	18.6	20.9	20.6
1	18.0	21.8	21.9	25.7

recognized that assumptions also need to be addressed when constructing confidence intervals.

Conclusions

The implementation and acceptance of the Advanced Placement Statistics program has been one of the most important developments in statistics education at the high school level in many years. In spite of the fact that the program is experiencing some “growing pains,” those involved with the program agree that it has been a great success. Many people have contributed to its current success–particularly the large number of high school teachers who have risen to the challenge of teaching a new and demanding course. They have combined enthusiasm and excitement with hard work to make the AP Statistics course a challenging and worthwhile course for their students. With the statistics community becoming more involved with the AP program and teachers gaining more experience with data analysis and some of the more abstract concepts of statistics, the future for AP Statistics looks bright indeed!

For more information on the AP Statistics Program, including the course description, released exam questions, and grading rubrics, check out the College Board web site at *www.collegeboard.org/ap/statistics* .

Outliers, from page 29

Outlier...s

Allan Rossman

For this issue's column my mind has been wandering to thoughts of word games and other arcane pursuits.

STATS in License Plates

Bradley Efron, Stanford University professor and developer of the bootstrap method, told me of a game that he plays to help pass the time while driving. He "collects" sequential three-digit license plate numbers. He started by looking for "001" as the last three digits on a license plate, then looked for "002" and has continued for years. When I told him that I wanted to mention his hobby in this column, he replied that he had just spotted "817" that morning. Brad mentioned that one frustration of this hobby is that it often seems like the numbers just past the one you're looking for come along much more often than the target number. He decided to study this apparent phenomenon at one point, so he kept track of sightings of the number in question and the one immediately past it. He found that occurrences were indeed about 50/50 and attributed the frustrating feeling to the fact that the next several numbers beyond the target, not just the one immediately past, contribute to the frustration. [**Assignment 1**: Get started on your own collection of license plate numbers! **Assignment 2**: If you know of similar hobbies about collecting numbers, please let me know.]

When I first heard this story, I wondered whether the proliferation of "vanity" license plates in recent years has made it harder for Efron to collect these numbers. Then I began thinking about whether I have even seen a vanity plate whose message was related to statistics. I don't believe that I have, but this led me to imagine what a statistics vanity plate might say. On the unimaginative end of the spectrum are STATGUY and STATGAL, but a more creative devotee of John Tukey's work might choose 2KEYFAN. An advocate of maximum likelihood estimation might say MLE4ME, and a statistician who analyzes data with prior and posterior distributions might economically select BAYCN ("bay-see-en"). A more generic statistical option might be STM8R ("estimator"). [**Assignment 3**: Please let me know if you have seen any statistics-related vanity plates. **Assignment 4**: Create and let me know of better (how could they not be?) statistics-related vanity plates than my suggestions.]

STATS in Anagrams

Playing with words in this manner has made me think of anagrams for some reason. An anagram is a rearrangement of the letters in a word or name to form other words or names. For instance, some of the anagrams that can be formed by rearranging the letters in my name include "nasal normals" and "also snarl man". Table 1 includes twelve anagrams resulting from the term "statistical science." [**Assignment 5**: Pick out and let me know your top three favorites from this list.]

Table 2 contains anagrams for ten well-known statisticians. [**Assignment 6**: Determine the name corresponding to each anagram.] If you would like some hints as to the identities of these statisticians, I will provide clues for half of the ten. One is currently President of the ASA. Another recently became Editor of the *Journal of Statistics Education*. A third is a best-selling textbook author. A fourth "collects" license plate numbers. A fifth was more famous as a nurse than as a statistician. Answers appear at the end of the column.

These anagrams were found using the program at *www.anagramfun.com*. [**Assignment 7**: Use this program to find anagrams of your own name, and send me your favorite. **Assignment 8**: Find anagrams of the names of other famous statisticians or of common statistical terms. Please send me your favorites.]

STATS in Philosophy

Playing these word games has turned my thoughts to matters philosophical, where one can play games not only with words but with ideas. I once reviewed a philosophy book by John Earman with the wonderful title *Bayes or Bust?* and the less captivating subtitle *A Critical Examination of Bayesian Confirmation Theory*. I enjoyed reading about applications of the Bayesian paradigm to

Table 1: Anagrams of "Statistical Science"

a stint ecclesiastic	assist acetic client	cats nastiest icicle	cite cacti saltiness
ecstatic lice stains	italics accents ties	tactile ascetic sins	classic taste incite
satanic sect elicits	elitist antic access	stale tactic iciness	insect tail ascetics

Table 2: Anagrams of Well-Known Statisticians

(a) Chef Friar Crashed	(b) Sigma Lowliest	(c) He Toy Junk
(d) Nasal Porker	(e) Smooth Trash	(f) Redone Barfly
(g) Obey Math Ass	(h) Frantic Slogan	(i) Often Legal Enriching
(j) Mood Varied		

philosophical problems. One principle that the author mentioned as a reasonable expectation for a theory of confirmation is that a confirming instance of a hypothesis should increase, or at least not decrease, the probability that the hypothesis is true. For example, if one hypothesizes that all of the taxicabs in a city are yellow, then observing a yellow taxicab (a confirming instance of the hypothesis) should not cause the probability of the hypothesis to decrease. While this principle seems eminently reasonable and desirable, it is easy to construct examples where it fails. I offer one based on the classic "matching problem" of probability theory.

Suppose that three executives bump into each other and drop their cell phones. Completely confused, each picks up a phone at random, so all six assignments of phones to people are equally likely. Consider the hypothesis: "Nobody gets the correct phone." [**Assignment 9**: Before reading further, determine the (prior) probability of this hypothesis.] I shall denote the six possible outcomes as 123, 132, 213, 231, 312, and 321, where *ijk* means that the first executive picks up phone *i*, the second phone *j*, and the third phone *k*. Only two of these six possibilities leave every person with the wrong phone, so the prior probability of the hypothesis is 1/3. Now suppose that the first executive tries her phone and finds that it really belongs to the second executive. [**Assignment 10**: Before reading further, determine the updated probability of the hypothesis in light of this evidence.] Since there are now only two equally likely possibilities left in the sample space (213 and 231), the probability is now 1/2 that the hypothesis is correct. Thus, this confirming instance (the first executive not getting the correct phone) has indeed increased the probability of the hypothesis from 1/3 to 1/2. Now suppose that the second executive tries his phone and finds that it belongs to the first executive. [**Assignment 11**: Before reading further, determine the updated probability of the hypothesis in light of this evidence.] We now know 213 to be the outcome of this mishap, so executive 3 has the correct phone and the hypothesis is false. Thus, even though we have had two confirming instances of the hypothesis (two people were discovered to have the wrong phone), the probability of the hypothesis has decreased to zero.

STATS in Stuffed Animals

Now that my brain is really tired from not only playing word games but trying to think philosophically for a few minutes, I will conclude this column with a childlike, but not a childish, example. Tom Short of Villanova University (whose anagram using "Thomas" appears in Table 2) has developed an activity that he presents to elementary school children to introduce them to some important ideas of statistics. Tom brings his collection of teddy bears, numbering in the dozens, to the children's classroom. He tells them that he is starting a company to make and sell teddy bears and that he needs to pick one bear for his advertisement. He tells them that he wants this bear to be as representative as possible of all the bears in the collection, and he asks for suggestions. Children typically call out their favorite bear at this point, and that bear has something special about it. Tom asks if the favorite bear is really typical, and the children realize that its special-ness actually makes it not typical. At this point Tom asks the students to be a bit more systematic and to think of bears that can be ruled out as candidates for the typical bear. Children respond by suggesting many bears that are

See Outliers, page 27