Stats

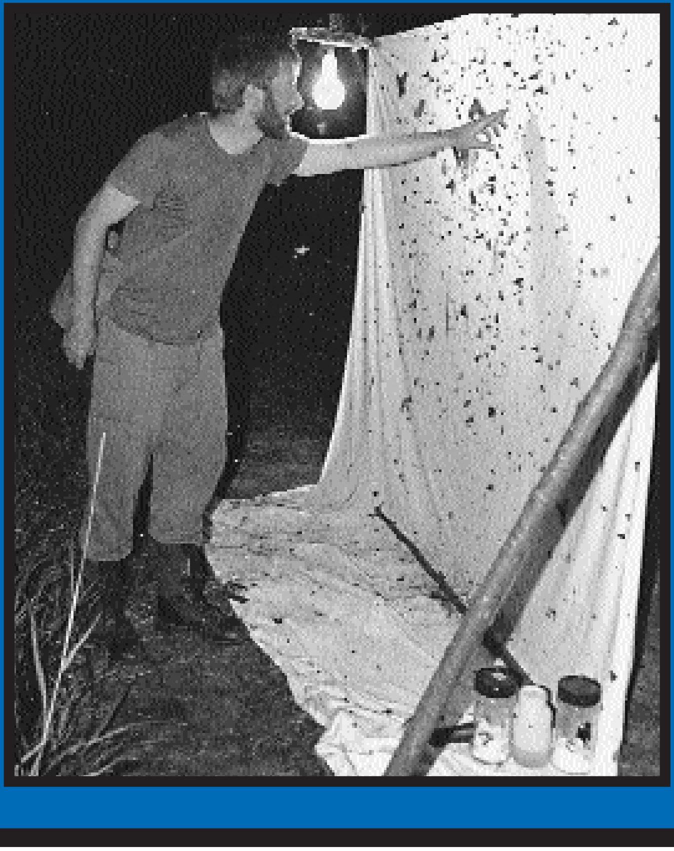
The Magazine For Students Of Statistics • Winter 1998 •

Bugs, Hollow Curves and Species-diversity Indexes

The Role of Statistics in Scientific Endeavor

An Evolutionary Model for the Optimization of a Maze

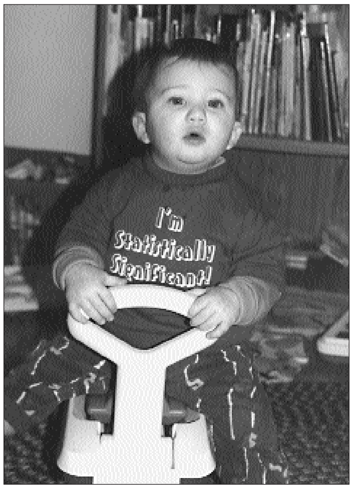
Life & Hard Times of a Statistician



We Have New Souvenirs

If we missed you in Anaheim, be sure to check out our new T-shirt

One year old Joey, son of Zoe and Chris Mullhaupt, rides off into the sunset. Joey is the grandson of Betty and Ed Spar. Ed is COPAFS Executive Director and a long time ASA member.



ADULT SHIRTS (Circle Size & color)			**Quantity**	**Total**
"Variability is the Spice of Life"				
Royal Moss	M L XL	$12.00	_______	_____
	XXL	$14.00	_______	_____
"Approximately Normal"				
Cumin (XL & XXL only)	M L XL	$12.00	_______	_____
Pebble (M & L only) Stonewashed Green	XXL	$14.00	_______	_____
"The Evolution of Statistics"				
Kelly Green Ash	M L XL	$12.00	_______	_____
	XXL	$14.00	_______	_____
"I□m Statistically Significant"				
Jade (L only)	M L XL	$12.00	_______	_____
Natural (L & XL only)	XXL	$14.00	_______	_____
"Uncertainty... Something You Can Always Count On"				
Jade Sunset Red Plum	M L XL	$12.00	_______	_____
	XXL	$14.00	_______	_____
"In God We Trust... All Others Bring Data"				
Teal Black	M L XL	$12.00	_______	_____
	XXL	$14.00	_______	_____
YOUTH/CHILD T-SHIRTS				
"I□m Statistically Significant"				
(Youth) Ash Navy	S M L XL	$8.00	_______	_____
(Child) Royal Red	2 4 5/6		$8.00 _	_____
"Dependent Variable"				
(Youth) Slate Black	S M L XL	$8.00	_______	_____
(Child) Pink Blue Yellow	2 4 5/6		$8.00 _	_____
Ball Point Pens	$3.00 each 4 for $10.00	6 for $15.00	_______	_____
VIDEOS				
"Statistical Science: 150 Years of Progress"			$20.00	_____

POSTAGE CHART

U.S. Orders	**Postage**
Up to $12.00	$4.00
$12.01–20.00	$5.00
$20.01–30.00	$6.00
$30.01–40.00	$7.00
$40.01–50.00	$8.00
$50.01–75.00	$10.00
$75.01–100.00	$11.00
Over $100.00	$13.00
Canada	$10.00
Outside North America	
$20.00	

Subtotal __________

VA residents add 4.5% __________

Postage (See Postal Chart) __________

Total $_________

METHOD OF PAYMENT

❑ I enclose a check or money order for $__________, made payable to ASA

Charge ❑ VISA ❑ MasterCard ❑ Amer. Exp. Card Number ______________________

Card Number ____

Expiration Date ___________ **Name as it appears on card** ____________________

Telephone Number ______________________________________

E-mail ______________________________________ **Fax** ____________

Ship to: Name ______________________________________

Address (No P.O. Boxes)__________________________________

City ____

State/ZIP or Postal Code/Country ____________________________

Orders must be prepaid. Send order form and payment to:

ASA Souvenirs
American Statistical Association
1429 Duke Street
Alexandria, VA 22314-3415 USA

Please make remittance payable in U.S. currency drawn on a U.S.

Exchanges cheerfully made—
but please be sure to include postage!

Winter 1998
Number 21

The Magazine For Students Of Statistics

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: http://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, Christine McLaren. 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.

Printed in U.S.A.

Cover photo flGaden Robinson finds a particularly rare moth that has settled on the collecting sheet (page 8).

Features

Departments

Editor's Column

STATISTICS and SCIENCE

Christine McLaren

Dear *STATS* Readers:

First, I am very happy to announce the winners of the *STATS* **Face the Facts Contest**. We asked readers to identify the cover photo that appeared on issue 20 of *STATS* (Fall, 1997). The following entrants correctly guessed that the photo was of **Cleveland, Ohio**: Chris Andrews, University of California, Berkeley, CA; Rick Cornez, University of Redlands, Redlands, CA; Yan Liu, The George Washington University, Washington, DC; Lisa Rybicki, Cleveland Clinic Foundation, Cleveland, OH; Tom Short, Villanova University, Villanova, PA; and Steve Wang, The University of Chicago, Chicago, IL. Additional entrants guessed that the photo was of Chicago, Pittsburgh, and Cincinnati: Andrea Hofer, University of Vienna, Vienna, Austria; Todd Schwartz, University of North Carolina at Chapel Hill, Chapel Hill, NC; and Don Reed, Georgia State University, Atlanta, GA. Each entrant received a prize of a one-time use panoramic camera! We thank Linda Quinn for submitting the photo. It also appears on the Cleveland Chapter Web Site with the URL, http://www.bio.ri.ccf.org/docs/ASA/cleveasa.html.

The three feature articles in this issue have a common theme: the application of statistics to scientific problems. In "The Role of Statistics in Scientific Endeavor," Graham McLaren (no relation) and Maria Carrasco-Aquino discuss statistics and its role in scientific research within the framework of the scientific method. In this expository article, they explain the logical steps that enable researchers to translate a scientific hypothesis to a statistical hypothesis, the role of statistical inference and the cyclic nature of scientific investigations.

In pursuit of the answers to his scientific questions, Gaden Robinson, an entomologist, left his desk at The Natural History Museum in London. He traveled to the valley of the Rampayoh river in the rainforest of Brunei in northern Borneo. The scientific purpose of his trip was to determine the diversity of moth species living in the rainforest. In his article, "Bugs, Hollow Curves and Species-diversity Indexes," Robinson gives us fascinating details of his scientific research, such as setting up a base-camp with hammocks, work- and cooking areas, and vapour lamps. We can imagine him listening to the hum of noisy cicadas while trapping and classifying moth specimens. Returning to London to analyze the data, he now explains the role of the logarithmic series in determining the distribution of number of individual moths among species.

In our third feature article, Curtis Stratman describes the independent research project that he developed during his high school science research class at Jefferson City High School in Jefferson City, Missouri. In his project, "An Evolutionary Model for the Optimization of a Maze," Stratman drew on ideas from his high school classes in mathematics, biology, and psychology. Stratman combined machine learning techniques with the genetic algorithm, a method that imitates the process of DNA recombination in sexual reproduction, to devise a technique for finding the optimal route through a maze. This article is the first one published in *STATS* that has been written by a high school student.

In this issue, we hear from Karen Kafadar, Professor in the Department of Mathematics at the University of Colorado-Denver. Kafadar describes statistical jobs that she has held in government, industry, and academe. In keeping with the theme of the issue, Kafadar tells about her work on scientific studies in the Statistical Engineering Division at the National Bureau of Standards (now the National Institute of Standards and Technology), Hewlett Packard, the National Cancer Institute, and the University of Colorado-Denver. Clearly a talented statistician, Kafadar shows her ability to apply statistics to a wide range of scientific problems.

We are very privileged to hear of the Life and Hard Times of J. Stuart Hunter. He describes his educational pathway leading to an appointment at Princeton. As an engineer and a statistician, Hunter says, "In many ways we are acolytes to the scientific method, lighting the way from conjecture to data acquisition, and through the pathways of analysis helping in the creation of new ideas and repetition of the learning cycle." He advises, "I can definitely declare that anyone with an interest in a science, or medicine, or engineering, or even politics who has an interest in handling data will find a career in statistics rewarding."

Yours in Search of Scientific Truth,

Christine McLaren

The Role of Statistics in Scientific Endeavor

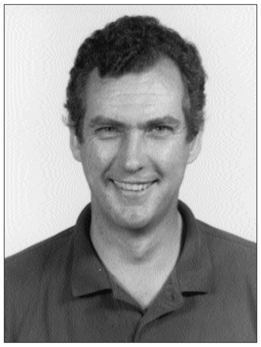
■ 1. Introduction

Scientific endeavor is a continuous cycle of clarification of the unknown through the development of theories and the generation of knowledge through the verification or modification of those theories. The process of generating, testing, modifying and verifying scientific theories is called the scientific method; and statistics, which is defined as the science and art of collecting, organizing, analyzing and interpreting data, is an indispensable instrument in this process. As such, statistics plays a vital role in scientific endeavor: it makes a major contribution to the efficient design of scientific investigations and to the applicability and validity of conclusions which are generated.

In this paper, statistics and its role in scientific research is discussed within the framework of the scientific method. Emphasis is placed on the logic and philosophy behind statistical techniques so that researchers can better understand their application and so that statisticians can identify their role in science.

■ 2. The Scientific Method

Although the techniques of investigation vary considerably from one scientific discipline to another, the scientific method represents a philosophy which is common to all. It is defined by the cycle of clarification of the unknown and generation of new knowledge through the continual search for improvement of tentative theories used to explain physical phenomena.



C. Graham McLaren

Maria Carrasco-Aquino

Graham McLaren is head of the Biometrics Unit at the International Rice Research Institute (IRRI). He first served as head of Biometrics in Zimbabwe, then as part of a British aid contribution to a World Bank project in Cameroon, and he joined the IRRI in 1993. He is a member of the ASA and the International Biometrics Society, where he served as council member from 1990-1993. McLaren's E-mail address is G.McLaren@cgnet.com.

Maria Carrasco-Aquino graduated from the University of The Philippines (UP) in 1985. She taught statistics as an Assistant Professor at UP and was a senior statistician with Dole Philippines. As a Senior Research Assistant at the International Rice Research Institute, she was responsible for training rice research scientists in applied statistics. Since 1995 she has moved into quality assurance and statistical process control with Texas Instruments and Cypress Semiconductor Philippines. Carrasco-Aquino's E-mail address is cca@cypress.com.

Example 1: An application of the scientific method.

Let us recall the controlled breeding studies of Gregor Mendel in 1865. The details of his research endeavor to discover the basic rules of heredity provide a good example of the scientific method.

Scientific Knowledge

Mendel began with the scientific knowledge that a cross between round-seeded peas and wrinkled-seeded ones produced round-seeded peas which no longer bred true among themselves.

Scientific Hypothesis

Mendel continued by developing the following theory to explain this observation. When two peas, one with round and one with wrinkled seeds, are crossed, they each donate to their offspring one particle determining the trait. One of the particles, say round, dominates the other so that in an individual with one of each type, the seeds will be round.

Statistical Hypothesis

The next stage was to consider the consequences of the theory in a practical setting. When the round-wrinkled crosses are bred among themselves about three quarters of their offspring should be round and one quarter wrinkled.

Experiment

This led to the design of an effective experiment. Breed the crosses and count the round and wrinkled offspring.

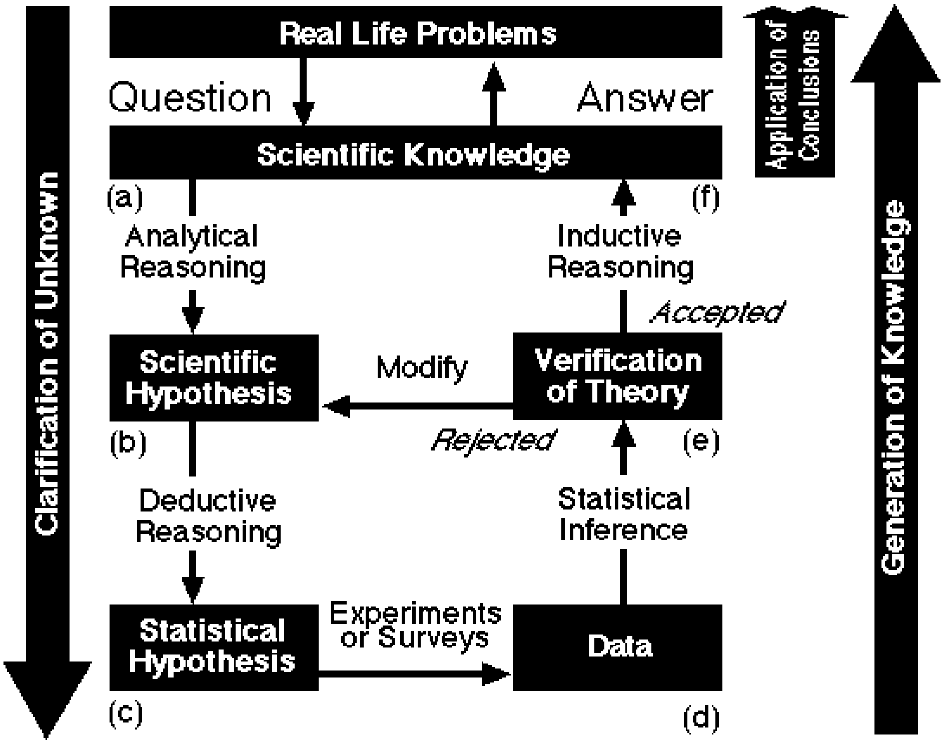


Figure 1. The Scientific Method

Data and Analysis

The data are the observed frequencies of the round and wrinkled offspring. Analysis was done to check whether or not they agree with the statistical hypothesis.

New Scientific Knowledge

By inference to other traits and other organisms, the theory of Mendelian inheritance began to be developed. Later observations showed that the simple diallel law was not always adequate and so started the cycle of scientific endeavor to refine the laws of inheritance.

This example illustrates the principal steps in the scientific method which applies very generally to scientific endeavor. The researcher, faced with real life problems or questions, wishes to find solutions or answers through the scientific method. As shown in Figure 1, the method involves the following logical processes—analytical reasoning, deductive reasoning, statistical inference, and inductive reasoning.

In the first step of the scientific method, (a) → (b), the question or problem is reduced to a scientific hypothesis. This is a theory about the state of nature giving rise to the phenomena being studied. This is the theory which the researcher wants to investigate.

Example 2: Defining a scientific hypothesis.

Suppose we have the following observation in our field: Rice leaves are yellowing in young, growing plants. **(Real life problem)** So we ask: Why? What could be the cause?

We seek answers to this question by first examining **what we know** about the problem of yellowing leaves. Previous studies on this issue can give some indication as to the probable cause(s) and consequences. Experience and literature indicate that yellowing of leaves in young plants may result from nitrogen (N) deficient soil and can cause considerable yield reduction if not corrected. **(Scientific Knowledge)** Hence, we contend that the yellowing of the leaves is due to soil N deficiency. **(Scientific Hypothesis)**

In this step we closely examine the problem and its manifestation in the context of current knowledge and experience. The purpose of this analytical reasoning is to formulate a tentative hypothesis that would explain the facts. This is the most important but least formalized step in the whole process.

Here are some suggestions for developing a satisfactory scientific hypothesis. It should:

- be formed in a way that is closely related to the problem to be solved,
- suggest or provide an answer to the particular problem which generated the inquiry,
- provide direction for the research,
- be stated as simply as possible, and
- be capable of verification or rejection.

In the path (b) → (c), the researcher tries to determine what would happen under a specific set of conditions, if the scientific hypothesis is true. This is basically a clarification and often a simplification of the question or problem. We call these consequences of the scientific hypothesis, the **statistical hypothesis**.

It is the statistical hypothesis that is verifiable and forms the basis for setting up the appropriate experiment or survey design. Analysis of the resulting data using statistical inference leads to acceptance or rejection of the statistical hypothesis. The reasoning at this stage must be clear and critical if reliable conclusions are to be drawn from the experiment. The type of reasoning involved in this step of the cycle, which leads from the general to the particular, is referred to as deductive reasoning.

Example 3: Translating a scientific hypothesis to a statistical hypothesis.

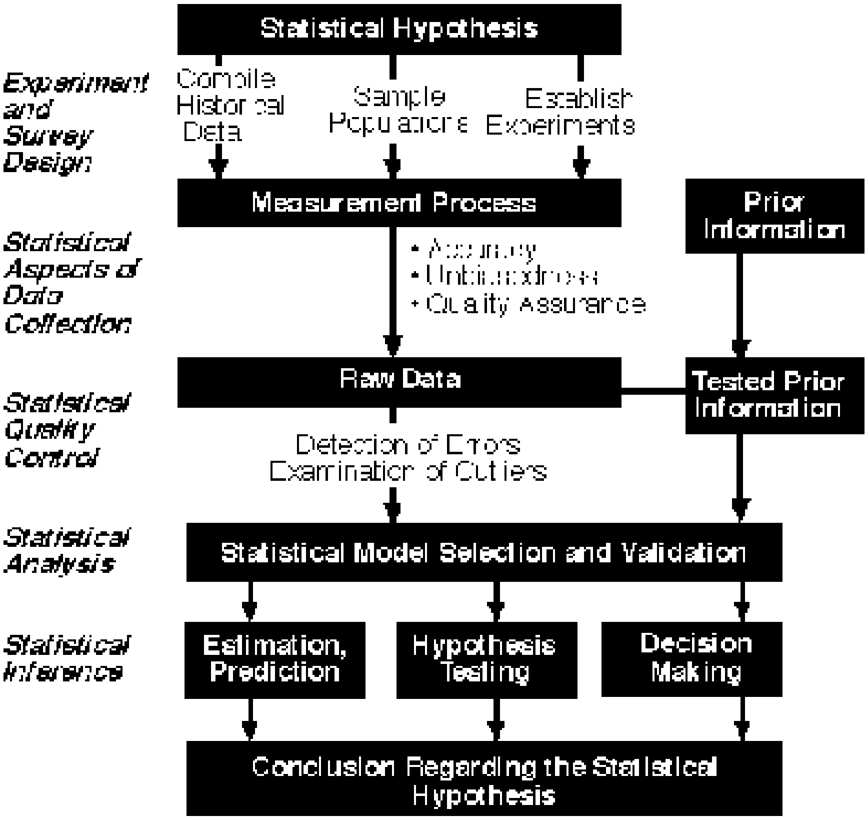
If we deduce the observable consequences of the scientific hypothesis following our example 2, the statistical hypothesis could be stated as follows:

If the scientific hypothesis regarding nitrogen deficiency is true then measured levels of N in the soil should be below critical levels indicated in the literature. **(Statistical Hypothesis)**

Let us now move on to the next steps in Figure 1. The path from (c) → (d) → (e) involves **the experimental process**, i.e., the designing and implementing of experiments or surveys to examine the predicted consequences. It is at this stage that statistics has a major role to play in the process by providing the logical and probabilistic framework for evaluating the tentative hypothesis. On the basis of the collected data, the researcher decides on the acceptability of the theory using statistical analysis and inference, (d) → (e).

The paths (e) → (b) and (e) → (f) lead to the **generation of a new scientific knowledge**. When the hypothesis and data fail to agree, the researcher is led to revise the former, path (e) → (b). However, when the data and hypothesis are in agreement, the generality of the new scientific knowledge is explored in order to arrive at some general principles that will apply to a wide class of situations represented by the experimental conditions, (e) → (f). This method of reasoning, from the specific to the general, is referred to as inductive reasoning. It is from this new generally acceptable scientific knowledge that we hope to find answers to the real life problems that motivated the cycle in the first place.

Figure 2. Statistics in the Experimental Process.



■ 3. Role of Statistics in the Scientific Method

Statistics has a major role to play in all stages of the scientific method. This is because it is involved with the definition and evaluation of hypotheses through the collection and analysis of data. In the paths (a) → (b) and (e) → (b) of analytical and inductive reasoning, the methods of descriptive statistics have their role to play. They provide powerful tools for suggesting questions to ask and formulating hypotheses. This is particularly useful in the study of large data sets, especially those routinely collected without specific research purposes in mind. Such data should also be examined for indications as to the hypothesis or theoretical model underlying the process which produced the data. Data examination may include exploratory techniques such as tabulations, summary descriptions, graphical analysis, and cluster analysis.

Statisticians play an invaluable role in this exploratory stage by working closely with researchers. A basic understanding of the subject area and excellent communication skills are important for the success of this collaboration.

The experimental process, paths (c) → (d) → (e), of the scientific method is intimately involved with many areas of statistics. A description of the statistical methods and thought processes in this part of the scientific method is depicted in Figure 2 and will now be discussed.

After clearly formulating the **statistical hypothesis**, relevant and valid data are accumulated from historical records, sample surveys or experiments in order to test the given hypotheses and provide indications for possible alternatives. Statistics provides the researcher with an array of methodologies to help in the design of an efficient and cost-effective data collection scheme which also ensures the **accuracy**, **unbiasedness**, and **quality** of the data.

In the area of **measurement process** the statistician's technical skills are needed. Close collaboration with researchers in the subject area and communication of statistical principles are also crucial.

The principles of **quality control** may prove to be of valuable assistance during and after data collection. Data ought to be routinely checked for the presence of errors, biases and outliers. The relevance of the data to the hypotheses under study also needs to be continually checked.

Statistics plays a crucial role in experimentation. Even in the best planned experiments, we cannot control all the factors that affect our observations and we can rarely make measurements without some noise or error from the measurement process. Hence, we have to make inferences based on imprecise sample data. To be of practical use, these uncertain inferences must be accompanied by probability statements expressing the degree of confidence the researcher has in the conclusions. To make certain that such probability statements will be possible, the experiments should be designed in accordance with the principles of statistical experimental design. These principles, together with the statistical hypothesis under study, dictate a statistical model relating the data to the statistical hypothesis through probability theory.

In other words, data have no meaning in themselves; they are meaningful only in relation to a statistical model of the phenomenon being studied. The interpretation of a set of data would be different, depending on what model was thought appropriate. In practice, some basic knowledge of the phenomenon under study is usually available to allow the researcher to specify a plausible statistical model.

Example 4: The statistical model in Mendel's breeding studies.

Let us consider Mendel's work as discussed in example 1. If plants with round seeds are crossed with ones having wrinkled seeds and the resulting hybrids are inter-mated to give what are known as the F_2 progeny, then the statistical hypothesis is that these will segregate in the ratio 1:3, that is, one quarter will produce rounds seeds and three quarters wrinkled. The experimental design was to randomly observe a number of F_2 plants and count the two classes. The statistical model is that any randomly observed F_2 plant has some probability, p, of producing round seeds so that the number of plants producing round seeds in our sample follows a binomial distribution. The statistical hypothesis is then expressed in terms of this model by taking $p = 1/4$.

The stage is now set for answering the researchers' queries. The process of drawing inferences or conclusions from the observed data will then make use of methods of statistical inference. These include methods for estimation, prediction, hypothesis testing and decision-making based on observed data and a specified statistical model. It is the application of probability theory in these methods that leads to confidence in precision of estimates or predictions and consistency in rejection or acceptance of hypotheses.

For example, if we assume our statistical hypothesis to be true, then our statistical model, with inferential methods, will allow evaluation of the probability of observing data which is less supportive of the statistical hypothesis than the probability of the data actually observed. If this probability is very small, we will conclude that either the statistical hypothesis, or the statistical model is wrong. If we find no evidence to indicate that the model is wrong, we reject the statistical hypothesis.

Example 5: The statistical analysis of Mendel's experiment.

If Mendel observed 100 F_2 plants in his experiment, the binomial model and the statistical hypothesis ($p = 1/4$) indicate that we expect about 25 to produce round seeds. If he observed 20 plants with round seeds, this deviates by five plants from what we expect. We need to ask whether this deviation is plausible under the statistical hypothesis. To do this we use the statistical model to calculate the probability of observing data that are less supportive of the statistical hypothesis. Clearly, results with fewer than 20 plants with round seeds are less supportive of the hypothesis $p = 1/4$ than the data actually observed; they tend to indicate that $p < 1/4$. Similarly, results which are more than 30 are less supportive of the statistical hypothesis, since they also deviate from our expectation by more than 5 plants and indicate that $p > 1/4$. Hence, we use the binomial probability distribution with N = 100 and p = 1/4 to compute the probability of observing less than 20 or more than 30 plants with round seeds. This probability is about 0.25 so the data are not at all unexpected and we can accept the hypothesis p = 1/4.

If we reject the statistical hypothesis, then the scientific hypothesis, which is supposed to hold more generally than in the specific experiment, is also unlikely to be true and must be re-considered.

■ 4. The Cyclic Nature of Scientific Investigation

An important aspect of the scientific method is its cyclic nature. Knowledge and understanding are accumulated through a never-ending cycle of scientific endeavor. For example, the theory of Mendelian inheritance soon required refinement to handle the inheritance of linked characters, and is still being refined today to accommodate more and more complex situations. We need to anticipate these cycles in our research planning, never viewing an experiment in isolation and always measuring or observing phenomena that can point the way to the next cycle. Sample problems can become very complicated during the inductive reasoning stage when we try to generalize from specific results.

Example 6: Refinement of knowledge in leaf yellowing.

Let us consider again examples 2 and 3 where we are interested in correcting the yellowing of the rice leaves. Suppose that the data we obtained from the soil samples indicated that the field is in fact N deficient. We conclude that this is the reason for the yellowing of the leaves. Still we may ask:

What is the degree of deficiency or how much N are we going to apply ?

To answer this question, we could try to determine the nitrogen rate to apply in our field by going back to rice literature which indicates that nitrogen rate is soil-and season-specific **(What we know)**. If we have no recommendation for the prevailing environmental conditions, we want to zero in on the N rate to apply to the field that will give us optimum yield. The literature indicates that response will be a convex curve with an optimum between 50 and 150 kg N/ha **(Scientific Hypothesis)**.

The exact form of our scientific hypothesis has important consequences for the design of our experiment. If the simple estimation of a convex curve and its optimum is sufficient, we only need three levels of N, two would not be sufficient. But in the same spirit of contingency planning that applied to the leaf N, we should have a fourth level to check the adequacy of our convex curve theory.

In a field trial **(Experiment)**, we may then choose N rates 0, 50, 100, and 150 kg N/ha as the **treatments** and we should observe the response of grain yield to N **(Statistical Hypothesis)**. The experimental **design** is to randomly allocate the treatments to the plots in block considering the direction of the field's fertility gradient, say with five replications (Randomized Complete Block Design).

The **statistical model** is that mean yield response to N follows a quadratic form with normally distributed errors. This model allows us to test the hypothesis about the convex response and then estimate the optimum N rate.

Data on grain yield will be obtained after harvesting and the yield of the four treatments will be compared. Suppose that the observed quadratic model indicates that the highest grain yield would come from 112 kg N/ha, then we postulate **(Inference)** that, with the given environmental conditions, 112 kg N/ha is likely to give optimal grain yield, and that this rate should correct the yellowing of the leaves.

■ 5. Conclusions

Scientific endeavor has been the principal method of generation of knowledge since the Renaissance and is likely to continue as such for centuries to come. We have seen how this is based on a scientific method or philosophy, and that statistics is central to this method.

Statisticians, therefore, have a responsibility of custodianship over this philosophy. An understanding of it, and of the role of Statistics in it, is crucial to knowing what we have to do, whether we are theoretical statisticians developing new techniques or applied statisticians using them.

The most important function of an applied statistician is in communicating to scientists this philosophy, and the role of statistics in it. To do this, we need to understand scientific subject areas and bring a statistical perspective to each step in the scientific process through analytical reasoning, deduction, statistical inference and induction.

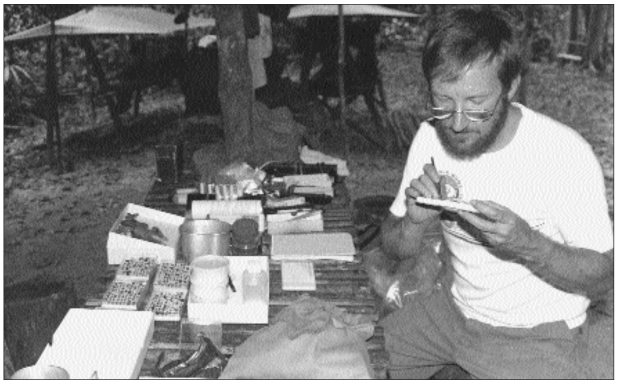
Bugs, Hollow Curves and Species-diversity Indexes

1. Introduction

The valley of the Rampayoh river in Brunei is a military training area. It is also one of the few remaining undisturbed tracts of lowland rainforest in northern Borneo. The helicopters that brought troops and four entomologists to the valley are long gone. The eight Gurkhas have established a comfortable base-camp of hammocks, work- and cooking areas. The insect scientists have been busy elsewhere and the cleared patch of ground that served as a helipad, surrounded by 150-foot high trees, is now dominated by a framework of poles supporting what appears to be a cinema screen but which is in fact a large white bed-sheet. As dusk falls the noise of crickets, cicadas and frogs becomes deafening. In this cacophony a regular beat develops as the generator warms up. The faint blue flicker in front of the sheet becomes a brilliant glow in just two minutes as the mercury vapour lamp reaches full power. Dazzled insects spiral in to the light and settle on the sheet. Chaos develops: the air is full of insects, hornets run up and down the sheet, biologists duck and weave, avoiding hornets, to search out their quarry, gently enticing tiny moths into glass tubes. Beard-owners curse more than the clean-shaven.

The following morning the chaos is reduced to order. The catch is humanely killed and pinned. Wings are spread to expose details of colour and structure to make identification as easy as possible. Specimens are packed into lightweight boxes ready for the journey back to London. Our prime purpose is taxonomic—that is, to provide specimens for use in research into the classification and relationships of these moths. Put another way, our insects on pins will eventually tell us what species live in this rainforest, how they are related, and what evolutionary route they took to get where they are today.

Gaden S. Robinson, DSc, entomologist at The Natural History Museum, London, England, is a specialist in the biology and taxonomy of small moths. His recent publications include A Field Guide to the Smaller Moths of South-East Asia, *a monograph on the clothes-moth family in Australia, and several papers on tropical moth diversity. He is currently developing a worldwide database of caterpillar hostplants.*



In camp—segregating species from samples collected the previous night. The "laboratory bench," constructed by the Gurkhas, is made from short poles. Awnings in the background shelter individual hammocks—the sleeping quarters.

A glance at the boxes of prepared specimens reveals a feature that is especially remarkable to a biologist from a temperate region—practically no two specimens are the same. There are hundreds of species, for this is one of the biologically most diverse habitats on earth—a haven of species-richness.

Some of our samples are taken not with taxonomy as the prime objective but in an attempt to measure that richness. They are samples that are as random as we can make them and all specimens are collected. These will provide the context for our work and give us some idea of the potential number of species that we would have to deal with in a complete inventory of the fauna. Perhaps more importantly, measurements of species-richness will allow us to compare and contrast the faunas of different localities and measure the effect of habitat disturbance on those faunas. It will allow us also to examine the phenomenology of species-richness and the way that it apparently varies with time and sampling technique. Back in the laboratory each moth in our samples will be labelled and the individuals sorted into species.

Months later, this sorting shows us what we would perhaps expect to see: many species are represented by just a single individual, fewer by two individuals, fewer again by three and so on. There are many "rare" species and comparatively few common ones. The ranked series of the distribution of individuals among species with increasing frequency is a hollow curve (Figure 1; Table 1).

Table 1. Distribution of individuals among species in a pooled ten-day light-trapped sample of small moths (Microlepidoptera) from lowland rainforest in Brunei. Total individuals (N) = 1230; total species (S) = 571. The hypothetical log-series distribution (α = 414) is shown for comparison.

Species with:	Observed	Expected (log series)
1 individual	320	309
2 individuals	116	116
3 individuals	49	58
4 individuals	33	32
5 individuals	15	20
6 individuals	19	12
7 individuals	6	8
8 individuals	4	5
9 individuals	2	3
10 individuals	0	3
Residual Species (>10 individuals) at 11, 13, 15, 16, 17, 19, 27	7	6

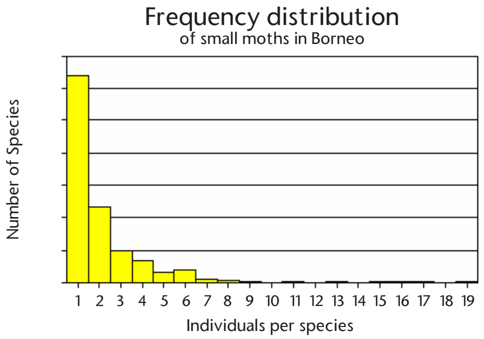


Figure 1. Distribution of individuals among species (data from sample in Table 1) - the classic “hollow

■ 2. History

The distribution of units among groups in the natural world has been a topic for investigation for more than 70 years. Hollow curve distributions apply universally in nature. They have been demonstrated in populations of plants, insects, birds, mammals and micro-organisms. But they do not apply just to the distribution of individuals among species, but also to higher ranking levels in biological classification—the distribution of species among genera and the distribution of genera among families. Indeed, hollow curves are universal—the frequency distribution of words in this article is a hollow curve (Williams, 1970).

Willis (1922) examined the frequency distribution of species among genera and recognised that here was mathematical regularity. He proposed the hyperbolic series as the best fit to the data he examined. But the fit was poor.

In the late 1930s, C.B. Williams embarked on an ambitious programme of nightly light-trapping of insects, notably moths, at Rothamsted Experimental Station in southern Britain. His design of trap, essentially a light above a funnel that led into a large jar charged with killing-agent, took samples that were cumulatively large—13,000-40,000 specimens per year. His trapping program continues today, using the same trap design in the same location, and provides now by far the world’s longest and largest biodiversity time-series data set. Williams noted (1937) that many features of his data were best analysed by treating insect numbers as a series of geometric relationships, but description and modelling of the hollow curve eluded him for several more years. In 1942, A.S. Corbet reopened the problem with the examination of the relationship between individuals and species in a collection (Corbet’s) of Malayan butterflies. Corbet’s collection was of course badly biased—butterfly collectors take a series of the common species then give up and concentrate on hunting rarities—but Corbet’s paper (1942) aroused Williams’ interest and he in turn was able to enlist the help of Sir Ronald Fisher in looking at not only Corbet’s data but also the first 4 years’ light-trapping results from Rothamsted—16,000 moths (individuals) in 240 groups (species). In a joint paper, Fisher, Corbet and Williams (1943) suggested that the observed frequency distribution could be fitted by a logarithmic series rather than a hyperbolic series. Diversity was suddenly an entity quantifiable using the parameter of the logarithmic series distribution.

■ 3. Diversity and its quantification

The diversity referred to here is now termed “alpha-diversity.” It is a measure of the species-richness of a particular group of organisms pertaining in a particular habitat or locality at a particular time. There are two ways to measure it. One is to enumerate all the species present, while the second is to measure the diversity of a sample or series of samples and extrapolate from these to provide meaningful comparisons and/or repeatable observations and/or testable hypotheses. The practical implications of the first alternative are staggering: it has taken an army of naturalists over 200 years to enumerate fully Britain’s Lepidoptera (moth and butterfly) fauna of 2500 species. In diverse tropical ecosystems full species inventories are clearly impractical except possibly for some vertebrate groups. Extrapolation from samples, despite the bias inherent in any sampling technique for whole organisms, is realistic.

At its simplest, for a sample containing a given number of individuals (say, moths), the diversity will be higher the greater the number of groups (species) in the sample. But we cannot take moth-samples at light that are conveniently all the same size. In order to be able to understand, compare and manipulate samples we need a measure of the species-richness in each that is independent of sample size - an index of diversity. The parameter of the logarithmic series is just this—it is now known as Williams' alpha (α) or, by some authors, Fisher's alpha.

Diversity has now come to mean many things to many people and its measurement and importance are areas of considerable debate (e.g., Patil and Taille, 1982). In fact, the logarithmic series has withstood the test of time and the buffeting of critics remarkably well. Williams (1944, 1947, 1964) showed that it applied to an enormously wide range of samples of different organisms of different sizes and with different life-styles taken over different periods of time. Critical testing by Wolda (1981, 1983) and Taylor (1978) provided additional justification. But Williams' alpha is not the only index of diversity and the reader should consult Wolda's papers and Magurran (1988) for a balanced view. My personal justification for using the log series and alpha still is the robust claim of the non-mathematician that "if it ain't broke, don't fix it" (Robinson and Tuck, 1996).

■ 4. The logarithmic series

So what is the logarithmic series and what is alpha? Given the crucial elements N, the number of individuals, and S, the number of species in the sample, the applicable log series can be calculated. The log series has particular properties that make it especially useful as a descriptor of samples. Firstly, it is relatively simple to generate log series and manipulate them using a computer. Secondly, populations (i.e., mixed-species assemblages) of different sizes can be generated using statistical estimates derived from samples so that samples can be "magnified" or "rarefied" and hypotheses explored. Thirdly, the parameter of the series, alpha (α), relates N and S directly and is a sample-size independent diversity index. Any particular log series is defined by the parameter α and knowledge of N and S, or by any combination of any two. The three are related by the formula:

$$S = \alpha\,(\log_e(1+(N/\alpha))) \qquad (1)$$

Alpha is most easily computed, given N and S, by iteration and substitution in this formula. Alternatively the tables given by Hayek & Buzas (1997) may be used. In the case of our moths from Borneo, the number of species represented by 1, 2, 3 .... r individuals is given by the elements in the series:

$$\alpha\,(N/(N+\alpha)),\ \alpha((N/(N+\alpha))^2/2),\ \alpha((N/(N+a))^3/3),\ \text{etc.} \qquad (2)$$

Fourthly, random samples of α log series are themselves a log series with the same α as the parent assemblage (although, of course, N and S will be different). The converse of this last point is that any sample must be from a putative mixed-species assemblage that has the same diversity (α) as the sample.

The large sample variance of α is given by:

$$V = \alpha^3(((N+\alpha)^2\log_e((2N+\alpha)/(N+\alpha))-\alpha N)/((SN+S\alpha-N\alpha)^2)) \qquad (3)$$

Sample diversity with 95% confidence limits is expressed by $\alpha \pm 2\sqrt{V}$.

■ 5. Applying the logarithmic series

We have used α to compare species-richness of small moths (Microlepidoptera and Pyraloidea) in different habitat types using "snapshot" samples accumulated over periods of 4-14 nights. During our work in Borneo we compared moth diversity in three distinct and different forest types: lowland

Table 2. Diversity of smaller moths in three different forest types in Borneo

	N (number of individuals)	S (number of species)	α (diversity)
Brunei–lowland rainforest	1230	571	414 ± 39
Sabah–montane rainforest	513	268	226 ± 35
Brunei–mangrove forest	870	253	120 ± 13

Gurkha infantryman gets an introduction to jungle entomology.

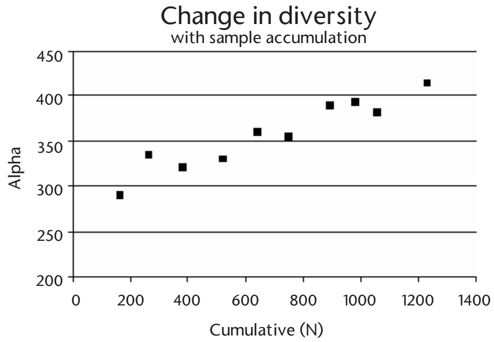
rainforest, montane rainforest and mangrove (swamp) forest (Table 2). Moth caterpillars, or most moth caterpillars, eat green leaves and are more or less host-specific. So it might be expected that floristically poor habitats, such as mangrove swamp (with perhaps 10-15 tree species in a hectare), have a less diverse moth fauna than a habitat rich in plant species such as lowland rainforest (with 230 tree species in one hectare). Indeed this is the case, and the moth fauna of montane rainforest, less rich in plant species than the lowland forest, is also less diverse. In fact the mangrove figure shown here is artificially high—our samples were from the landward edge of the swamp and are "contaminated" by species from nearby secondary forest. More recent samples from traps that were "buried" deep in the swamp show a much lower diversity of about 25.

Using the log series, it is simple to extrapolate and derive estimates of the total number of species that might be inventoried in, say, two years' sampling in a particular habitat. The observed value of α is substituted in the formula above along with a reasonable estimate for the number of individuals (specimens − N) that could be collected in the space of two years. For our lowland site in Borneo we estimated (Robinson and Tuck, 1993) that two years' trapping would yield about 160,000 specimens comprising at least 3750 species of moths. In practice this is an underestimate. If nightly trap-samples are pooled the diversity of the cumulative sample rises and over the first few days of trapping this rise is quite rapid (Table 3; Figure 2). This effect is caused by habitat heterogeneity. Variation in climate from night to night affects flight range and direction; metaphorically speaking our samples in the first few days of a trapping programme are drawn progressively from more and more tiles in the environmental mosaic. In the longer term there are seasonal changes that add to diversity.

Table 3. Change in diversity with sample accumulation (moths at light - lowland rainforest)

Days	N	S	α
1	164	130	290 ± 109
1-2	263	194	334 ± 90
1-3	383	252	320 ± 64
1-4	522	313	330 ± 53
1-5	641	368	360 ± 51
1-6	749	403	355 ± 45
1-7	894	464	389 ± 45
1-8	982	492	393 ± 43
1-9	1056	506	381 ± 39
1-10	1230	571	414 ± 39

Figure 2. Change in diversity with sample accumulation (moths at light, lowland rainforest)



The logarithmic series can also be used to compare samples in terms of their species-composition (i.e., the number of species shared between pairs of samples). Obviously, if α differs substantially between two samples they must be from different "parent" populations. But if α is similar, do these samples represent the same or different assemblages? The [expected] number of species shared between a pair of samples is simple to calculate—add the individuals in the two, then use the formula (1) to calculate the expected number of species (S_3) in a theoretically combined sample. The predicted number of shared species is given by $S_1 + S_2 - S_3$ where S_1 and S_2 are the [observed] numbers of species in the two samples. But the calculation of the variance in this operation eluded us. So we resorted to empiricism. The observed field situation can be mimicked by generating a large log series population within the computer with the same α as observed in the field; bootstrap random sampling of individuals from this artificial population is carried out to produce a series of replicate pairs of samples of the same size as the field samples. The number of shared species is counted for each pair, and mean and standard deviation is calculated.

We used this technique to compare samples taken 1 km apart in lowland rainforest in Borneo and also compared the species composition between mangrove forest, lowland rainforest and montane rainforest (Table 4). As might be expected, there is very close similarity between the two lowland samples but very little between other pairs. This confirmed what was obvious by inspection—that the insect faunas of the different major forest types in Borneo are radically different in terms of their species composition. Closer inspection of the data from the two lowland sites showed that the discrepancy in species-composition was accounted for by the smallest moths (the Microlepidoptera) that could be assumed to have limited dispersive powers. When

Table 4. Comparison of samples of Microlepidoptera from different forest types in terms of shared species ("master population" sampled has alpha equal to that of the two samples combined); the expected number is that expected were the two samples drawn from the same population

Shared Species: Sites Compared (forest type)	Observed	Expected	O/E
Lowland 1 × Lowland 2 (1 km apart)	151	175 ± 20	0.86
Lowland × Montane (180 km apart)	27	223 ± 13	0.12
Lowland × Mangrove (50 km apart)	37	299 ± 17	0.12
Mangrove × Montane	2	86 ± 6	0.02

Table 5. Frequency distribution of number of genera of hostplants eaten by 1183 genera of North American Lepidoptera (excluding butterflies). N = 9759; S = 1183. The hypothetical log-series distribution (α = 352.4) is shown for comparison.

Number of plant genera used as foodplants	Number of caterpillar genera (observed)	Expected (log series)
1	342	340
2	184	164
3	110	105
4	84	77
5	54	59
6	41	47
7	47	40
8	33	33
9	13	28
10	22	25
11	14	22
12	16	19
13	21	17
14	22	15
15	8	14
16	12	13
17	7	11
18	9	10
19	5	10
20	14	8
21-30	54	58
31-40	28	29
>40	43	39

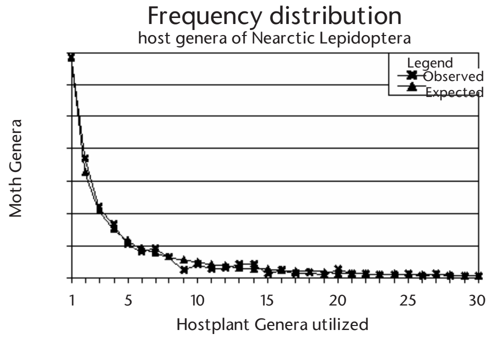


Figure 3. Frequency distribution of number of genera of hostplants used by 1183 host genera of North American Lepidoptera (excluding butterflies) with hypothetical logarithmic series shown for

we analysed the data for the comparatively strongly-flighted pyraloid moths that comprised about half the sample, the observed shared species (103) matched the expected (100 ± 13).

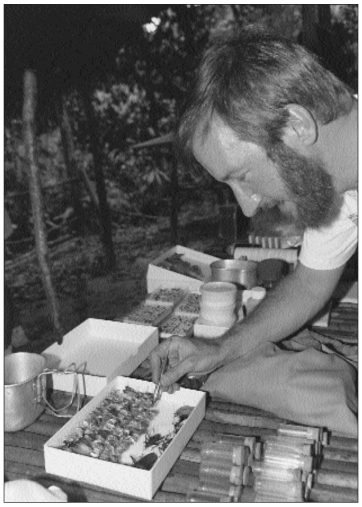
We suspect that, in the mosaic of a tropical forest, cumulative samples from two adjacent sites will show convergence in terms of their species-composition (the observed/expected ratio of shared species will increase) as time progresses. The rate of this convergence and the change in rate with the distance between the sampling-sites may be one way in which we can quantify the spatial accumulation of biotic diversity—"beta diversity." But this is a project for the future.

6. Spooky?

We have moved on from the rainforest: our latest project is very different—compiling a worldwide database of caterpillar hostplants—what eats what. As this project progressed one old, familiar feature began to emerge. Many caterpillar species eat just one genus of hostplant, rather fewer eat two, fewer eat three, and so on. The most comprehensive data available to us at present is for the caterpillars of North American moths. The frequency distribution of host specificity for these is a hollow curve (Table 5; Figure 3) and one that is a remarkably good fit to the logarithmic series.

7. Why a logarithmic series?

Although logarithmic series are found so abundantly in biological systems the mechanism underlying their universality is by no means clear. Hollow curves that approximate to log series may be empirically generated by growing "trees" whose branch tips bifurcate randomly through time: the "tree" is subjected to regular random "killing" of the growing tips. The model is that of evolution and extinction. The tree is "sampled" by sectioning at a fixed level in time; the number of living tips



Entomologist Gaden Robinson pinning and packing moths in the Brunei rainforest.

subtended by each cut stem is counted. The resulting frequency distribution of extant tips among stems approximates to a logarithmic series. So it may be that the logarithmic series and similar hollow curves are a summation of the effect of random extinction points superimposed upon a similarly random pattern of binary development or binary "choices."

Projects in biology invariably generate more questions than they answer. Investigating frequency distributions that generate hollow curves to which diversity indexes may be related is no exception to this version of Murphy's Law. Meanwhile, when you next clean the bug bodies out of a light-fitting, pause a moment and reflect that you are almost certainly throwing a logarithmic series into the trash.

R.A. Fisher observed that if successive, independent, equal samples are taken from homogeneous material, then *N* (no. of individuals observed) is distributed as a Poisson (*m*) whose parameter *m* is the number expected. If the material is heterogeneous, or if samples are of unequal sizes, we have a mixture of Poisson distributions with differing values of *m*, that is m_i. He showed that an important extension of this Poisson theory arises when the m_i have a known distribution. Since the values of the m_i's are positive, the simplest distribution that can be assumed is of Eulerian form. He then showed that the resultant series is related to the negative binomial expansion. However, in its application to the number of individuals from *S* different species, the observed *N* in any sample cannot be 0. Consequently he derived the log series as the limiting case of the negative binomial.

References

Corbet, A.S. (1942), "The distribution of butterflies in the Malay Peninsula," *Proceedings of the Royal Entomological Society of London*, A.16, 101-116.

Fisher, R.A., Corbet, A.S. and Williams, C.B. (1943), "The relation between the number of species and the number of individuals in a random sample of an animal population," *Journal of Animal Ecology*. 12, 42-58.

Hayek, L. C. and Buzas, M.A. (1997), *Surveying natural populations*. New York: Columbia University Press. Appendix 4, pp. 415-486.

Magurran, A. E. (1988), *Ecological diversity and its measurement*. London: Croom Helm.

Patil, G. P. and Taille, C. (1982), "Diversity as a concept and its measurement," *Journal of the American Statistical Association*. 77, 548-567.

Robinson, G. S. and Tuck, K. R. (1993), "Diversity and faunistics of small moths (Microlepidoptera) in Bornean rainforest," *Ecological Entomology*. 18, 385-393.

Robinson, G. S. and Tuck, K. R. (1996), "Describing and comparing high invertebrate diversity in tropical forest - a case study of small moths in Borneo," pp. 29-42 in *Tropical Rainforest Research*—Current Issues (D. S. Edwards, W.E. Booth, and S. C. Choy, eds.), Dordrecht, Netherlands: Kluwer Academic Publishers.

Taylor, L. R. (1978), "Bates, Williams, Hutchinson—a variety of diversities," in *Diversity of Insect Faunas* (L. A. Mound and N. Waloff, eds.), London: Royal Entomological Society.

Williams, C. B. (1937), "The use of logarithms in the interpretation of certain entomological problems," *Annals of Applied Biology*. 24, 404-414.

Williams, C.B. (1944), "Some applications of the logarithmic series and the index of diversity to ecological problems," *Journal of Ecology*. 32, 1-44.

Williams, C.B. (1947), "The logarithmic series and its application to biological problems," *Journal of Ecology*. 34, 253-272.

Williams, C.B. (1964), *Patterns in the Balance of Nature*. London: Academic Press.

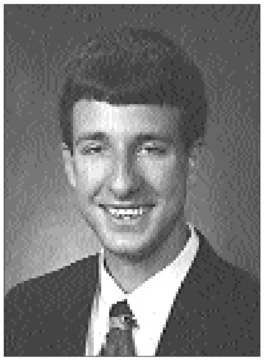
Williams, C.B. (1970), *Style and vocabulary: numerical studies*. London: Griffin.

Willis, J.C. (1922), *Age and Area*. London: Cambridge University Press.

Wolda, H. (1981), "Similarity indices, sample size and diversity," *Oecologia, Berlin*. 50, 296-302.

Wolda, H. (1983), "Diversity, diversity indices and tropical cockroaches," *Oecologia, Berlin*. 58, 290-298.

An Evolutionary Model for the Optimization of a Maze



Curtis S. Stratman

1. Introduction

1.1 Background

I am a class of 1997 graduate of Jefferson City High School in Jefferson City, Missouri. The school offers a class that meets one day a week in the mornings where the students prepare independent research projects to be presented at local science fairs, including the International Science and Engineering Fair and the Missouri Academy of Sciences. When thinking about a subject, my long time interest in computers led me to consider a project in machine learning. During a college visit at the University of Missouri-Rolla I came into contact with Cihan Dagli and he agreed to be my mentor. The classic concept of running mice through mazes was something I wanted to try reproducing on a computer. He said I might want to consider genetic algorithms. This paper is the result of my work of combining computers and biology by way of mathematics.

Ideas from many of my high school classes have gone into this project. Mathematics, biology, and psychology all serve as foundations. Several times I found an idea for this project during a daily lecture in one of these classes. The applications of this project extend even further, into fields such as physics and engineering. This project would not have been possible for me without a solid and diverse background to work from.

1.2 Problem Solving

Some problem situations cannot be solved using a definite procedure. An example of this would be finding the solution to maze games. We can watch little children as they attempt to find the solution. In the beginning they have no knowledge on how to approach the problem. As they gain experience, they learn strategies that help them find the path from the beginning to the end of the maze.

This is the same approach machine learning takes toward such problems. Instead of finding one exact answer to the problem by completing a series of defined steps, strategies are programmed so that a near optimal answer is found by trial and error. Knowledge and success are built upon to find increasingly better solutions to the problem. The goal is not necessarily to find the one exact answer, but rather to find one of the best. In this paper I will describe the approach I took to solving a maze problem. Starting with the idea of using the genetic algorithm, which is based on probabilities, I attempted to apply it to finding the shortest route through the maze and then tested the model's performance.

2. The Genetic Algorithm

2.1 The Natural Basis for the Genetic Algorithm

The way strategies are developed is by attempting to reproduce natural processes. For example, neural networks use a structure similar to how scientists believe the neurons in the brain might process information. These networks are then designed specifically to perform tasks such as facial recognition, speech recognition, predicting stock market trends, etc. It is important to note that these models are not meant to duplicate exactly what occurs in nature or to prove how a natural system works. This research simply uses nature as a basis for ideas and the models are developed further with mathematics.

The genetic algorithm uses sexual reproduction as its basis. Every living organism

Curtis S. Stratman is currently a student of the University of Missouri-Rolla studying physics. In 1997 he graduated 5th in his high school class from Jefferson City High School and was chosen as one of two Missouri delegates to attend the National Youth Science Camp held near Bartow, West Virginia.

contains DNA which is composed of chromosomes which in themselves are composed of individual genes. These genes produce the characteristics that an individual organism has, for example blue eyes. For organisms that reproduce sexually, the offspring receives a combination of genes from both of its parents.

Charles Darwin theorized that evolution is a result of sexual reproduction. In nature individuals compete against one another in their environment for survival. The concept of survival of the fittest states that those individuals with the best characteristics tend to live while the others die. The living individuals produce offspring, which through sexual reproduction have a new and unique combination of characteristics from their parents. This generation of offspring also compete against each other for survival and produce offspring of their own. As the process of improvement versus extinction continues, populations of organisms containing increasingly better qualities are formed.

2.2 Holland's Genetic Algoritm

The genetic algorithm was invented by John Holland (1975). He derived a mathematical proof, called the Schemata Theorem, which serves as the fundamental theorem of genetic algorithms. This theorem is too extensive to describe in this paper, but for those interested I would recommend reading Goldberg (1989) and also Jennison and Sheehan (1995). Both are available at many university libraries.

Holland designed the genetic algorithm to imitate the process of sexual reproduction and natural evolution. Each organism, or individual, has a single chromosome which is represented as a string of characters, each character being analogous to a single gene. An initial population is generated randomly to ensure no initial bias is present. This is referred to as generation 0. Each individual is then placed into the "environment" where its fitness value is determined. Fitness is a numerical value that represents the individual's ability to survive and produce offspring in its environment. Once these values are computed for each individual, the population of individuals reproduces by way of the genetic algorithm which follows rules that are similar to those that natural DNA follows in sexual reproduction. This gives us generation 1. This process, calculating fitness and reproducing, is repeated until a specified generation number is met. It is hoped that as the individuals that are best-fit pass their qualities on to the next generation through sexual reproduction, the population as a whole will evolve toward the desired solution.

The simplest form of the genetic algorithm contains three steps: reproduction, crossover, and mutation. Each is designed in such a way as to imitate the process of DNA recombination in sexual reproduction. In reproduction it is decided which individuals will produce offspring for the next generation. The probability of a single individual being selected is proportional to its fitness value. Selections are made with replacement so one individual may be copied into the mating pool more than once. Individuals are chosen until the size of the mating pool equals the original population size.

Crossover recombines the chromosomes to form new individuals. Each individual in the mating pool is paired with one other. A value called the crossover probability is determined by the designer of the model. It is determined which pairs will undergo crossover based on this probability. For example, if the crossover probability is set at 60%, then about six out of every ten pairs will undergo crossover. Pairs not selected are unaltered during crossover. For selected pairs, a split point within the chromosome is randomly selected. The characters to the right of the split point of each chromosome are simply swapped with those of its pair. An example of crossover is shown in Figure 1.

Figure 1. An example of crossover.

Before	Split Point = 5	After
11111111		11111000
00000000		00000111

Mutations are made to the population in order to introduce new information. This is done to keep the genetic algorithm from getting into a rut, so to speak. Perhaps all the individuals containing a key characteristic died out because the other characteristics they contained provided them with poor fitness. It is hoped mutations will reintroduce this material. With traditional binary strings mutation is done by changing a 0 to a 1 and vice versa. The characters selected for mutation are based on the mutational probability, another value determined by the designer of the model. If the mutational probability is set at 1%, then about one out of every one hundred characters will be mutated.

■ 3. The Maze Problem

3.1 Application of the Genetic Algorithm

So how can these ideas be used to solve a problem? That was the question I had to answer. I was able to research up to this point, but everything from here was up to me to figure out. Somehow I had to make the basic elements represent something within the problem situation where the individual's chromosome represented a possible solution to the problem and its fitness was the measure of that solution's value.

I chose to have the chromosome represent a path. It was a string of base-4 numbers (i.e. 01230130213203...) because there are four directions of movement in 2- dimensional space. Each character represented a direction of movement. 0 was up, 1 was right, 2 was down, and 3 was left. The chromosome was a series of single steps that when read sequentially from left to right guided the individual through the maze.

The method I devised for calculating fitness was the hardest step for me in the process of designing this model. The reason for this is because a maze is typically thought of as a visual image, not something that can be represented as a simple mathematical function. My first attempt was to simply count the number of steps each individual took before reaching the ending point and subtract that number from the length of the chromosome so higher numbers would indicate better performance. The problem with this method was that almost none of the initial population ever found the ending point and therefore had fitness values of 0. I then decided fitness would have to be calculated as an individual value for each step and then those values combined into a single number.

The maze was divided in grid fashion. Each occupiable position was assigned an *xy*-coordinate. The top left corner was the coordinate (1,1). *X* values increased as you moved to the right. *Y* values increased as you move downward, opposite that of the traditional Cartesian coordinate system.

The first factor I considered important was the distance the individual was from the ending point. There needed to be a value to subtract from as the distance to the ending point decreased so that fitness increased as the individual approached the ending point. I set up a ratio of the distance between the current position of the individual and the ending point as compared to the distance between the starting point and the ending point. Instead of using straight line distance I used a simple difference in *x* plus the difference in y to calculate the distance. This is sometimes referred to as taxicab geometry and was valid in this model since the individual could not move in a diagonal path. Since this value would always be less than one when the individual was closer to the ending point than the starting point was to the ending point, the value to subtract from was one. Note that this factor can return a negative value. Also note that the individual did not move after reaching the end-point. The remaining steps' values were calculated with the individual on the ending point. The formula for this factor is

$$\left[1-\frac{|X_E-X_C|+|Y_E-Y_C|}{|X_E-X_S|+|Y_E-Y_S|}\right] \qquad (1)$$

where X_E and Y_E = X and Y Coordinates of Ending Point,
X_S and Y_S = X and Y Coordinates of Starting Point, and
X_C and Y_C = Current X and Y Coordinates

The other factor I considered important was how fast the individual approached the ending point. The early steps are more important to optimize because there is no point in optimizing the last twenty steps when the first twenty are wasted. For this factor I set up another ratio, this time with the step number compared to the length of the chromosome. This factor should decrease with each step; and since this ratio will never be greater than one, the ratio was subtracted from one. The formula for this factor is

$$\left[1-\frac{Step}{LChrom+1}\right] \qquad (2)$$

where *Step* = Current Chromosome Position {1, 2, 3, ..., *LChrom*}, and
LChrom = Chromosome Length

The factors in (1) and (2) were multiplied together to give the formula used for calculating the value of each step. The values returned for each step were then summed. This value was then divided by the best possible value achievable if there were no walls within the maze. This

$$\sum_{K=1}^{K'-1}\left[1-\frac{K}{LChrom+1}\right]\left[1-\frac{|X_E-X_S|+|Y_E-Y_S|-K}{|X_E-X_S|+|Y_E-Y_S|}\right]+\sum_{K=K'}^{LChrom}\left[1-\frac{K}{LChrom+1}\right] \qquad (3)$$

where $K' = |X_E-X_S|+|Y_E-Y_S|$,
LChrom = Chromosome Length,
X_E and Y_E = X and Y Coordinates of Ending Point, and
X_S and Y_S = X and Y Coordinates of Starting Point

maximum value was calculated using the formula:

When this value was multiplied by 100, this final fitness value was the percent efficiency of the individual. Any individual whose final fitness value was less than 0 was assigned a fitness value of 0. This process for calculating fitness provided the genetic algorithm with the information it needed to evolve the population in such a way as to find an optimal path. We must look at the goal for this model and then examine how it must be achieved. Since the goal was to find the shortest route possible between the starting and ending points within the maze, it is obvious to see the role that the distance factor played in the fitness calculation. The shorter the route taken, the higher the fitness value. Therefore during the reproduction step of the genetic algorithm the individuals with shorter paths were more likely to be chosen. To understand how the speed factor influenced the process we must consider the process of how the problem must be solved. Since the value of every move was dependent upon every move that preceded it, the first moves had to be optimized as early as possible. Otherwise if the last moves had been optimized and then an earlier value was changed, those moves might no longer be applicable. By placing a higher weight on the first moves, the speed factor caused them to converge earlier.

3.2 Mutations

Since the chromosome is composed of base-4 numbers, the mutation step of the genetic algorithm could not simply swap 0s and 1s. I chose instead to swap 0s with 2s and 1s with 3s.

3.3 Other Learning Strategies

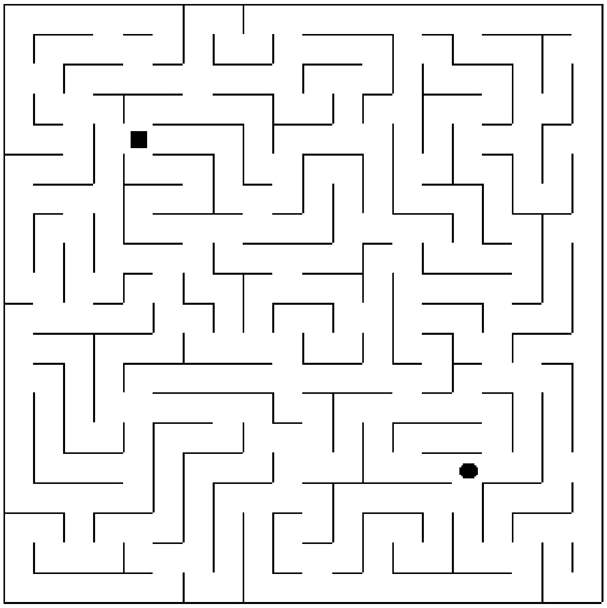
I chose to incorporate two additional learning routines in this model. The first was operant conditioning. Operant conditioning is defined as the conditioning that results from one's actions and the consequences they cause (McMahon et al., 1990). Obviously the model would be more efficient if the individuals did not waste their time attempting to move through the walls. If live creatures were used, such as mice, one might electrify the walls in order to condition them. Eventually they would stop making the attempt. To represent this learning in a mathematical model, I wrote a routine to simply remove every character in the chromosome that caused no movement. All the characters to the right of that location were then shifted to the left by one position and the end of the chromosome was regenerated randomly to preserve its original length.

The second routine was cognitive learning. Many animals have shown the use of strategies for exploring a maze without tracing the same territory more than once (Olton, 1978, 1979). I did not want to allow the individual to make a move that returned it to the position that it had previously left; in other words, not letting it make a U-turn. This was done by not allowing a move to the right to be followed by a move to the left or vice-versa and the same with upward and downward. To accomplish this the values 0 (up) and 2 (down) along with the values 1 (right) and 3 (left) were not allowed adjacent to each other within the chromosome. When an occurrence of these values being adjacent was found, both characters were removed. As in the above routine, the characters to the right of that location were shifted and the end of the chromosome was regenerated randomly.

Both routines assisted the genetic algorithm by removing noise from the data, but the second routine incidentally had an even more significant effect. It drastically reduced the dimensions of the search. Without the routine any character in the chromosome could be any of four values. The total number of possible chromosomes was therefore $4^{(\text{Chromosome Length})}$. With the routine the first character could still be any of four values but the subsequent characters could only be one of three values. For example, if the first character were a 0, then the next character could be a 0, 1, or 3 but not a 2. The permutations were therefore reduced to $4*3^{(\text{Chromosome Length - 1})}$.

These routines also increased the role of mutation. Instead of just altering a single character

Figure 2. The maze.



in the chromosome, such a change could cause several characters to be removed by these routines. The other characters on the chromosome would also be shifted to a new location earlier on the chromosome. A single mutation therefore could incidentally result in drastic change in the path taken by the mutated individual.

■ 4. Testing the Model

I personally drew the maze shown in Figure 2. The starting point is represented as a square and the ending point as a circle. Note that there is not one single solution like mazes found in game books. The number of possible solutions increases with every turn. Therefore our goal was not to find the only path but rather one of the shortest paths possible from the starting point to the ending point.

Table 1 shows an example run of this model through five generations. I arbitrarily selected the values 50 for population size and 50 for chromosome length. Goldberg (1989) suggested the values 60% for crossover probability and the reciprocal of the population size (in this case 2%) for mutational probability. By glancing at the best chromosomes of the progressing generations, one can see that individuals were being optimized beginning with the first moves and then a few more following moves in each successive generation. This follows the design of the model. One can also see by looking at the maximum fitness values that in this example the population was very quick to approach a near optimal path. (Recall that the maximum fitness value attainable is 100).

Table 1. Summary of example run.

		Fitness		
Generation	Best Chromosome	Minimum	Maximum	Average
0	1112222222211000333010112212 22100333032322233012123	0.00	46.04	20.04
1	111221100111122210333212230 01122230030103011030322	0.00	56.85	33.73
2	111221100111222112111232230 11000112321032332111222	5.62	75.56	38.30
3	111221100111222122233211223 23301223301012210301032	11.20	77.76	44.84
4	111221100111222122221001211 22223211121232301030011	15.80	81.24	47.90
5	111221100111222122221232111 22323010300110103300333	0.00	85.64	49.82

Figure 3 shows the path taken by the individual from generation 0 with the maximum fitness. This is the generation where all the individuals were created randomly. The

Figure 3. The path taken by the best individual from generation 0.

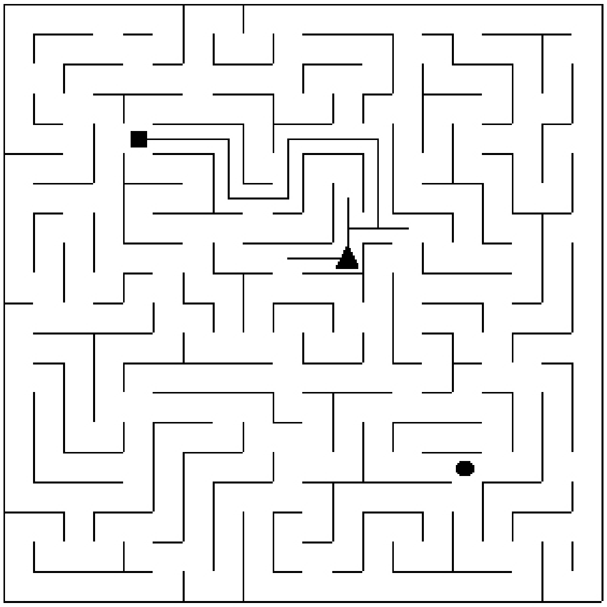
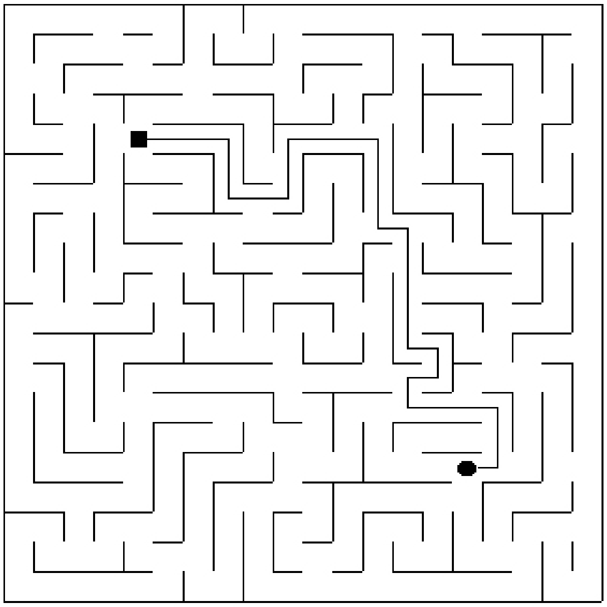


Figure 4. The path taken by the best individual from generation 5.



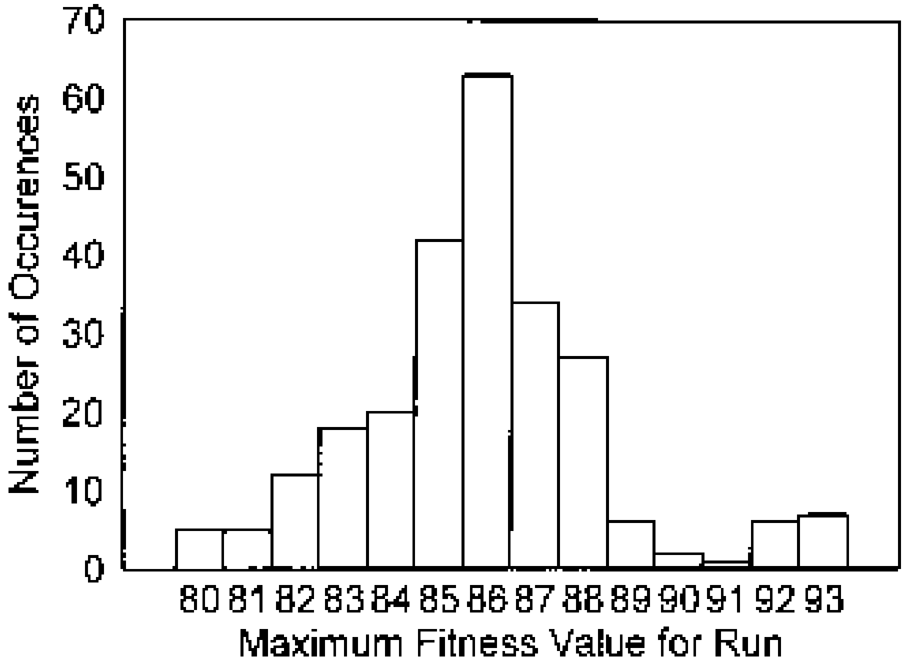


Figure 5. Histogram of the maximum fitness values.

triangle in Figure 3 denotes the individual's stopping point. It is likely that the individual in Figure 3 received the maximum fitness for that generation because of its good beginning path. Those values after all were weighted the highest. This individual did not reach the end because it ran out of steps on the chromosome, in this case 50. Figure 4 shows the best individual from generation 5. Notice that this individual reached the ending point. This path is a near optimal one.

To assure this was not a one time incident I tested the model in 250 separate runs, letting the population evolve through 50 generations in every run instead of just 5 generations. The histogram of the maximum fitness values of the 250 runs is shown in Figure 5. The mean of these values is 85.82 with a standard deviation of 2.56. This mean is higher than the fitness value of the individual shown in figure 4. This means that on average they found a path even more optimal than the one shown. Based on this observation and the small variation in fitness values, (that is, the model rarely had trouble finding a path with fitness less than 80), I therefore concluded that this model was successful in optimizing the given maze.

5. Conclusion

This project does not prove anything about genetics or evolution. It is simply a model which finds an optimal route through a maze. But this concept does have many real world uses besides playing games. By putting weights on certain paths of the maze, routes could be optimized by speed, distance, or cost. This would be useful for drivers or engineers planning street routes. The model could be simply modified to optimize 3-dimensional mazes by changing the chromosome to base-6 numbers. This would be useful when planning the wiring or piping of a building. In fact an infinite number of dimensions could be added by increasing the base by two for each additional dimension. The implications of that are beyond my imagination.

References

Goldberg, D. E. (1989), *Genetic Algorithms in Search, Optimization, and Machine Learning*, Reading, MA: Addison-Wesley Publishing Company, Inc.

Holland, J. H. (1975), *Adaptation in Natural and Artificial Systems*, Ann Arbor, MI: University of Michigan Press.

Jennison, C. and Sheehan, N. A. (1995), "Theoretical and Empirical Properties of the Genetic Algorithm as a Numerical Optimizer," *Journal of Computational and Graphical Statistics*, 4, 296-318.

McMahon, F. B., McMahon, J. W., and Romano, T. (1990), *Psychology and You*, St. Paul, MN: West Publishing Company.

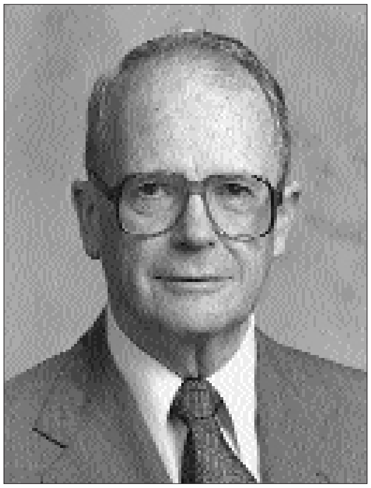
Olton, D. S. (1978), "Characteristics of Spatial Memory," in S. H. Hulse, H. F. Fowler, and W. K. Honig (Eds.), *Cognitive Aspects of Animal Behavior*, Hillsdale, NJ: Erlbaum.

——- (1979), "Mazes, Maps and Memory," *American Psychology*, 34, 583-596.

Acknowledgements

I would like to thank Timothy Adkerson, Martha Harris, Dennis Nickelson, and Terry Sontag as the teachers of the Science Research class at Jefferson City High School, the program through which I completed this project. I would also like to thank Ruthi Sturdevant for her assistance with the manuscript and Cihan Dagli of the University of Missouri-Rolla as my mentor.

Life & Hard Times of a Statistician



J. Stuart Hunter

When our editor Christine McLaren asked if I would write a review of my rites of passage into the world of statisticians, I quickly agreed. I enjoy reminiscing and there is always the chance someone might be interested.

In high school I was the youngest in my entire class and on graduation day in 1940 I turned 17. Being the youngest has its disadvantages, forget athletics, but if you are eager you overcompensate. So, I was active as a debater (won the state championship for boys), was the year book editor and busy in many school clubs. My grades were OK, but never great. I was good in math, but far from the best. I aspired to become a lawyer or a CPA (certified public accountant), but financial circumstances dominated any immediate considerations of college. My first job was with the Prudential Insurance Company as a mail-boy. For over a year I delivered messages and closely watched the lawyers and accountants at work. Frankly what they did day in and day out didn't look all that interesting. But the actuaries seemed an exciting bunch. They earned a lot more money too. I soon found myself punching keys on a Freiden calculator computing monthly annuities and thinking that perhaps, one day, I'd try the beginning actuarial exams. It was during that period that I started attending night school at Union County Junior College (UCJC or Ucey-Juicey). I took engineering math, the tougher of the two math sequences offered, along with a collection of liberal arts courses. And then the great war erupted and everyone's life was dramatically changed.

After basic training the army offered a program called ASTP (Army Specialized Training Program) in which soldiers were sent to a college to take engineering and science courses. To qualify one had to pass a series of exams. My decision to take engineering math at night school proved to be the crucial element in my being accepted into the ASTP program and sent on to N.C. State to take courses in electrical engineering. Later, and after a second dose of basic training, I ended up at the Army's Aberdeen Proving Grounds in Maryland and a member of a chronograph team. We traveled all over the Philippines determining the muzzle velocities of field artillery pieces with a clock that could count (the wonder of its day) to 10^{-5} seconds. There was data in abundance to analyze and I suspect I became, quite unconsciously, something of a statistician even then. I returned to N.C. State after the war to complete my Electrical Engineering degree in 1947.

To this day I can hardly believe my good luck. Across the campus Miss Gertrude Cox was establishing a statistics department and R.L. Anderson, who had taught me advanced calculus while I was a GI, was a faculty member. And so I decided to take some statistics courses. My life has never been the same since.

I think it very important that statisticians not be shy

I soon got a Masters degree from N. C. State in Engineering Mathematics with a Statistics minor and then promptly took a full time job as manager of the computing laboratory at the Institute of Statistics. This was back in the days of the IBM sorter, collator and tabulator. Doing simple tasks such as taking sums of squares and cross products could consume hours. I also found myself doing a lot of consulting with graduate students helping them complete their computations for their theses. After a couple of years I tired of working on the other guy's thesis. "Heck, I can do that." And so I resigned my job and became a full time Ph.D. student.

I spent my summers working for Frank Grubbs at the Aberdeen Proving Grounds. I now proudly recall one occasion when I stayed awake for 26 hours running the ENIAC (the world's first electronic computer). One day Dr. Grubbs asked what I was planning to do for a Ph.D. thesis. I replied, "Oh, something in factorial designs." He wondered if I had read the paper by Box and

Wilson (Box, G. E. P. and Wilson, K. P. , (1951), “On the experimental attainment of optimum conditions,” *Jour. Royal Stat. Soc.* Series B 13.) “Yes,” I replied, “And isn’t that linkage between regression and experimental design fascinating.” Miraculum: that autumn the Army Research Office proposed that Miss Cox invite George Box to Raleigh for one year as a research professor with Frank Grubbs as program supervisor.

Those were the days. I was newly married. I worked under Box on a great thesis topic: rotatable designs. (Box, G. E. P. and Hunter, J. S. (1957) “Multi-factor experimental designs for exploring response surface,” *Annals of Mathematical Statistics*, 28, 195-241.) After graduating in 1954 I secured a fine industrial job at American Cyanamid as “consulting statistician.” Further, my boss, Herbert Grosskopf, encouraged me to give short courses and public seminars. He was anxious that the arts of statistically designed experiments be announced to the industrial world. Then in 1958 came an invitation to become a member of the Statistical Techniques Research Group at Princeton University with George Box and John Tukey as the co-directors. Not too long after I became the first editor of *Technometrics*, then moved on to Wisconsin for several years and back to Princeton for another twenty-five years. I am now enjoying an active retirement.

You may be surprised to learn that I have never, formally, been a member of a statistics department. My academic appointment at Princeton was in the School of Engineering and Applied Science, initially in the Chemical Engineering Department and later in Civil Engineering. Thus I’m a hybrid: part engineer and part statistician, or perhaps more accurately a statistics stem grafted onto an engineering root. And I’d like to propose that being a “hybrid” is a very good idea. The field of statistics becomes positively lively once you can engage the language and thinking processes of a specialized subject along with that of statistics. I can definitely declare that anyone with an interest in a science, or medicine, or engineering, or even politics who has an interested in handling data will find a career in statistics rewarding.

I have always been a good teacher. A gift. In addition to my on-campus lectures I guess that I have taught a hundred or so industrial short courses, not counting a TV series of 32 one-half hours. In addition, I have always enjoyed spreading the word that statistics is exciting, useful and fun and to this day I seldom turn down invitations to speak to local groups. I mention this because I think it very important that statisticians not be shy. Too many in society think of us as dull number-crunchers. They are wrong! Thus, whenever we get the chance to talk about what it is we do, we should accept and make it as interesting as possible.

Over the years I’ve collaborated on a host of different projects. Before satellite observation was possible, aircraft flying overseas were on their own for long distances and I was part of the team that determined the overseas flight corridors between the USA and Europe—Asia. I helped in the design of the automobile airbag, consulted on many many chemical pilot plant investigations, conducted a local election survey, helped in the study of pollution from coal fueled power generators using radio-isotopes, was part of a team that reviewed the repeatability and reproducibility of EPA’s national laboratories, testified before Congress on the statistical aspects of environmental monitoring, and have consulted in varied industries: chemicals, metals, glass, pharmaceuticals, auto, micro-electronics, ... My latest consultations concern the development and manufacture of very small but powerful batteries which, if successful, will greatly influence the future of the electric car. In all these activities my role has been that of a consultant or participant and although I may not lead a project I do try to pull a strong oar. There are not many dull moments.

What message is there in all of this ruminating for the young nascent statistician? To be sure, luck plays a role in one’s career, and I have had my share. But so does being happy with who you are and with what you are doing. My main message is that statisticians are not mathematicians laboring in some secluded field, nor number librarians keeping data in neat files. We are professionals who recognize the important difference between data and information and thus have the unique task of shaping the links between ideas and their quantitative measures. In many ways we are acolytes to the scientific method, lighting the way from conjecture to data acquisition, and through the pathways of analysis helping in the creation of new ideas and repetition of the learning cycle.

Frankly, the profession of statistics is unlike any other in the excitement it can provide. What other professional can you name who is equally welcomed (sometimes feared) by the research worker, the doctor, lawyer, politician, or everyday citizen immersed in data? What other philosophy provides the language and logic for elucidating the scientific method? No other. We statisticians are a unique clan.

Welcome aboard.

Karen Kafadar

When Rick Burdick, associate editor for *STATS*, asked me to submit an article for the column "On The Job," he suggested that I describe several typical "days" from my past experiences which have ranged from government to industry to academe. Each position offered its own challenges, rewards, and, most of all, learning experiences which, when taken together, contributed to an overall package that has been valuable in teaching and research.

My first permanent job following my doctorate was in the Statistical Engineering Division at National Bureau of Standards (NBS), now National Institute of Standards and Technology, supervised by Dr. H.H. Ku. This position was a haven for a new Ph.D. The division's expertise was well respected by Bureau scientists as well as by other government agencies, so challenging problems were at one's doorstep. Projects required tools from all areas of statistics—designing an experiment to measure the force of gravity to one more decimal place, deriving calibration curves for weights and measures, assessing the uncertainty in a Bureau-certified amount of a chemical or compound in a standard reference material used to calibrate customers' measuring devices—and I learned an enormous amount about statistical practice from my highly qualified colleagues. A typical day might involve the analysis of data from one project, one or more meetings with scientists from other projects describing their needs and objectives (usually at the scientists' laboratories), and weekly or semi-weekly statistics seminars with either outside speakers or in-house statisticians or scientists presenting their work for comment and feedback. The atmosphere was lively and fun but very active. There were hallway conversations where ideas were exchanged and office doors were wide open, to encourage collaboration, where people could be seen working at their desks to derive models, design studies, analyze data, or prepare reports.

Many interesting projects arose from other government agencies. On one occasion, the National Archives and Record Service (NARS) solicited NBS assistance in designing a cost preservation survey of the vast array of historical documents in their archives. Keith Eberhardt was (and still is) an expert in surveys and sampling, and, consistent with our division chief's philosophy of two persons per project, included me as the backup statistician on the project. Keith and I made several trips to NARS in downtown Washington for meetings to determine the sampling frame, develop the appropriate design, and to conduct (personally!) the pilot survey. That experience taught me why one should never do a survey without a pilot; we learned a tremendous amount from the pilot data. I also dragged Keith across the street to the National Gallery of Art during lunchtime — he was always a good sport and indulged my insistence on experiencing a bit of culture while we had the opportunity!

My experience with primary standards provided me with excellent training as a statistician in Hewlett Packard's Stanford Park Division which manufactured electronic instruments, such as power meters and signal generators, "traceable to NBS" (measurements calibrated against NBS' standards). The atmosphere in industry was quite different from a government research laboratory. Things happen very fast in industry: a process that was running last week may be entirely different this week if the supplier had to substitute different materials needed to manufacture the components. Much more dependence on other processes and departments required more cooperation among those involved. Rarely did we have the luxury to derive the "perfect" solution, as time and cost were crucial considerations. (My brother in industry was once told, "The best is impossible. Second best is too expensive. Go with the third and get on with it.") Production floors were incredibly active and full of noise with busy workers operating complex machinery. While the research and development laboratory was less noisy than the production floor, the open partitions, versus structured offices at NBS, led to a more active atmosphere.

Perhaps it was the nature of my job, but, somewhat surprisingly, I had many more meetings at HP than I have ever had while working for the federal government. Some meetings involved managers who needed guidance on measuring productivity in their departments, others involved production line supervisors requesting assistance in identifying influential variables on part defect rates, and still others involved R & D engineers designing products to meet specific performance objectives. Some projects were short; I wrote several memos each week with summaries of the consultation and recommendations for further actions, ranging from the usual "How large a sample will I need?" to specification limits from measurements on several sources of variation.

One of my favorite stories involved a project manager who needed specification limits on the power output of an instrument developed in his group. He had 100 measurements on only one lab prototype instrument, not on a sample of representative instruments that had been manufactured on the production line. I explained that he needed at least one, if not several, more instruments, so he measured a second one: "I did what you said and got different answers. Now what do you want me to do?" Despite the sometimes less than full appreciation for measurement uncertainty, the precision and quality of the measurements in an engineering setting such as those from NBS and HP is a real blessing—just ask anyone who deals with data on real people, subject to many sources of extreme variation.

That experience taught me why one should never do a survey without a pilot

A typical day at HP would start usually with meetings or with returning phone calls to people who had left notes on my desk asking to set up a meeting. The most active time of the day was 9-3; after 3, many of the production workers who started at 6 or 6:30 had gone home, and others sat down to some quiet work. It was not uncommon to see people working well past 7 in preparation for the next day. Statistical questions arose throughout the division, from production and R & D to marketing and personnel, and beyond, across the company to other divisions. Adopting a "lean and mean" approach, there were often only 1 or 2 statisticians per division (of usually 1000-1500 people), so we statisticians often contacted our colleagues in other divisions to share experiences, programs, and other tools.

During my seven years at HP, I had the good fortune of being appointed Associate Editor of *Technometrics* by then-editor William Meeker at Iowa State. So some of my "typical" days at HP, and since (to present day), involve editorial work: reading manuscripts, assigning and writing to referees, summarizing reports and preparing an overall recommendation to the editor. I am very grateful to Bill Meeker and to his successors for giving me the opportunity to stay current with the latest statistics research while providing an important service to the profession. HP of course did not really budget time for its employees to conduct such outside research, but I had a terrific and very supportive supervisor, Julius Trager, who endorsed this kind of career development for his statisticians.

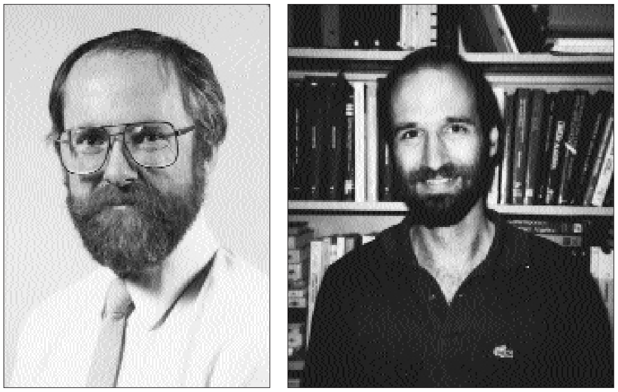
Both of these positions, together with a brief research fellowship in the Biometry Branch at the National Cancer Institute, provided me with real-life experiences that I enjoy sharing as a professor at the University of Colorado–Denver . Faculty interests in our department lie in primarily one of four areas in applied mathematics: computational math, discrete math, optimization, and probability/statistics. My "typical" day here is not too different from that of David Moore (see *STATS* 1995, 13, 26-27): grade assignments, hold student office hours, review notes for the day's lecture, attend faculty or committee meetings, solicit advice on administrative matters from our program assistant, handle editorial matters, return phone messages, check postal and electronic mail, and, on a rare day of luxury, scan articles from the latest statistics journal to arrive in the mail. The biggest difference is that most of my classes are taught after 4 P.M. So I learned a trick from Dick Jones, head of the graduate program in Biometrics at CU's Health Science Center: reserve mornings for research-related activities, and schedule meetings with students, colleagues, and clients for the afternoons.

One of the best aspects of our department is the potential for cross-fertilization with collaborators in various areas of mathematics. I am just starting to learn about some of the problems facing the computational mathematicians and hope to contribute a statistical component to their mathematical models of physical phenomena.

I feel very fortunate to have had a wide variety of experiences in my professional career. Many, if not most, university departments are cautious about hiring people from government or industry, even if they have maintained a publication record. And yet it is precisely those kinds of experiences which have made it easier for me to teach statistics to students and to communicate with clients, both within and outside of the University. I would encourage every student to take advantage of these opportunities as they arise. As academic departments acquire greater respect for applied statistics, they may be more inspired to hire faculty from industry or government so students can learn from their experiences and be better prepared for their own careers.

Bob Stephenson
Professor of Statisics
Iowa State University

Jeffrey A. Witmer
Professor of Mathematics
Oberlin College



Bob Stephenson **Jeffrey A. Witmer**

Dear Dr. STATS,

I am a science teacher in a middle school and teamed with a math teacher. She has been telling me about how statistics is a big part of the K-12 mathematics curriculum. Since I do a lot of hands-on experiments that involve observation and measurements, she thinks that my science classes are a natural place to expose students to the ideas of statistics. Where can I go to get information on how to use statistics in my science classes?

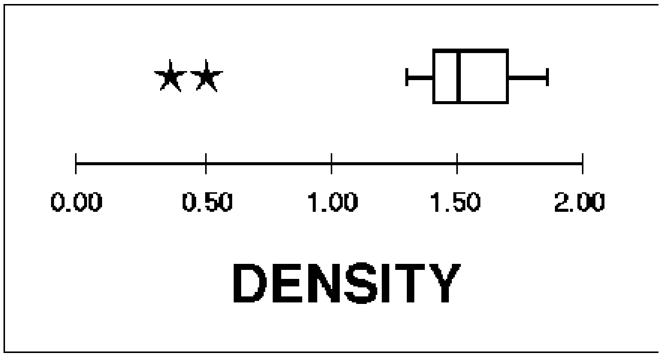
Signed,
All this data, and no where to go

Dear All This Data,

The Quantitative Literacy (QL) project is an effort on the part of statistics and mathematics educators to improve how statistics is viewed and taught by high school mathematics teachers. Since the mid 1980s, the American Statistical Association's Center for Statistical Education has organized Quantitative Literacy workshops at various places around the country. These efforts have been in concert with the National Council of Teachers of Mathematics movement to revamp mathematics instruction with their Curriculum and Development Standards. I took your question to Jeff Witmer, Professor of Mathematics, Oberlin College, a member of the QL project team and a task force member for the Science Education And Quantitative Literacy (SEAQL) project (see *STATS*, No 12, Fall 1994 pg. 20). Here is what he had to say.

Science students routinely collect large amounts of data that are used to answer specific questions. In a typical science class, each student (or team of students) completes a procedure and determines some sort of answer, for example, the density of a substance. Rarely are class data compared to anything other than an accepted value, as found in a reference book. SEAQL seeks to foster genuine exploration of data in science laboratory activities that promote a view of science as exploration and modeling, rather than primarily as confirmation of facts that are already known.

Let's look at the density example a little more closely. In physics classes in high school, or even physical science in middle school, a common experiment is determining density using water displacement. This experiment provides a great opportunity to collect and analyze class data. Below is a boxplot of 24 determinations of the density of a nut and bolt; measurements are in grams per milliliter.



Clearly, two of the observations are unusual (outliers). The class can investigate and discuss what laboratory experience led to these two values. The other observations suggest that the density is near 1.50 g/ml, but there is considerable variation in these values as well. We do not know the "correct" answer here, nor can we look up the answer in a reference book. Rather, we have explored the situation by experimenting.

Jeffrey A. Witmer is Principal Investigator of the SEAQL project, an NSF Teacher Enhancement grant. A past-Editor of STATS, *he has presented over a dozen workshops for middle and high school science and math teachers. His E-mail address is jeff.witmer@oberlin.edu.*

The collection, display and discussion of data for the whole class, or even among several classes, helps students develop an appreciation of inherent variation, measurement bias and precision. Indeed, experience suggests that students are more inclined to try to be accurate and precise in their measurements when they know that their data will end up as part of a class boxplot—no one wants to be an outlier!

Other science activities that lend themselves naturally to the sort of data collection and analysis promoted in SEAQL are:

—Determining the acceleration due to gravity using two or more methods

—Investigating the relationship between temperature and volume of a gas at constant pressure (Charles' Law)

—Simulating radioactive decay through the use of a shoe box, corn seeds and sunflower seeds

and many others. Prototype activities have been developed for biology, physics, chemistry, earth science and general science (typically taught in middle school to students age 12-14).

The outreach mission of SEAQL is to bring this information to science teachers through workshops. In SEAQL workshops, which last from two to four weeks, teachers are taught data analysis techniques using the *Exploring Data* QL book (*Exploring Data*, Revised Edition, by J. Landwehr and A. Watkins, Dale Seymour Publications) and are given experience using these techniques with data that are generated during the workshop. The teachers participate in science labs in biology, physics, chemistry, earth science and general science that are, for the most part, familiar to the teachers. They then use QL ideas in analyzing the data.

Other aspects of the workshops include instruction in the use of graphing calculators and calculator-based lab equipment, such as a temperature probe for gathering data during a heat of reaction experiment, discussion of non-standard labs that teachers have used with success, group projects in which participants gather and analyze data of their own choosing, time for teachers to prepare lesson plans as they consider how they will use SEAQL in their science classes, and brief consideration of statistical aspects of experimental design.

SEAQL is funded through 1998 by a grant from the National Science Foundation. If you would like more information about SEAQL please contact: The ASA Center for Statistical Education, 703-684-1221.

JERRY KEATING
Professor of Statistics
University of Texas
at San Antonio

STATS Quotes

"By a small sample we may judge the whole piece"
—*Miguel de Cervantes Saavedra (1547-1616)*

STATS from the Front Page

On October 28, 1997, the Dow Jones Industrial Average (DJIA) dropped 554.26 points: the largest single day point loss in the market's history. Trading was halted. However, market analysts advised investors that this was not a market collapse because the drop only represented a 7.2% drop. *{As you know the DJIA is an example of a time series. To what statistic are the analysts referring investors?}* This was probably not much solace to Bill Gates the Executive Director of Microsoft Corporation whose stock dropped $1,760,000,000 in value. And you thought that *you* had a bad day!

Answer: Signal to Noise Ratio

The 10 worst days for the DJIA in terms of percentage loss are given below:

1) October 19, 1987	22.61%	
2) October 28, 1929	12.82%	**Market Collapse leading to the Great Depression**
3) October 29, 1929	11.73%	
4) November 6, 1929	9.92%	
5) December 8, 1899	8.72%	
6) August 12, 1932	8.40%	
7) March 14, 1907	8.29%	
8) October 26, 1987	8.04%	
9) July 21, 1933	7.84%	
10) October 18, 1937	7.75%	

October has a rather ignominious distinction in that half of the 10 worst percentage losses of the DJIA occurred in October.

STATS from Science

As I watched the recent movie release, *Volcano*, with that scientific subtitle: *The Coast is Toast*, I wondered what was the worst volcano in history? Due to lack of measurement devices in prehistoric times we cannot piece together the worst cataclysm unless of course you support "The Big Bang Theory" of the origin of the universe. Nonetheless, here are some record values on another volcano that you can add to your disaster portfolio.

On August 27,1883, Krakatoa (Krakatau) erupted in the Indonesian arc. Anecdotal evidence includes that the explosion was heard on Rodriguez Island some 4,653 km across the Indian Ocean. At least 36,417 persons were killed most by giant sea waves, which reached heights of 40 m above sea level. Blue and green suns were observed as fine ash and aerosol reached as high as 50 km into the stratosphere. The volcanic dust lowered global temperatures as much as 1.2 degrees Celsius. Certainly, this eruption was not Krakatoa's worst, which most likely occurred around 416 AD.

STATS Parodies

We're adding a new section, *STATS* Parodies, to **OUTLIER ... s**. In this issue we feature two parodies by Professor Mark Glickman, Statistics Professor in the Mathematics Department at Boston University. His first parody is of the **Rolling Stones**' mega-hit, "Satisfaction," and the second is a parody of the **Beatles**' hit, "Nowhere Man." Mark is pictured on the next page tuning up his guitar, getting ready for class.

To get the full effect of Mark's contribution listen on Mark's web page at

http://math.bu.edu/people/mg/music.html.

If your system is a bit slow like mine be patient, it takes a while to upload the music.

Title: "Satisfaction"
Words: Mark Glickman
Music: Jagger/Richards ("Satisfaction")

I can't get no statisfaction,
I can't get no statisfaction.
'Cause I try and I try and I try and I try.
I can't get no, I can't get no.

When I'm sitting down at lecture,
And that man begins to explain to me
That you must pay close attention
When you're fitting regression
To heteroskedasticity!
I can't get no, oh no no no.
Hey hey hey, that's what I say.

I can't get no statisfaction,
I can't get no statisfaction.
'Cause I try and I try and I try and I try.
I can't get no, I can't get no.

When I'm working on my homework,
And I'm filled with great uncertainty,
So I choose the pooled procedure,
And my teacher points and laughs at me
'Cause our *p*-values don't agree!
I can't get no, oh no no no.
Hey hey hey, that's what I say.

I can't get no statisfaction,
I can't get no statisfaction.
'Cause I try and I try and I try and I try.
I can't get no, I can't get no.

When I'm handed my diploma
For my hard earned Bachelor's degree,
And the Dean says I cannot leave
Until I give an explanation
How to compute a correlation.
I can't get no, oh no no no.
Hey hey hey, that's what I say.

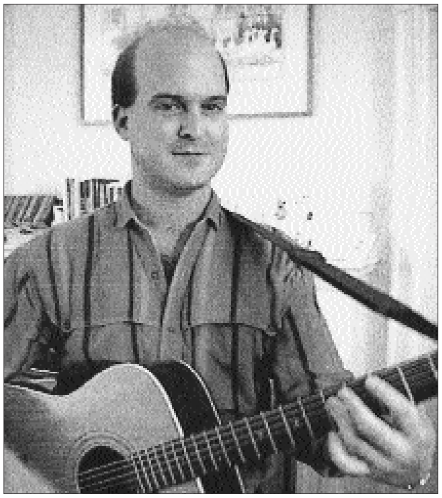
I can't get no statisfaction,
I can't get no statisfaction.
'Cause I try and I try and I try and I try.
I can't get no, I can't get no.

Title: "ANOVA Man"
Words: Mark Glickman
Music: Lennon/McCartney ("Nowhere Man")

He's a real ANOVA man
Designing all his sampling plans
Calculating mean-squared errors and *p*-values.

Wants to test for equal μ's
Knows which tables he must use
All his samples he will choose at random.

ANOVA man, please listen;
Where's the data that you' re missing
ANOVA man, what kinds of bias can you
withstand?



Writes down two hypotheses;
Hopes to reject the first of these;
Needs to list out all degrees of freedom.

ANOVA man, try harder,
Don't give up you're smarter;
ANOVA man, how come your students don't
understand?

At 0.05 he rejects
Ignores the size of his effects
Now he's stuck - he's got selection bias!

ANOVA man, please listen;
Where's the data that you're missing
ANOVA man, what kinds of bias can you
withstand?

He's a real ANOVA man
Designing all his sampling plans
Calculating mean-squared errors and *p*-values.

All you headbangers stay tuned to our next issue for Peter Westfall's parody of "Takin' Care of Business" by the Bachman Turner Overdrive. If you have original songs or parodies of popular hits send them to me at the address given on pg. 1.

STATS Pop Quiz

I. There are four colored balls in a bag: two red balls, one black and one blue. If you draw two balls at random, and then you're told that one of them is red, what is the likelihood that the other ball is also red?

Taken form *Mind Bending Puzzles*
by Terry Stickles.
Answer: 20%

II. The following data are the electrical consumption in kilowatt-hours of the Keating household for the months of March and April from 1986 through 1992.

Year	92	91	90	89	88	87	86
March	973	923	944	1,128	852	948	884
April	900	998	875	1,136	919	970	1,059

Which observation would you omit to create a sample with a skewness as close to that of the original sample as possible? Which observation would you delete to maximize the skewness? Can you generalize your findings?

Answers: *Maintain skewness:* *973*
Maximize skewness: *1,059*

STATS Revisited

In response to an earlier problem that I gave in *STATS* (Spring 1997, Number 19, p. 29), let me provide the following explanation:

Question: A certain blood test was 99% accurate in declaring the presence of the HIV virus among a group of persons known to be HIV+. Among a control group of persons known to be HIV-, the same blood test had a false-positive rate of 2%. If 0.3% of a population are HIV+, what is the probability that a randomly chosen person in this population is HIV+ given that the blood test is positive?

Explanation: Consider a representative group of 100,000 persons so that exactly 300 are HIV+. You would expect that 297 (99%) of these HIV+ members would test positive based on the blood test. Of the 99,700 who are HIV-, we expect the blood test to falsely declare 1,994 members (2%) as positive. Then you have an expected total of 2,291 members who test positive but only 297 are indeed HIV+. The subsequent ratio 297/2,291 produces a probability of 0.129638.

STATS Funnies

Can you identify the statistical terms in the two cartoons given below? These were taken from "Lower Bounds on Statistical Humor" by Alan H. Feiveson, Mark Eakin and Richard Alldredge. The artwork is by Kathlene Senghaas and Mark Eakin.

Answer: Contingency Table

Winners of the *STATS* Face the Facts Contest

These readers recognized the cover photo of Cleveland, Ohio (STATS, Fall 1997, Number 20):

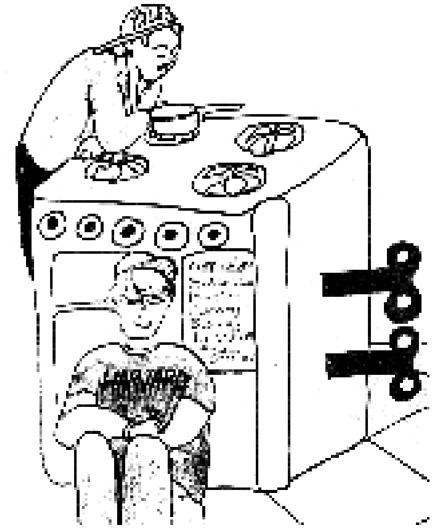
Chris Andrews, University of California, Berkeley, CA;
Rick Cornez, University of Redlands, Redlands, CA;
Yan Liu, The George Washington University, Washington, DC;
Lisa Rybicki, Cleveland Clinic Foundation, Cleveland, OH;
Tom Short, Villanova University, Villanova, PA;
Steve Wang, The University of Chicago, Chicago, IL.

Readers sending in other excellent suggestions (Chicago, Pittsburgh, and Cincinnati):

Andrea Hofer, University of Vienna, Vienna, Austria;
Todd Schwartz, University of North Carolina at Chapel Hill, Chapel Hill, NC;
Don Reed, Georgia State University, Atlanta, GA

New STATS Species-diversity Contest

For the research described in "Bugs, Hollow Curves and Species-diversity Indexes," Robinson used empirical methods to estimate the variance of the number of shared species between pairs of samples (p. 11).



Answer: Tukey (two-key) Studentized Range