STATS

The Magazine For Students Of Statistics

Winter 2002 • Number 33

Production

Megan Murphy
email:
megan@amstat.org

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

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

STATS is published for beginning statisticians, including high school, 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 Editors; addresses of the Editors 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.

Features

Departments

Editors' Column

Beth Chance **Allan Rossman**

We are very excited to be the new editors of *STATS*, and we look forward to continuing the fine tradition established by previous editors. We especially thank our predecessor Jerry Keating for his years of service, for maintaining the high quality of the magazine, for inviting us to join the team of *STATS* editors under his leadership, and for all the assistance he has provided during the transition period.

We remain committed to the original mission of *STATS*: to provide a first-rate publication that speaks to the interests and needs of *students* of statistics. As previous editors have done, we define "student" quite broadly. We include under that classification not only high school, undergraduate, and graduate students, but also teachers of statistics who never stop learning and finding new ways to help their students learn, and also professional statisticians as well. We hope that *STATS* will continue to entertain as well as to educate all of these groups.

In this premier issue of our editorship, the lead article is a very timely one. Particularly since the events of September 11, 2001, the need for effective security in situations ranging from airports to theme parks, from computer accounts to bank accounts, is paramount. The use of biometric identification devices is becoming more important and more common. Michael Schuckers' article introduces readers to these devices and to some of the statistical issues that accompany their development and use.

A second article concerns the difficulties in obtaining realistic estimates in complicated surveys. Counting animals may sound simple, but Willard Losinger's article highlights some of the challenges involved and how a method called "raking" can help adjust estimates in a large-scale farm survey.

An article from the Spring 2001 issue of *STATS* described the use of randomization tests to study infant handling by female baboons. We are delighted that this article inspired Cliff Lunneborg to analyze the data further, and we present his sensitivity analysis in this issue.

Larry Lesser, a published songwriter as well as a statistics educator, has penned (and performed) several clever and entertaining lyrics based on popular tunes that deal with statistical issues. Be forewarned, reader — you may find yourself singing those songs as you read them.

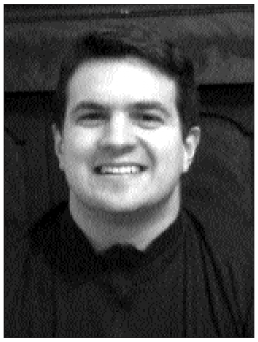
Three new members of our editorial board are contributing the first installments of their features in this issue. Gretchen Davis of Santa Monica High School and UCLA takes over responsibility for the AP Statistics column that aims to provide articles of interest to students and teachers of this course. Her feature presents a clever and enlightening way to think about the often elusive concept of degrees of freedom. Chris Olsen of George Washington High School in Cedar Rapids, Iowa, debuts a new feature called μ-sings: Statistics in the Media that provides recommendations and reviews of books, movies, and other aspects of popular culture that deal with statistical issues. In this issue Chris provides a very informative and entertaining review of a book titled *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. Robin Lock of Saint Lawrence University also joins our editorial board and initiates a new feature called The Statistical Sports Fan, in which Robin will examine applications of statistics to sports and other recreational pursuits. In this premier article he examines scoring and pointspread data from the National Football League and suggests how you might succeed if you participate in office pools.

We would also like to take this opportunity to introduce the other members of our editorial board. Jackie Dietz of North Carolina State University continues her association with *STATS*, joined by Patti Collings of Brigham Young University and Dave Fluharty of Continental Teves in Auburn, Michigan.

Another new feature in this issue is Data Sleuth, in which we present examples of the detective work involved in statistics. This issue includes a contribution from Rick Burdick, based on a mystery that arose from data analysis as part of a student project. We also provide a mystery from one of our favorite datasets that we hope will appeal to baseball fans.

We hope that readers of *STATS* will enjoy this issue as much as we have enjoyed working with our editorial board and authors in putting it together. We would very much appreciate your feedback on this issue, your suggestions for future articles and features, and especially your contributions. Please contact us at the e-mail

Some Statistical Aspects of Biometric Identification Device Performance



Michael E. Schuckers

A biometric device or biometric identification device is one that captures a physiological 'image' and uses that image to permit or deny access. The access being controlled could be to a computer account, to a room, or to a theme park. The goal of these devices is to provide a more accurate and secure method for physical or logical access. Everyone has a story or has heard one of someone forgetting his or her password and not being able to 'log in' to an account. Almost everyone also has a story about losing a set of keys. Biometric devices are meant to be an improvement over keys and passwords, since these latter devices can be lost or forgotten. In theory, your "biometric" cannot be lost or stolen because it is specific to you. Note that this is a different usage of the word biometrics. Biometrics, as the statistics community commonly uses it, refers to statistical or mathematical analyses of biological phenomena, as in the journal *Biometrics*.

A wide variety of biometrics are in use. A biometric is the physiological image that is used to determine identity. A partial list includes fingerprints, hand geometry, finger geometry, hand vein patterns, ear geometry, face recognition, voice recognition, retinal scans, iris patterns, handwriting, keystroke dynamics, and walking gait. A more complete list can be found in, for example, Jain, Bolle, and Pankanti (1998).

Since the tragic events of September 11th, biometric identification devices have received a great deal of attention and scrutiny. These devices potentially add an additional layer of security for personal identification. As such, they are a greatly desired commodity at this time. For several years the Department of Defense has been investing resources in testing the effectiveness of these devices. More recently, the Federal Aviation Administration and the airline industry have begun consideration of the use of biometric devices for increased travel security. Congressional hearings have also been held to determine the appropriate role for the federal government in promoting and developing the use of biometric devices.

Statistics has an important role to play in the development and evaluation of biometric devices. Each time a device is used, it must reach a decision about whether to grant or deny access to that person. Thus, there are two types of errors that can ensue: granting access to a person who does not deserve it, and denying access to a person who does deserve it. These are similar to Type I and Type II errors, and different biometric devices can be compared based on their error rates. Estimating variability in these error rates is an open statistical problem, because the assumptions of the conventional binomial model are satisfied only in very restrictive settings. Alternative methodologies have been proposed, but as yet none has achieved widespread acceptance.

This article provides background information on biometric identification devices and then explores some of these statistical issues.

Biometric Characteristics and Subsystems

The goal of a biometric device is to accurately determine whether or not you are who you say you are. Several factors go into a 'good' biometric device. Jain et al.(1998) suggest that a biometric should possess the following characteristics: universality, uniqueness, permanence, collectability, performance, acceptability, and circumvention. Universality means that as many people as possible should have the biometric in question. Not every person has a right index finger, so that a biometric device based solely on this will not be universal. Next, uniqueness implies that each person should have a different version of the biometric. Fingerprints are generally thought to be unique. Permanence is the condition that the biometric should not change over time. A biometric device based on facial recognition is not ideal in this sense because people change their hair, they grow

Michael E. Schuckers (mschuke@stat.wvu.edu) *is an assistant professor in the Department of Statistics at West Virginia University and is an active member of the Center for Identification Technology Research at West Virginia University. His research interests are in hierarchical models and Bayesian methodology.*

Figure 1. A fingerprint image.

beards and they get wrinkles. The ease with which a biometric can be captured is its collectability. It might be possible to create a biometric device based upon your electroencephalogram (EEG), but it would be difficult to capture that information quickly and easily. On the other hand, a fingerprint or an iris is fairly exposed and, therefore, easily collectible. Performance measures how easy a particular biometric is to use and implement. Acceptability is the degree to which the there is public acceptance of the biometric for identification purposes. Fingerprints are a prime example of a biometric with high acceptability, since they have been used for centuries as a method of identification. Finally, circumvention is the amount of effort required to fool the system. Signatures are notoriously easy to reproduce, whereas fingerprints are far more difficult to copy.

The basic biometric system contains five subsystems: data collection, transmission, signal processing, a decision-making algorithm and a database (Wayman, 1998). The data collection mechanism is a sensor of some kind. For facial recognition, the data collection mechanism is a camera. For keystroke dynamics, the data collection mechanism is a keyboard. The information from the data collection mechanism is then transmitted to the signal-processing unit. As part of the transmission, techniques such as signal compression may be implemented on the presented biometric. The signal-processing unit then extracts the relevant details from the transmitted image and compares that image to one or more stored images, or templates, of the biometric from the database. The decision-making phase then decides whether or not the presented biometric is ‘close enough' to the stored template to be considered a match. Several aspects of this process, particularly the decision-making step, have a statistical flavor. In this paper, I'll focus on the statistical aspects of the performance of the decision making process.

Figure 2. A fingerprint-based biometrics identification device.

Matching Performance

One of the most important aspects of any biometric device is its matching performance. The matching performance is usually measured in terms of false accept and false reject rates. (Within the biometrics industry, these are sometimes referred to as the false match and false non-match rates, respectively (U.K. Biometrics Working Group, 2000)). I will refer to users enrolled in the database as genuine users and to those not enrolled in the database as imposters. Thus, the matching performance describes how well the system allows access to genuine users and denies access to imposters.

When an individual presents his or her biometric, the ‘image' is processed and matched against one or more stored templates from the database. The number of comparisons depends upon the mode that the device uses. There are two basic modes of operation. The first is verification, or one-to-one mode. In this mode, some identifier such as a name or an ID number is given to the system and it verifies that your biometric matches the biometrics stored under your name. The second mode of operation is identification, or one-to-many mode. Under this scenario, the biometric system compares the presented biometric to the entire database looking for a match. Though these two modes of operation have very different methodologies, their performance is measured in the same way.

In either of these modes, the result of the matching algorithm is a match score, T. The match score is a measure of distance between the stored template and the presented biometric. Thus, low scores indicate that the stored and presented biometrics are similar. A comparison is then made between the match score T and a threshold, τ. If $T \le \tau$, then the system decides that a match has been made and permits access. This is called an accept. If $T > \tau$, then the system decides that a match has not been made and denies access. This is termed a reject. The statistical aspects of biometric device performance focus on the rate at which errors are made in this process. These errors are akin to the Type I and Type II errors that are encountered in hypothesis testing. A Type I error is a false reject and a Type II error is a false accept.

Error Rates

To make this discussion more precise, consider the population of match scores for all attempts by genuine users, and let $f_{gen}(x)$ represent the density of this distribution. Similarly, consider the population of match scores for all attempts by imposters, and let $h_{imp}(y)$ be the density for this distribution. Then the false rejection rate (FRR) is the probability that T is greater than τ, given that T comes from the distribution of genuine user scores. The false acceptance rate (FAR) is the probability that T is less than τ, given that the T comes from the distribution of imposter scores. Symbolically,

$$FRR = P(T > \tau \mid T \in \text{Genuine}) = \int_{\tau}^{\infty} f_{gen}(x)dx$$

and

$$FAR = P(T \leq \tau \mid T \in \text{Imposter}) = \int_{-\infty}^{\tau} h_{imp}(y)dy.$$

The threshold τ can be set so that we have some control over the values that the FAR and FRR will take. However, note that as τ increases FRR will decrease and FAR will increase. Likewise, as τ decreases FRR will increase and FAR will decrease. In a practical setting, we are often interested in estimating the FAR and FRR for a particular biometric device. Given τ and samples from both genuine users and imposters, we can create estimates for the FRR and FAR, in the following way:

$$\bar{p}_{FRR} = \frac{\#(T > \tau \mid T \in \text{Genuine})}{\#\text{Genuine}}$$

and

$$\bar{p}_{FAR} = \frac{\#(T \leq \tau \mid T \in \text{Imposter})}{\#\text{Imposter}}$$

where $\bar{p}_{FRR}$ is an estimator of the FRR, $\bar{p}_{FAR}$ is an estimator of the FAR, #Genuine is the total number of genuine user scores and #Imposter is the total number of imposter scores. Thus the estimated FRR is the percentage of genuine user scores that fall above the threshold τ. Likewise, the estimated FAR is the percent of imposter scores that fall below the threshold τ.

ROC Curves

As mentioned above, for any given τ, we can estimate FAR and FRR. By varying τ, we can get different values of FAR and FRR. Plotting the different values that FAR and FRR take produces a function called a Receiver Operating Characteristic (ROC) curve. ROC curves are frequently used in engineering applications. The ROC curve is a concise graphical summary of the performance of the biometric device. Figure 3 gives an example of an ROC curve. Note that the ROC curve must always begin and end with the points (0,1) and (1,0). This can be seen by letting $\tau = \infty$ and $\tau = -\infty$, respectively.

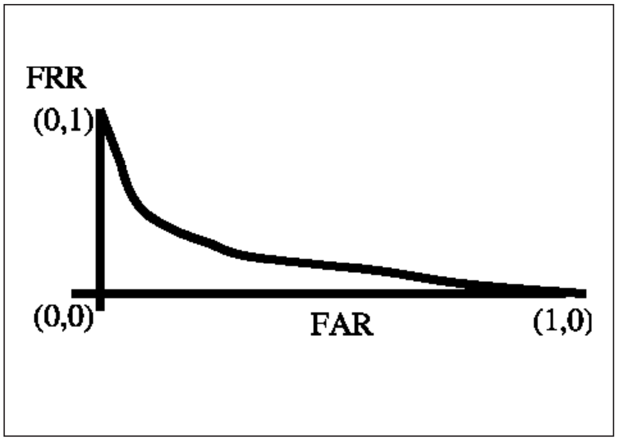


Figure 3. Example of a Receiver Operating Characteristic Curve

One measure of overall performance that is commonly used for biometric devices is the equal error rate (EER). This is the value that the ROC curve takes when it passes through a line of slope 1. That is, it is the point where the FAR is equal to the FRR. The EER is a single measure to assess the overall performance of a biometric device, but biometric devices are rarely set at τ = EER. Rather, τ is chosen based upon the conditions under which the device will be used. If security is the most important consideration, then τ will be chosen to give a low FAR. If the number of people moving quickly through the system is an important consideration, then τ may be set to give a low FRR.

Confidence Intervals for FAR and FRR

To statistically assess the performance of a biometric device, we would like to be able to create confidence intervals for FRR and FAR. To do this, we need to describe the usual techniques for testing a biometric device. For simplicity suppose that we are estimating the FRR. Estimation of the FAR would follow by a similar process. Traditionally, what is done is that M individuals are tested for each of n_i attempts, where $i = 1, 2, 3, \ldots, M$. The number of rejects from the

$$\sum_{i=1}^{M} n_i$$

attempts is then recorded. If we let X be the number of false rejects from a fixed number of attempts, it would seem that X might be a Binomial random variable. However, consider the conditions necessary for a Binomial experiment:

i. There must be a fixed number of trials, n.

ii. Each trial must result in one of two possible outcomes.

iii. Probability of success, p, must be constant for all trials.

iv. Each trial must be statistically independent of the others.

If we let $n = \Sigma n_i$, then condition *i* is met. Condition *ii* is clearly met since the biometric device decides to either accept or reject. It is known that each individual has his or her own probability of success and that these probabilities are different from individual to individual, so condition *iii* is not met when $M > 1$. However, if $M = 1$ and the trials are independent (condition *iv*) then X is a Binomial random variable. Recall that this means that we would only be testing one person. Thus, the conditions under which the Binomial distribution applies are severely limiting.

An Alternative Model

The biometrics community recognizes that the Binomial distribution cannot be used to create confidence intervals for FAR and FRR except under the restrictive conditions mentioned above (U.K. Biometrics Working Group, 2000). At present there is no accepted methodology for assessing the variability in estimated FAR's and FRR's. One alternative to the Binomial that has been proposed for assessing biometric performance is the Beta-binomial distribution (Schuckers, 2001). This Beta-binomial model assumes that each individual *i* has his or her own probability of an error, p_i. Let X_i be the number of failures from n_i attempts. Further, let p_i come from a population of probabilities that can be described by a Beta distribution. Then we have the following model:

$$X_i \mid p_i, n_i \sim \text{Binomial}(n_i, p_i) \quad \text{for } i = 1,2,3,\ldots, M.$$

$$p_i \mid \alpha, \beta \sim \text{Beta}(\alpha, \beta).$$

Then,

$$\pi = E[p_i] = \frac{\alpha}{\alpha + \beta}$$

and

$$Var[p_i] = \pi(1 - \pi)\frac{1}{(\alpha + \beta + 1)}.$$

This is a hierarchical model, as described in Stangl (2001). Our interest is now on π, the mean error rate for the population. We can treat the p_i's as nuisance parameters and integrate them out:

$$f(X_i \mid \alpha, \beta, n_i) = \int f(X_i \mid p_i, n_i) f(p_i \mid \alpha, \beta) dp_i.$$

Then,

$$X_i \mid \alpha, \beta, n_i \sim \text{Beta-binomial}(n_i, \alpha, \beta)$$

with

$$E[X_i] = n_i \pi \text{ and } Var[X_i] = n_i \pi (1 - \pi)\frac{(\alpha + \beta + n_i)}{(\alpha + \beta + 1)}.$$

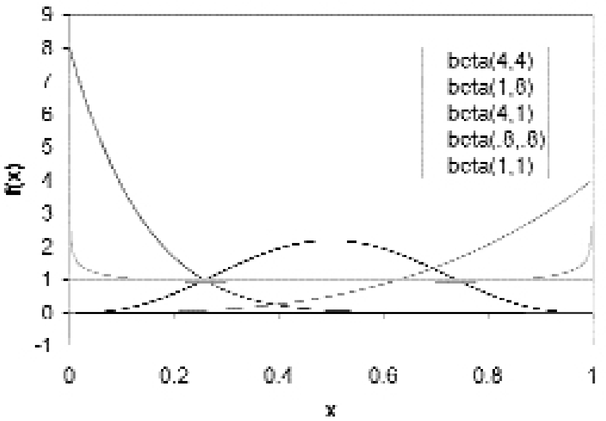


Figure 4. Beta distributions.

Numerical methods, such as maximum likelihood estimation, are used to estimate π and to estimate the variability in π. See, for example, Silvey (1975) for a discussion of maximum likelihood methods. Under regularity conditions, these methods then allow for the creation of a confidence interval for π.

The Beta distribution is a very flexible family of distributions for modeling the population of probabilities. The Beta distribution works well if the population of p_i's is unimodal, has a J-shape, or a reverse J-shape. (See Figure 4 for some of the various shapes that a Beta distribution can take.) One of the difficulties in using the Beta-binomial distribution is that there is a body of literature in the biometrics community asserting that the population of p_i's, particularly for FRR, may be bimodal. For FRR, the two modes are made up of subpopulations known as ‘goats' and ‘sheep.' Sheep are those individuals who have fairly low variability in their biometrics, whereas goats have a much higher level of variability in their biometrics. Hence, it seems likely that these two groups would constitute separate populations with different mean rates of a false reject. These populations were first acknowledged in a paper on speaker recognition that is affectionately called “Doddington's Zoo” (Doddington et al., 1998). Under these conditions, the Beta-binomial is not an appropriate distribution. One possibility is to model the population as a mixture of two Beta-binomial distributions (Everitt and Hand, 1981).

Summary

In the preceding sections, I have introduced some of the statistical aspects of assessing the performance of biometric identification devices. As the need for security increases it is likely that the use of these devices will expand. Many of these are already in use. For example, Disney World now uses a hand geometry system. In a post-September 11 world, biometric identification devices are likely to become more prevalent. Statistics will continue to play a large role in this development.

There are several statistical issues central to biometric identification devices. Here, I have focused on the matching performance of these devices which is critical to their acceptance and their viability. Each time an individual presents his or her biometric, it is either accepted or rejected by the system. The decision to accept or reject is based upon a matching algorithm that produces a match score. If the score is below some threshold, the attempt is accepted. Similarly if that threshold is exceeded then the attempt is rejected. Errors in this process are classified as false accepts and false rejects. Estimating and making confidence intervals for these rates is one of the most important issues in the biometrics community. Related to this is the pressing need to determine the sample sizes needed in order to create confidence intervals of a certain size. These and other statistical issues ensure that statisticians will have a substantial impact on the future of biometric identification devices.

References

Doddington, G., Liggett, W., Martin, A., Przybocki, M., and Reynolds, D. (1998), "Sheep, Goats, Lambs and Wolves: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation," in *Proceedings of 5th International Conference of Spoken Language Processing*, ICSLP 98, Sydney, Australia. Paper 608 on CD-ROM.

Everitt, B.S. and Hand, D.J. (1981), *Finite Mixture Distributions*, London: Chapman and Hall.

Jain, A., Bolle, R., and Pankanti, S. (eds.) (1998), *Biometrics: Personal Identification in Networked Society*. Boston: Kluwer Academic Publishers.

Johnson, N. L. and Kotz, S. (1970), *Continuous Univariate Distributions* 2, New York: John Wiley and Sons.

Schuckers, M. E. (2001), "Using the Beta-binomial Distribution to Assess the Performance of a Biometric Device," submitted to *International Journal of Image and Graphics*.

Silvey, S. D. (1975), *Statistical Inference*, New York: Halsted Press.

Stangl, D. (2001), "A Primer on Hierarchical Models," *STATS: The Magazine for Students of Statistics*, 32, 3–9.

U.K. Biometrics Working Group (2000), "Best Practices in Testing and Reporting Performance of Biometric Devices," on the web at www.cesg.gov.uk/biometrics.

Wayman, J. L. (1998), "Technical Testing and Evaluation of Biometric Identification Devices," in *Biometrics: Personal Identification in Networked Society*, eds. Jain et al., Boston: Kluwer Academic

A Look at Raking for Weight Adjustment

Willard C. Losinger

The National Animal Health Monitoring System (NAHMS) is a relatively new program of the United States Department of Agriculture (USDA) (Hueston, 1990). The first NAHMS national study was the 1990 National Swine Survey, which provided information on farm biosecurity practices, facility characteristics, swine diseases, and routine preventive/treatment practices (USDA, 1992). Since then, we have generally done two national studies per year, revisiting a livestock species every five years. The 1995 National Swine Study concentrated on health management practices in the grower/finisher phase of production (Losinger et al., 1998). The NAHMS Swine 2000 study was the third NAHMS national study of swine producers, and provided not only new information on swine diseases and management, but served as a basis for profiling changes in the swine industry (based on information from the previous two NAHMS surveys). More information on the NAHMS is available at the web site *http://www.aphis.usda.gov/vs/ceah/cahm/*.

In many U.S. livestock industries, the population of animals tends to be highly concentrated on a small percentage of the farms (Losinger, 1997). In the U.S. swine industry, in particular, increasing returns to scale have been associated with a rapidly increasing concentration of pigs onto fewer, larger operations (Losinger et al., 1999). Some producers have farrow-to-finish operations, raising pigs from birth until they are ready for slaughter. Some producers specialize in the farrowing phase of production, sending pigs (after they are ready to leave their mothers) to other farms that specialize in fattening pigs for market. In selecting participants for NAHMS national studies, there is a certain trade-off between representing animals and representing livestock producers. To select farms for participation in NAHMS national studies, farms are generally grouped into strata based on farm-size (i.e., number of animals on the farm) within states. Larger farms (which account for the majority of animals) are sampled at a higher rate than smaller farms (which are more numerous, but have a much smaller fraction of the animals) (Losinger et al., 2000).

Initial Sampling Weights

From the farms that participate in a survey, we generate estimates that apply to all of the farms and to all of the animals in the participating states. Since farms have different probabilities of selection based on their size, we can't just take the average value from among the respondents and say that this is what average farms do or that this is how average pigs are managed. Generally, each small farm in your sample represents a lot more farms than each large farm in the sample and so must be assigned a sample weight equal to the number of farms in the population that a farm in your sample represents for estimation purposes. Initially, this is just the inverse of the sampling fraction within each stratum. For example, if your sampling rate for a particular stratum is one in ten, then each sampled farm in this stratum represents a total of ten farms in the population (itself, plus nine other farms). If your sampling rate in another stratum is one out of two, then each sampled farm in this stratum represents two farms in the population. Since large farms have a higher sampling rate than small farms, large farms receive lower sample weights than small farms.

As a simple hypothetical example, suppose that Stratum A consists of 20 small farms and that two of them are sampled (call them A1 and A2), for a sampling rate of one-in-ten. Suppose further that one of these two farms has a sick pig. Finally, suppose that Stratum B consists of two large farms, that one farm is sampled (call it B1), and that it does have a sick pig. Thus, within the sample, two out of three farms (66.7%) have sick pigs. However, each farm sampled in Stratum A represents ten farms in the population, and the one farm sampled from Stratum B represents two farms. Therefore, our adjusted estimate is that 12 farms (the ten represented by farm A1 and the two represented by farm B1) of the 22 in the population (54.5%) have sick

Willard Losinger (wlosinger1@netscape.net) *is a statistician with the US Department of Energy's New Brunswick Laboratory, and has been a federal statistician for 12 years. Other statistical experience includes a couple of years in the pharmaceutical industry, and two years as a US Peace Corps volunteer, where he served as Chief of the Survey Division of Tonga's Statistics Department.*

pigs. Similarly, we could estimate the number of farms employing this or that practice (for example, using a particular vaccine or feed additive).

Animal-Level Weights

In addition to learning about practices of farms, we also want to be able to make estimates about the individual animals. Examples include the number of sick pigs, the number of pigs that receive a particular vaccine or feed additive, and death rates. Animal-level weights are created by multiplying the farm's weight by the number of animals that the farmer reported, in order to estimate the number of animals in the population that the animals on a participating farm are representing. Continuing the above example, suppose that farm A1 has ten pigs, one of which is sick, that farm A2 has twenty pigs, none of which is sick, and that farm B1 has 100 pigs, 90 of which are sick. The two farms sampled in Stratum A would then have pig-level weights of 100 and 200, respectively (10 pigs times the ten farms represented by A1, 20 pigs times the ten farms represented by A2). The sample farm from Stratum B would have a pig-level weight of 200 (100 pigs times the two farms represented by B1). Thus, we would estimate a total of 500 pigs in the population. Applying the pig-level weights to the percentage of sick pigs on each sampled farm, we would estimate that 38% (100 × .1 for farm A1, 200 × 0 for farm A2, and 200 × .9 for farm B1, divided by the total estimate of 500 pigs) are sick. If we had merely estimated the percentage of sick pigs based on the animals sampled, we would have calculated (1 + 0 + 90)/(10 + 20 + 100), which is 91/130, or about 70%. This is much different from the previous estimate, indicating that pig-level adjustments are necessary to obtain a more accurate view of the animal populations.

Response Adjustment

In fact, even these sample weight adjustments are not sufficient because, when we implement a survey, we invariably find that not every producer is still in business or willing to participate in the survey when visited by the enumerator. Therefore, weights need to be transferred from sampled farms that would have been eligible but refused to participate in the survey, to farms that participated. This is accomplished by creating a response adjustment equal to the sum of weights of eligible farms divided by the sum of weights of respondents, generally either within the original sampling strata or within poststrata (i.e., strata defined after the data have been collected) (Losinger et al., 1998). Typically, if a stratum has fewer than 20 respondents, then farms within this stratum are combined with farms from another stratum (in a similar region or farm-size group) to form a new poststratum. Weights of nonrespondents are set to zero, and weights of respondents are multiplied by the response adjustment. Sometimes, low participation rates in a few parts of the country can have a greater impact on the resulting weights than the initial sampling rates did.

In our hypothetical example, suppose that the farm with ten pigs (A1) had chosen not to participate. We would transfer its sample weight to the participating farm with 20 pigs, giving the participating farm with 20 pigs (A2) an adjusted weight of 20. Our estimate would then be that two (9.1%) of the 22 farms in the population have sick pigs, and that 30% (20 x 0 + 90 x 2 = 180 divided by 20 x 20 + 100 x 2 = 600) of pigs are sick.

This small example illustrates that the sample weights can have a dramatic effect on the population estimates. However, we do have more information. The National Agricultural Statistics Service (NASS) publishes the number of operations and the number of pigs by state and size groups. We can use these inventory numbers as a way to verify our weights, by seeing if the resulting numbers “match.” If they do not, we can make further adjustments to our weights.

Inventory Adjustment

For NAHMS surveys, the traditional inventory-adjustment method has been to force inventory estimates to match the NASS numbers by state and size groups. First, a NAHMS inventory estimate was computed (for each state-by-size group cell) by summing the animal-level weights (within each state-by-size group cell). Then, each participant's weight was multiplied by the ratio of the NASS published inventory to the NAHMS inventory estimate by state and size group (Losinger et al., 1998). Then, it was necessary to examine the distribution of the adjusted weights. At this stage in particular, we had to be extremely wary of the impacts that the inventory adjustments were having on both farm-level and animal-level estimates. Frequently, a small number of respondents ended up with extremely large weights (after the inventory adjustment) compared to the majority of the respondents in the sample. If we didn't do anything about it, then the population estimates would have been heavily dependent on the responses given by the respondents with large weights. Generally, respondent weights exceeding a particular value were truncated to a maximum value, and their excess weight was redistributed among all respondents within their poststratum (Losinger et al., 1998). Basically, within each poststratum, each participant's inventory-adjusted weight was multiplied by the ratio of the sum of the untruncated inventory-adjusted weights to the sum of the truncated inventory-adjusted weights. Thus, even the truncated inventory-adjusted weights received the adjustment

Outline of Steps in the Weight Creation Process for National Animal Health Monitoring System (NAHMS) data.

1. **Initial sample weight:** the inverse of the sampling fraction within each sampling stratum [initial sampling weight = 1/sampling rate]
2. **Response adjustment:** transfer weights from eligible non-respondents (i.e., farms that were selected in the sample and that would have been eligible to participate in the survey, but that refused or somehow failed to participate in the survey) to farms that participated in the survey [response adjustment = (sum of weights of eligible farms) / (sum of weights of poststratum)].
3. **Inventory adjustment**: force estimates of inventory (i.e., numbers of animals) to match figures published by the National Agricultural Statistics Service (NASS). We evaluated our traditional method and raking [inventory adjustment = (NASS estimate)/(NAHMS estimate) within each cell or raking between marginals].

The traditional method of performing the inventory adjustment had the following steps:

a. Compute NAHMS inventory estimates by cells based on state and size groups (using weights from Step 2 and inventory figures provided by participants).
b. Multiply each participant's weight by the ratio of NASS published inventory to NAHMS inventory estimates (within each cell).
c. Smooth excessively large weights by truncating to a maximum value, and redistributing excess weights to participants within the same poststratum.

Raking for inventory adjustment had the following steps:

a. Compute NAHMS inventory estimates by state (using weights from Step 2 and inventory figures provided by participants).
b. Multiply each participant's weight by the ratio of NASS published inventory to NAHMS inventory estimates (by state).
c. Use weights from step b. to compute inventory estimates by size group.
d. Multiply each participant's weight (from step b.) by the ratio of NASS published inventory to the new NAHMS inventory estimates (by size group).

and ended up with weights greater than the truncation limit. We referred to this procedure as "smoothing."

To perform the inventory adjustment for the Swine 2000 study, we decided to try an alternative weight adjustment method called "raking" (Deming and Stephan, 1940). We had 2,499 participating farms with 100 or more pigs in the 17 states included in the study, with a total of 8,024,131 pigs. Table 1 shows the marginal totals for the numbers of pigs and operations that we sought to represent with the sample (i.e., the population from which we were sampling). NASS provides the total number of operations, which often have multiple farm sites. We computed estimates for farm sites rather than for operations, and we use the terms "participant" and "respondent" to refer to a participating farm site.

Table 2 provides some summary statistics based on the traditional method of adjusting weights for inventory within each of the 85 state-by-size group breakouts. At this stage, this was the result of multiplying each participant's weight by the ratio of the NASS published inventory to the NAHMS inventory estimate within each of the 85 state-by-size groups. Within each state-by-size group, the weighted total numbers of pigs matched the NASS numbers exactly (no surprise there), but the weighted number of farms was often off by quite a bit. Three farms (within a group that had relatively low participation) wound up with extremely huge weights (more than 17,000—all other weights were less than 3,000). The next phase in the traditional weight adjustment method would be to "smooth" the more outlandish weights by truncating their weights to some maximum reasonable number, and then reallocating their excess weight to farms in a similar part of the country and in a similar size group. We always had to pay very close attention to what was happening with the weights and resulting estimates, and to make judgments about various tradeoffs involved in choosing one cutoff versus another. Then, if estimates of the numbers of farms were way off, some arbitrary compromises had to be made between misrepresenting the number of farms and misrepresenting the number of animals.

With raking, we do not adjust weights for cells individually. Instead, first we adjust all participant weights to match one set of marginal totals, and then the other set of marginal totals. Then, we go back and do it again ("raking" back and forth) until we achieve convergence (i.e., very little change from one iteration to the next).

Raking Analysis

In this case, first we adjusted all weights so that the weighted sum of pigs would match the NASS-published estimates by state (across all five size groups). Then, we adjusted the weights so that the

Table 1. Total number of operations and pigs (for operations with 100 or more pigs) on January 1, 2000, in the 17 states included in the NAHMS Swine 2000 Study.

State	Number of Operations	Number of Pigs
Arkansas	420	676,200
Colorado	110	885,550
Illinois	4,440	4,108,500
Indiana	3,800	3,250,500
Iowa	12,100	15,453,500
Kansas	1,000	1,376,100
Michigan	900	990,000
Minnesota	5,300	5,643,000
Missouri	2,100	3,004,250
Nebraska	3,550	2,905,750
North Carolina	1,800	9,552,000
Ohio	2,500	1,330,000
Oklahoma	300	2,255,650
Pennsylvania	950	1,013,250
South Dakota	1,800	1,205,400
Texas	100	809,100
Wisconsin	1,100	533,600
Size Group		
100–499 pigs	20,490	4,314,150
500–999 pigs	8,820	5,092,700
1,000–1,999 pigs	6,205	7,206,750
2,000–4,999 pigs	4,815	12,591,850
> 5,000 pigs	1,900	25,786,900
Total	42,230	54,992,350

Source: *http://www.usda.gov/nass*

weighted sum of pigs would match the NASS-published estimates by size group (across all seventeen states). Then, we went back and adjusted again by state, and then by size group, and so on, until we did a total of ten adjustments. Convergence happened pretty quickly (Table 3). We could have easily stopped prior to ten iterations, but went on to ten to see what would happen. The variability in the resulting weights was much less than with the traditional weight adjustment method; indeed, the maximum weight was only 232 instead of over 26,000! The raking method corrected for the number of pigs and had nothing to do with the number of farms during the process. However, an examination of the state-by-size group weighted number of farms and pigs showed that we were reasonably close to the NASS-reported numbers (much better than what we had usually experienced with our traditional weight adjustment methods). Moreover, we didn't have any extreme weights to smooth and fuss with at all.

One pitfall with raking is that you have to pay attention to falling weights—especially weights that fall below one. You know that a farm in your sample represents at least itself in the population—it cannot represent any less than itself. Therefore, one weight that fell slightly below one was rounded up to one. We used the resulting weights to estimate numerous survey parameters related to swine health and management, and to provide information that will ultimately be used to improve swine production practices in the United

Table 2. Summary of results using the traditional inventory weight adjustment method of multiplying each farm's unadjusted weight by the ratio of the number of pigs (reported by NASS) to the sum of the weighted number of pigs (i.e., the sum of the products of each farm's unadjusted weights and number of pigs) within each of the 85 state-by-size group poststrata.

	Mean	Minimum	Maximum
Unadjusted Weight	10.19	1.10	124.19
Adjusted Weight	171.44	1.71	26,404.84
Adjustment Factor	16.64	0.74	777.60

Table 3. Summary of results using "raking" to adjust weights to match marginal totals of numbers of pigs by state and size group.

	Mean	Minimum	Maximum
Unadjusted Weight	10.19	1.10	124.19
Weight after:			
First Adjustment	14.08	1.21	162.38
Second Adjustment	16.05	1.05	232.19
Third Adjustment	16.13	1.00	232.46
Fourth Adjustment	16.07	0.99	232.09
Fifth Adjustment	16.10	0.99	232.61
Sixth Adjustment	16.07	0.98	232.22
Seventh Adjustment	16.08	0.98	232.34
Eighth Adjustment	16.07	0.98	232.25
Ninth Adjustment	16.08	0.98	232.28
Tenth Adjustment	16.07	0.98	232.26
Weight Adjustment Factors:			
First Iteration	1.37	0.59	2.08
Second Iteration	1.03	0.80	1.43
Third Iteration	1.00	0.93	1.07
Fourth Iteration	1.00	0.99	1.01
Fifth Iteration	1.00	0.99	1.01
Sixth Iteration	1.00	1.00	1.00
Seventh Iteration	1.00	1.00	1.00
Eight Iteration	1.00	1.00	1.00
Ninth Iteration	1.00	1.00	1.00
Tenth Iteration	1.00	1.00	1.00

States.

With raking, convergence is not necessarily guaranteed. It is possible that a given set of circumstances would lead to divergence (i.e., a bouncing back and forth between two or more points) rather than convergence. Thus, the statistician will have to pay attention to what is happening from one iteration to the next.

Conclusion

Raking weight adjustments (back and forth across marginal totals) demonstrated superiority to our traditional method (of adjusting weights to match totals within individual cells) for performing inventory-adjustments on our weights. Estimates of numbers of farms were better with raking than with our traditional method (which were often way off and required compromises between accurate estimates of numbers of farms and numbers of animals), and we didn't end up with enormous weights (due to low participation in a few cells) that had to be smoothed. Sometimes, our confidence with the results of smoothing was not high because of errors in judgment that might have occurred in deciding where to truncate and smooth. With raking, low participation in any particular cell is generally not a problem, as long as there are enough participants to contribute to the marginal total. With raking, you do have to watch out for falling weights, and you certainly don't want to allow any participant's final weight to remain below one. Whatever weight adjustment method is used, statisticians in this line of work do need to be aware of potential pitfalls, do need to examine distributions of weights at each stage of adjustment, and do need to pay close attention to impacts on estimates. While weight adjustment is a highly specialized branch of survey statistics, “raking” has been used for weight adjustment for over 60 years, and it would be fascinating to see whether raking might have some application to other branches of the statistics profession.

References

Deming, W.E., and Stephan, F.F. (1940), “On a Least Square Adjustment of a Sampled Frequency when the Expected Marginal Totals are Known,” *Annals of Mathematical Statistics*, 11, 427–444.

Hueston, W.D. (1990), “The National Animal Health Monitoring System: Addressing Animal Health Information Needs in the USA,” *Preventive Veterinary Medicine*, 31, 1–14.

Losinger, W.C. (1997), “The Lorenz Curve Applied to Livestock Populations,” *Chance*, 10(2), 19–22.

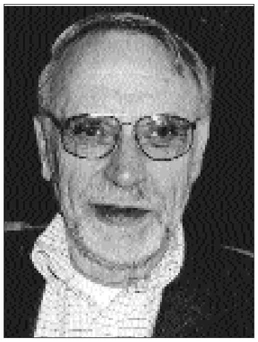
Losinger, W.C., Bush E.J., Hill, G.W., Smith, M.A., Garber, L.P., Rodriguez, J.M., and Kane, G. (1998), “Design and Implementation of the National Animal Health Monitoring System 1995 National Swine Stud,” *Preventive Veterinary Medicine*, 34, 147–159.

Losinger, W.C., Dalsted, N.L., Sampath, R.K., and Salman, M.D. (1999), “Returns to Scale in the Production of Finisher Pigs in the United States,” *Investigación Agraria: Producción y Sanidad Animales*,14, 71–84.

Losinger, W.C., Traub-Dargatz, J.L., Sampath, R.K., and Morley, P.S. (2000), “Operation-Management Factors Associated with Early-Postnatal Mortality of US Foals,” *Preventive Veterinary Medicine*, 47, 157–175.

United States Department of Agriculture. (1992), National Swine Survey Technical Report. Document N106.692. USDA:APHIS:VS, CEAH, 555 South Howes St., Fort Collins, CO 80521.

Infant Handling by Female Baboons: A Sensitivity Analysis



Clifford E. Lunneborg

In the Spring 2001 issue of *STATS*, statistician Thomas Moore and anthropologist Vicki Bentley-Condit described the use of permutation tests to evaluate the hypothesis that adult female baboons tend to "handle the infants of females who are ranked the same as or lower than themselves." Data for the analyses were provided by Bentley-Condit, who had observed a troop of Kenyan baboons, including 23 females and 11 infants, over an 11-month period in 1991–1992. In this paper an alternative analysis is proposed, one exploring the sensitivity of their findings to the inclusion of individual animals.

Moore and Bentley-Condit (2001) are to be congratulated for providing a careful development of permutation test methodology and for illustrating a novel application. Permutation (or randomization) tests deserve much greater emphasis than they are presently given, both in the curriculum and in practice (Ludbrook and Dudley, 1998). The Moore and Bentley-Condit article certainly will aid in developing more interest for these techniques on the part of students and instructors.

A caution was sounded by Moore and Bentley-Condit on the interpretation of their permutation test results. Rightly, they noted that statistical significance was suspect, given that the randomness required by the test was not present. That is, the baboons observed were neither randomly sampled from some larger population, nor, of course, were the females randomly assigned ranks, nor—perhaps more relevant to the hypothesis evaluated here—were the exposures of females to infants randomly controlled by the researcher. Having cautioned the reader, a permutation test p-value of 0.0166 is reported for a test statistic, $S = 472$, which counts the excess of the number of handlings of infants of same or lower rank females over the number of handlings of infants of higher rank females.

This certainly is not the first occasion on which a frequentist p-value has been assigned to an outcome in the absence of any randomness in the design of the study to which that p-value can be linked. The practice is endemic. Rarely is the reader warned, as Moore and Bentley-Condit have done, that the p-value is suspect. What, indeed, is to be made of a finding of statistical significance here, if one interprets the p-value of 0.0166 as significantly small? We cannot conclude anything about the infant-handling behavior of female baboons not observed in this study. A population inference of that kind would require random sampling of baboons. Nor can we conclude that the relative ranks of a female and the mother of an infant about whom a handling choice was made "caused" the choice. A causal inference would be possible had an experimenter randomly controlled the exposures of females to infants or of infants to females. Absent these traditional inferences, what can we conclude from an "observational" p-value? Some would argue that a small p-value is testimony to the strength of an outcome, if only in a particular nonrandom sample. I find that uncompelling. First, the p-value is no more than a monotonic transformation of the magnitude of the test statistic, e.g., in this study a larger S would have resulted in a smaller p-value. The p-value tells us nothing new. Second, and more important by far, the p-value is highly subject to over-interpretation. Readers are far too likely to attach traditional interpretations to the p-value than to take the nonrandom nature of the study into account.

Observational studies are carried out. Many of these studies are important scientifically. In the field, Bentley-Condit could not have carried out a randomized experiment. Those field observations have undoubted value and deserve to be reported in a manner that conveys their "significance." Significance testing and related techniques such as the estimation of a confidence interval help the researcher—and readers of that researcher's report—understand the magnitude of an effect observed in a random sample or randomized study. What might similarly aid researchers to understand the "significance" of observational findings?

The finding of an observational study often takes the form of a summary statistic, one that aggregates observations across a number of sources—clinics, lab-

Clifford E. Lunneborg (cliff@ms.washington.edu) *is Emeritus Professor, Statistics and Psychology, University of Washington, Seattle, Department of Statistics, Box 354322, University of Washington, Seattle, WA 98195-4322.*

Table 1. Frequency of Interactions Between Infants (rows) and Adult Females (columns)

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

oratories, political subdivisions, schools, time periods, etc.—each source contributing multiple observations. That summary statistic, almost certainly, is interpreted as holding across sources. Consequently, I have argued (Lunneborg, 2000) that it is important to verify that the summary statistic in a nonrandomized study fairly reflects the whole of the data and, in particular, that it is not unduly influenced by the inclusion in the study of one particular source of observations. A simple technique for doing this is to form subsamples of the data, leaving out one source of observations at a time, and recompute the summary statistic on each subsample.

The data set analyzed by Moore and Bentley-Condit lends itself to such a subsample analysis. Each baboon, whether female or infant, supplied multiple "handling" observations. Is the overall conclusion (that females in this troop tend to handle infants of the same and lower maternal rank) strongly dependent on the behavior of a particular female or infant?

Table 1 is repeated from Moore and Bentley-Condit. Rows correspond to infants and columns to female handlers. Cell entries give the number of interactions observed between an infant and a handler. Where the handler is the infant's mother, the cell entry is set to zero.

Rows and columns of the table have been arranged to correspond to the social ranks, 1: High, 2: Mid, or 3: Low, of handlers and of the infants' mothers. The sensitivity analysis I propose carrying out may be particularly useful where the observational data, through the necessary lack of a design, are unbalanced. Here, for example, only one infant (KG) has a High maternal rank and some adult females (e.g., ST and WK) are observed interacting with infants much more frequently than are other females (e.g., LS and LY).

Moore and Bentley-Condit overlaid a 3 x 3 grid on the cells of Table 1 to group together adult-infant interactions by ranks of handler and infant's mother:

a : Infant 1, Handler 1 ***b*** : Infant 1, Handler 2 ***c*** : Infant 1, Handler 3
d : Infant 2, Handler 1 ***e*** : Infant 2, Handler 2 ***f*** : Infant 2, Handler
3
g : Infant 3, Handler 1 ***h*** : Infant 3, Handler 2 ***i*** : Infant 3, Handler 3

They then counted the number of interactions within each of the nine cells of this grid and combined these counts in a summary statistic reflecting their research hypothesis:

$$S = a - b - c + d + e - f + g + h + i.$$

A positive value of *S* signals an excess of interactions of adult females with infants of the same or lower maternal rank over those with infants of higher maternal rank. The value of their comparison statistic was *S* = 472 in the direction of their research hypothesis.

How strongly is the magnitude of *S* influenced by individual infants or handlers? In a subsampling sensitivity analysis I would withdraw one animal at a time and recompute this statistic. However, the magnitude of *S* is very much dependent on the total number of interactions. Removing any animal reduces the number of interactions and, consequently, could be expected to reduce the size of *S*. That would cloud the interpretation of the subsampling analysis. As a result, I propose first to alter slightly the definition of the summary statistic. I will divide the Moore and Bentley-Condit count by the total number of interactions in the table:

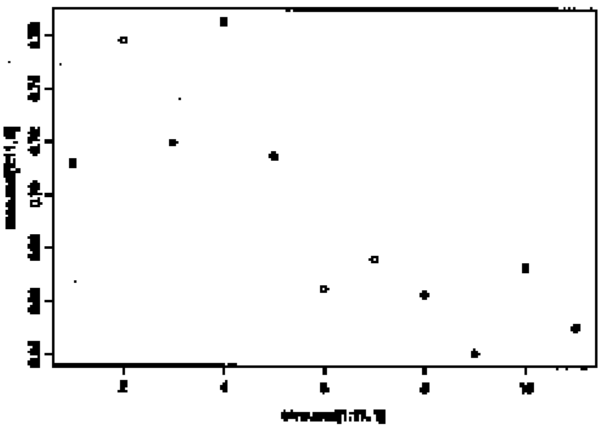


Figure 1. Values of Test Statistic S′, Omitting Data for One Infant in Turn[1,2]

1.Values of S′ are given along the y-axis.
2. x-axis identifies the omitted Table 1 row: 1 (KG) through 11 (PK).

$$a+b+c+d+e+f+g+h+i$$

$$-\frac{b+c+f}{a+b+c+d+e+f+g+h+i}$$

My S' is a difference in proportions: the proportion of interactions that are with same or lower maternal rank infants minus the proportion that are with higher maternal rank infants. This difference in proportions has the same permutation distribution as the difference in frequencies of interactions used by Moore and Bentley-Condit. That is, had they chosen to scale in the way I have, their results would have been unaffected. For an analysis based on all infants and females in this study (Table 1), the difference in proportions S' takes the value 0.694. In other words, roughly 85% of the interactions are concordant with the research hypothesis and 15% are discordant.

Figure 1 shows the difference in proportions leaving out one infant at a time. These differences, plotted on the *y*-axis, range closely and more or less evenly about 0.694, from 0.641 to 0.766.

This is the result to be expected if no single infant strongly determined the overall summary. If there were a strongly influential infant, the subsample differences would remain close to the total table difference, except when the influential infant's contribution was removed. That particular subsample difference would be markedly different from those for both the total and the other subsamples (Lunneborg, 2000).

Here, the biggest shifts away from the total value of S' are associated with the omission of infants 2 (HZ) or 4 (NK), the two infants most frequently handled. Neither of these shifts is large enough to change the conclusion about the influence of rank on infant handling.

A similar picture of no strong individual influence emerges in Figure 2 when, one at a time, an adult female's interactions are removed from the table. These differences in proportion are, if anything, more closely grouped around the overall result, ranging from 0.666 to 0.758. Withdrawal of females 10 (ST) or 11 (WK) appear to have the greatest impact on the value of S'. These females were involved in more infant handlings than their peers.

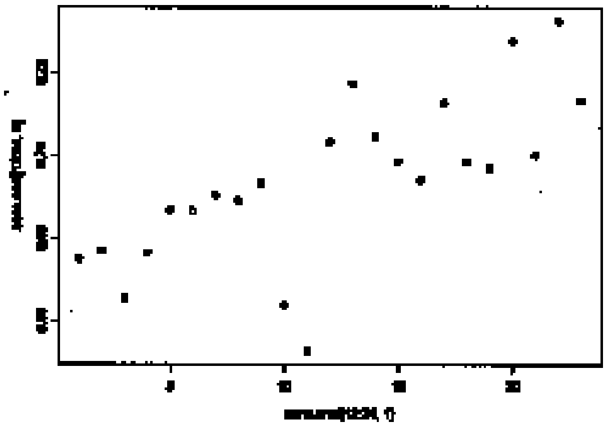


Figure 2. Values of the Test Statistic S', Omitting Data for One Adult Female in Turn[1,2]

1.Values of S'are given along the y-axis.
2. x-axis identifies the omitted Table 1 column: 1 (KM) through 23 (RS).

A difference on the order of 0.694 between the proportion of interactions that are concordant with the research hypothesis and the proportion that are discordant appears to provide a fair summary for this troop of baboons. That summary is not strongly biased by the observations made on particular adult females or on particular infants.

Although there is no evidence of a single infant or a female adult unduly influencing the overall summary, there is something of a pattern to Figures 1 and 2. Leaving out the infant of a High (or Mid) rank mother increases the value of S', while leaving out the infant of a Low rank mother decreases S'. Similarly, dropping a higher ranked female decreases S', while dropping a lower ranked female increases S'.

Certainly the positive value of S' characterizes infant-handling behavior in this troop of baboons fairly and thus supports the researchers' substantive hypothesis—without the need for any p-value. However, the exact value of the statistic would appear to be dependent on the distribution of ranks among the adult females in the troop, as well as the distribution of maternal ranks of the infants.

References

Ludbrook, J. and Dudley, H. (1998), "Why Permutation Tests are Superior to *t* and *F* tests in Biomedical Research," *The American Statistician*, 52, 127–132.

Lunneborg, C.E. (2000), *Data Analysis by Resampling: Concepts and Applications*, Pacific Grove, CA: Brooks-Cole.

Stat Song Sing-Along!

Larry Lesser

On the way to writing my 1994 statistics education dissertation, I completed all coursework required for a statistics Ph.D., worked as a statistician outside academia, and combined backgrounds in statistics, mathematics and education, not unlike Miller (2000). I recently also integrated my songwriting background when I published the first juried comprehensive articles (Lesser 2000; Lesser 2001a) on how music and song can be used in the teaching of mathematics and statistics. The latter article uses songs from many genres to provide vehicles for generating descriptive statistics, testing hypotheses, analyzing data, and analyzing statistical terms and themes in song lyrics.

Now grab a guitar (do statisticians who "play a mean guitar" use "X-barre chords"?) or sing these new lyrics over vocals-removed karaoke discs or even without any accompaniment:

I recently debuted (Lesser 2001b) "The Gambler," which may be sung to the tune of the Don Schlitz song of the same title (that yielded Kenny Rogers a #1 country hit and a TV miniseries!). My interest in finding creative ways to educate general audiences about the lottery dates back to a highly-publicized course I created on the psychology and probability underlying the then-new Texas Lottery (Elliot 1993; Lesser 1997).

‡The Gambler·
lyrics copyright 2001 Lawrence Mark Lesser;
reprinted by permission; all rights reserved

On a warm summer's evenin', on a train bound for nowhere,
I met up with a gambler — we were both too tired to sleep.
So he told me how he planned winnin' lottery prizes
'Til, as a math teacher, I just had to speak:

"Son, I've made a life out of readin' students' faces,
Checkin' comprehension by the way they held their eyes.
And I can see your blackboard is erased in some places—
Give me some peanuts and I'll give ya some advice.

First, your instant scratch-off tickets give 1 in 5 chances,
But that don't mean that 1 in 5 will win.
'Cause ev'ry ticket's sep'rate, like a new flip of a coin:
It has no mem'ry how your wallet's gotten thin!

And you track those weekly draws, you say ya got a system—
You call some numbers "hot," you deem others "due;"
But I insist, they each have the same chance—
If you're gonna play the game, boy, ya gotta know what's true!

CHORUS: You gotta know when you pick 'em, what's superstition,
And where strategy is there to be had,
Or you'll learn why lotteries seem like
Tax on folks who don't know much math!

Now all sets of numbers are equally unlikely,
More rare than death by lightning, still there's somethin' you should know:
If you should happen to win that big jackpot,
You'll win more money if you picked it all alone!

So avoid those numbers that more folks are playin':
Like sevens and birthdays and sequences, too.
'Til this song gets famous, you'll have the advantage—
Maybe you'll thank me with a share of your loot!"
(Repeat Chorus)

My fascination with the "birthday problem" (Lesser 1999) inspired this next ditty (Lesser 2001b). The lyric contrasts the often-confused events of "some people matching" and "someone matches with ME" and may be sung to the tune of Mildred J. Hill and

Lawrence (Larry) Lesser (lesserla@mail.armstrong.edu) *follows his mu's as an Associate Professor at Savannah's Armstrong Atlantic State University. Learn more about his academic and musical interests at www.math.armstrong.edu/ faculty/lesser/.*

Patty Smith Hill's “Happy Birthday to You”:

‡Birthday Song·
lyrics copyright 2001 Lawrence Mark Lesser;
reprinted by permission; all rights reserved

Happy birthday to you —
Bring another 22:
Then we'll have even chances
Of a match in this room.....
Or many more!

Happy birthday to me —
Bring another two-fifty-three:
Then I'll have even chances
Someone matches with ME.....
Or many more!

As the staff statistician for the Texas Legislative Council during its redistricting project in 1990–91, I utilized Census data and gained experience to appreciate some of the issues raised over the last several years in discussions about statistical adjustment for undercount. As Moore (1998, p. 10) relates, “On August 24, 1998, a federal court panel ruled that the use of statistical sampling for Congressional apportionment violates the Census Act. It is important to note that statistical and scientific issues played no role in the decision.” The lyric at the end of Lesser (2001a) concisely articulates the controversy and may be sung to the tune of John Denver's #1 hit “Annie's Song”:

‡Taking Leave of Our Census·
lyrics copyright 1998 Lawrence Mark Lesser;
reprinted by permission; all rights reserved

You fill out the Census
Once ev'ry decade.
It's quite a sample;
It tries to count all!
It can't help but miss some,
Some more than others—
But can we adjust it
And follow the law?

If that “census” pun wasn't too much, you may get your fill of puns in this statistician's “breakup” song, which I sing as a standard 12-bar blues, with the words in parentheses spoken (rather than sung) during the final 2 bars (measures) of each group of 12.

‡Statistician's BLUEs·
lyrics copyright 1994, 2001 Lawrence Mark Lesser;
reprinted by permission; all rights reserved

I've been mean-in' to tell ya 'bout my last co-relation,
I've been median to tell ya 'bout my last co-relation:
She wasn't from Haiti, but she was variation!
(unexplained and uncontrolled!)

I saw her with ANOVA man, and they were not discrete,
I saw her with ANOVA man, and they were not discrete—
I went proba-ballistic and let out a Pearson scream!
(those deviates!)

Told her, “If you're gamma data me, mu beta change your mode.
If you gamma data me, mu beta change your mode.
Chi-square you'll be inference, if you random that road!
(you'll be skewed!)”

Called up my dad: “Hi Pa! This is testing my heart!”
Yeah I told my dad: “Hypothesis testing my heart!”
He said, “What's your expectation? Ya met her at an X-bar.”
(“You're right, Dad! Simulator!”)

She was my significant other — significant at point-oh-three,
She was my significant other — significant at point-oh-three,
But alpha get her soon — as sample as can be!
(That'll Fisher! Time serious!)

References

Elliot, D. (1993, August 28), “Professor Seeks to Even Odds of Lottery,” *Austin American-Statesman*, B1, B3.

Lesser, L. (1997), “Exploring Lotteries with Excel.” *Spreadsheet User*, 4(2), 4–7.

Lesser, L. (1999), “Exploring the Birthday Problem with Spreadsheets,” *Mathematics Teacher*, 92(5), 407–411.

Lesser, L. (2000), “Sum of Songs: Making Mathematics Less Monotone!” *Mathematics Teacher*, 93(5), 372–377.

Lesser, L. (2001a), “Musical Means: Using Songs in Teaching Statistics,” *Teaching Statistics*, 23(3), 81–85.

Lesser, L. (2001b), “Formula for a Hit: Using Songs in Teaching Mathematics and Statistics,” invited closing keynote presentation, Lowcountry Mathematics and Science Hub Day, Beaufort, SC, October 2001.

Miller, J. B. (2000), “Student Voices: On Becoming a Statistics Educator,” *STATS: The Magazine for Students of Statistics*, 28, 28–29.

Moore, D. (1998), “Federal Courts Rules on Census Sampling” *Amstat News*, 257, 10.

AP Statistics

Some Thoughts about Degrees of Freedom

Gretchen Davis

When we hear "degrees of freedom," do we think of a popular folk song from the 60's, a self-help book for the recently divorced, a sailboat moored in the marina, the different ways that a given molecule can change its energy, or the number of independent components minus the number of estimated parameters?

Introduction

Students in AP Statistics may be perplexed when the concept of degrees of freedom is introduced as the denominator in the calculation of the sample variance. It may be helpful to link what students already know about dimension and coordinate geometry to what they are learning about estimating parameters using statistical samples. This approach was suggested in the first section of the classic four-part article written by Helen Walker and published in the *Journal of Educational Psychology* in 1940.

Dimension and Motion

Imagine a pesky fly moving freely in three-dimensional space. If we restrict its path to a tabletop, then we reduce its freedom of flight to a two-dimensional path. And if we confine the fly to walking on a wire, its freedom of movement is reduced to a one-dimensional path.

Locus and Coordinate Geometry

We consider the locus or path of ordered points (x, y) in the xy plane, where the point is free to move anywhere in two-dimensional space. When we add the restriction that the sum of its coordinates is seven $(x + y = 7)$, we reduce the locus to a line. If we have two simultaneous restrictions that the sum of the coordinates is seven and the difference is three $(x + y = 7$ and $x - y = 3)$, we reduce the locus to a single point (5,2), which is the intersection of the two lines.

Expanding the discussion to ordered triples (x, y, z) in three-dimensional space, the point is free to move anywhere in the space. When we add the restriction that the sum is seven $(x + y + z = 7)$, we are now on a two-dimensional plane within the space. If we have the two restrictions $(x + y + z = 7$ and $x + y - z = 3)$, the locus is reduced to a line. Finally, imposing the simultaneous constraints $(x + y + z = 7$ and $x + y - z = 3$ and $x - y - z = 1)$ reduces the possible common solutions to a single point (4, 1, 2).

Walker generalized the analogy to ordered n-tuples in n-dimensional space. We reduce the dimension of our locus by one, each time we impose a restriction. If r conditions or restrictions are imposed, our dimension is reduced from n to n-r.

Connections to Statistical Samples

Students should be familiar with dimension, locus, and coordinate geometry from their earlier studies. How do these concepts connect to statistics?

Specifically, how are the degrees of freedom related to why we divide by n-1 in calculating a sample variance or standard deviation? When we measure n items in a sample from a population of interest, the n items are free to take on any values and can be anywhere in the n-dimensional plane. Variance measures the dispersion of these values from the mean. To calculate the sample variance, we must first estimate the population mean with the sample mean. This imposes a restriction that the items in the sample can take any values as long as their sum divided by n equals the value of the sample mean. This restriction reduces the dimension of the space in which the items can vary from n to $n - 1$. Thus, we estimate one parameter from our sample and lose one degree of freedom. Similarly, when we consider the least squares regression line, we need to estimate two parameters, the slope and the intercept. The equations for the sample slope and the sample intercept impose two restrictions on the data values, and we lose two degrees of freedom.

Gretchen Davis (davis@stat.ucla.edu)*, who is the Visiting High School Teacher in Statistics at UCLA this year, serves as one of the Table Leaders for AP Statistics.*

Table 1.

	Column D	Column E	Column F	Column G	Column H	Row Total
Row A						n_1
Row B						n_2
Row C						n_3
Col Total	n_4	n_5	n_6	n_7	n_8	Sum

Table 2.

	Column D	Column E	Column F	Column G	Column H	Row Total
Row A	f_1	f_2	f_3	f_4	x_1	n_1
Row B						n_2
Row C						n_3
Col Total	n_4	n_5	n_6	n_7	n_8	Sum

Connections to Two-Way Tables

Table 3.

	Column D	Column E	Column F	Column G	Column H	Row Total
Row A	f_1	f_2	f_3	f_4	x_1	n_1
Row B	f_5	f_6	f_7	f_8	x_2	n_2
Row C						n_3
Col Total	n_4	n_5	n_6	n_7	n_8	Sum

Table 4.

	Column D	Column E	Column F	Column G	Column H	Row Total
Row A	f_1	f_2	f_3	f_4	x_1	n_1
Row B	f_5	f_6	f_7	f_8	x_2	n_2
Row C	x_3	x_4	x_5	x_6	x_7	n_3
Col Total	n_4	n_5	n_6	n_7	n_8	Sum

For a two-way, or contingency, table of r rows and c columns, the same principle holds. Consider Table 1 with 3 rows and 5 columns. The row and column totals in this table are represented by n_1 through n_8, and Sum = $n_1 + n_2 + n_3 + n_4 + n_5 + n_6 + n_7 + n_8$.

If we start placing values in the first row, we obtain Table 2. Since the sum of the first row is n_1, we can consider the first four values as free (denoted by f_1 through f_4), but then the last value is determined (denoted by x_1).

Similarly, in the second row, four of the values are free and the last is determined (see Table 3). By the time we get to the third row, those values are all determined in order to obtain the correct column totals (see Table 4).

In the entire table, seven of the cells are determined by the other cells, leaving eight of the cells (denoted by f_1 through f_8) free to vary. Thus, the degrees of freedom for a 3 by 5 table is 8. In general, for a table with r rows and c columns, we lose c degrees of freedom because the cells in the last row are determined by the entries in the other $r - 1$ rows, and we lose $r - 1$ degrees of freedom because the cell in the last column of each row is determined by the previous values in that row. Therefore, of the rc cells in the table, $rc - c - (r - 1) = (r - 1)(c - 1)$ are free to vary.

Conclusions

For AP Statistics students, it may help to illustrate degrees of freedom by the analogy between restrictions of physical motion of a point in space and the loss of degrees of freedom. In general, each time we impose a restriction (estimate an unknown parameter from our known sample data), we lose one degree of freedom.

Further Readings

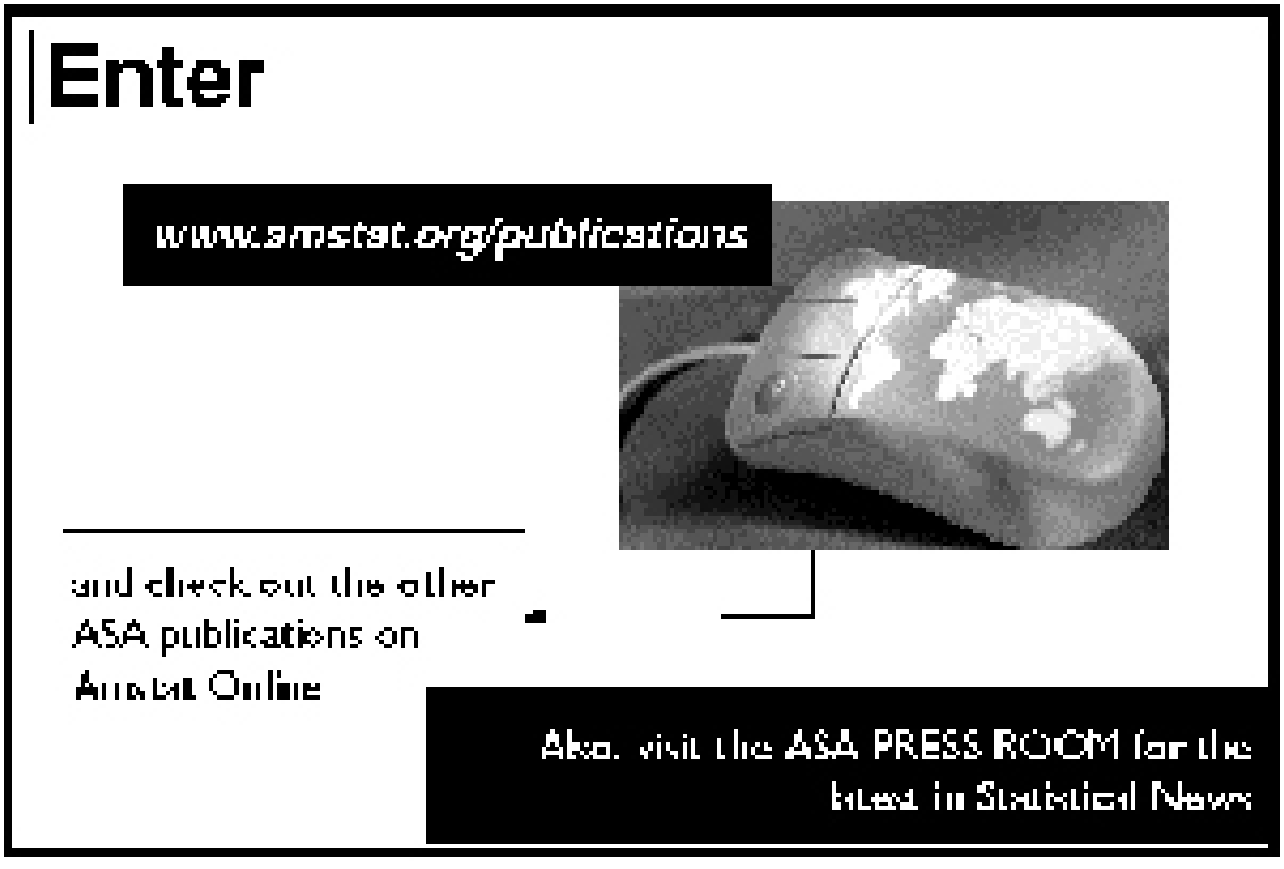
The Walker article continues to discuss topics which are beyond the scope of AP Statistics but which may be of interest to continuing students and teachers: the representation of a statistical sample by points in *n*-dimensional space, the importance of the concept of degrees of freedom, and illustrations of how to determine the number of degrees of freedom appropriate in certain common situations.

Reference

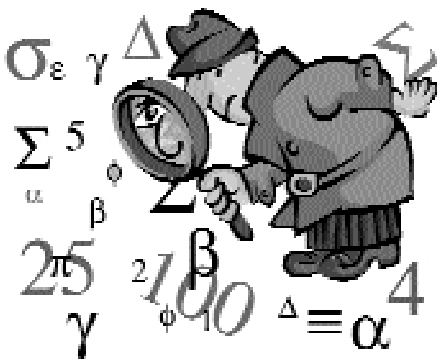
Walker, H. M. (1940), “Degrees of Freedom,” *Journal of Educational Psychology*, 31, 252–269. Reprinted in *Readings in Statistics for Behavioral Scientists* (ed. J.A. Steger; Holt, Rhinehart, and Winston, 1971) pp. 346–363.

Acknowledgement

I am grateful to Chris Olsen at George Washington High in Cedar Rapids, Iowa, for sharing the Helen Walker article.



Data Sleuth



This feature invites you to solve mysteries involving data. We hope that you will find this feature to be fun and enlightening, and we encourage you to send us your own submissions of data mysteries.

Mystery 1: The Yolk's on You!

Contributed by Rick Burdick, Arizona State University

Students in my quality analysis course at Arizona State University conducted a project to study variability in the weights of eggs purchased in local supermarkets. The study compared grade AA large eggs with grade AA extra large eggs for two store brands (Fry's and Abco) and one premium brand (Hickman's). The eggs were all purchased on April 4th, 1999 and weighed (in grams) on electronic scales made available to the students by the chemistry department. After weighing all the eggs (one dozen for each combination), the students prepared the table of group means and the interaction plot shown below:

Table 1. Average Weights (in grams)

	Fry's	Abco	Hickman's
Large	64.21	58.01	59.47
Extra Large	63.46	62.79	64.13

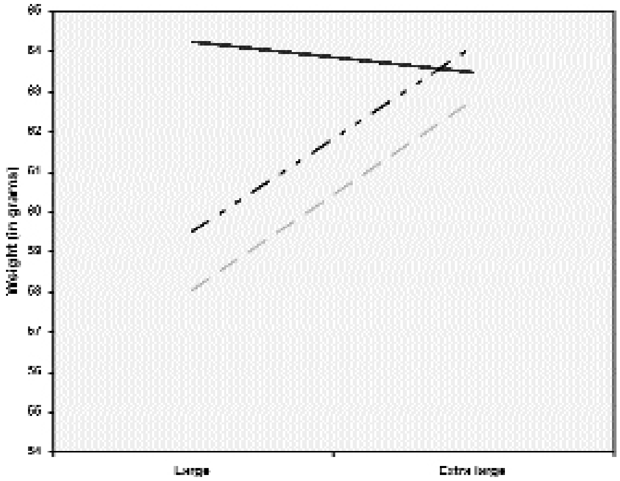


Figure 1. Interaction Plot of Treatment Means

Question 1: Which store's egg weights seem to follow a different pattern than the other two stores?

Question 2: Which size (large or extra large) seems to have the anomalous weight for that store?

Question 3: Based on the information provided, suggest a plausible explanation to resolve this mystery.

(The solution appears on page 24.)

Mystery 2: Buy Me Some Peanuts and Cracker Jacks!

Contributed by Beth Chance and Allan Rossman

The *Team Marketing Report* newsletter annually reports on the average cost of a family of four attending a Major League Baseball game. This report calculates a "fan cost index" by examining prices of parking, soda, beer, hot dogs, programs, and caps, in addition to the costs of adult and child tickets. The stem-and-leaf plots below present the distribution of prices for the 30 Major League teams in the 2000 season on the program, cap, and parking variables:

```
Stem-and-leaf of Program  N = 30
Leaf Unit = 0.10

2  0
2  5
3  000000233
3
4  000000000000
4
5  000000
5
6  0
```

```
Stem-and-leaf of Cap   N = 30
Leaf Unit = 0.10
--------------------------------
 7   6
 8   0
 9   009
10   0000000002
11
12   000000000000
13   0
14   0
15   0
--------------------------------
```

```
Stem-and-leaf of Parking  N = 30
Leaf Unit = 0.10
--------------------------------
 4   0
 5   000000
 6   00000056
 7   00059
 8   00000
 9
10   00
11   0
12   0
--------------------------------
```

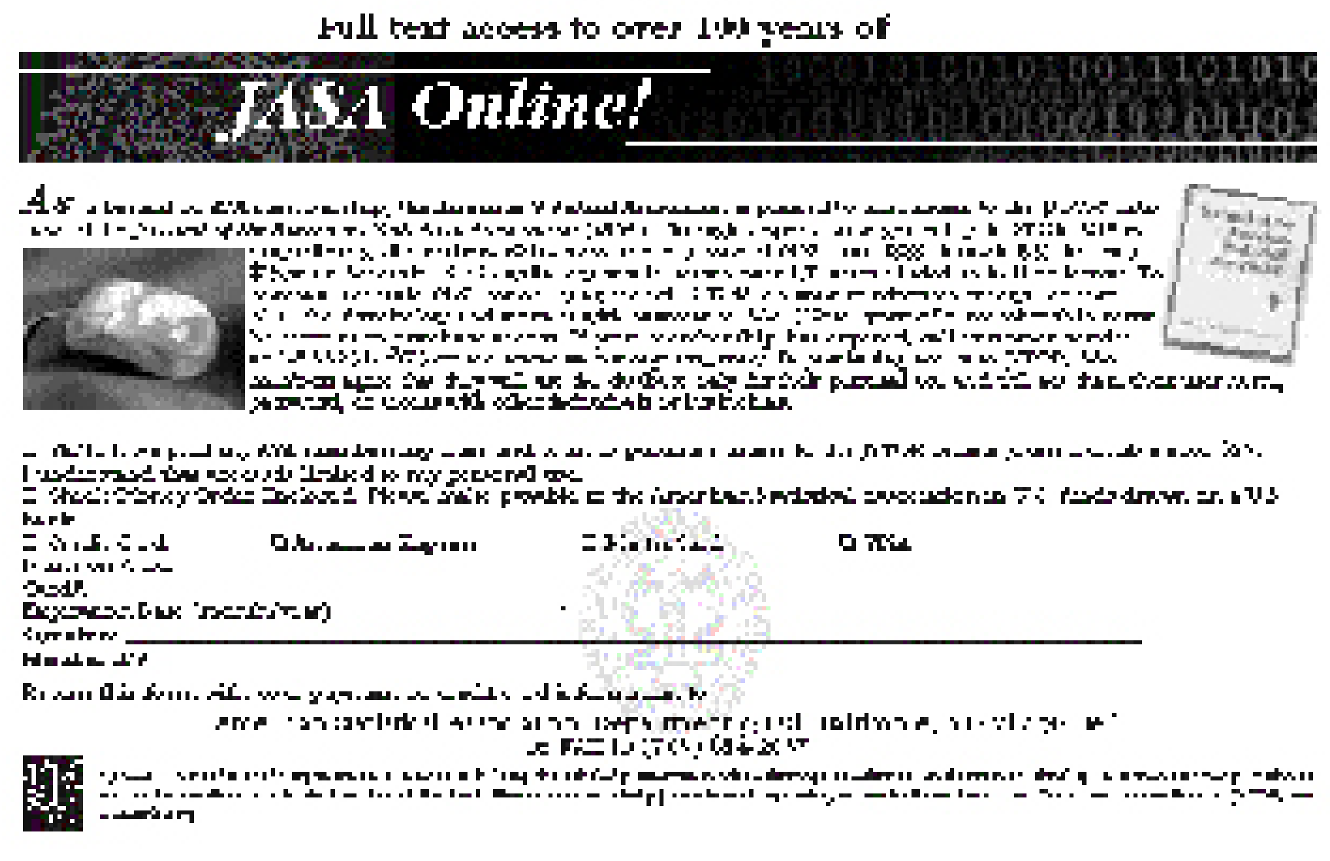
Question 1: What value does the leaf unit have for the large majority of these cases? Explain why that makes sense.

Question 2: For the leaves with different values than the answer to #1, some of them equal 2 and some equal 5. Explain why this also makes sense.

Question 3: Eliminating the answers to 1 and 2 leaves (ooops, pardon the pun!) the following the leaf values: 3 and 3 for program, 6 and 9 for cap, 6 and 9 for parking. Come up with a plausible explanation for why there are two teams with such odd leaf values for every variable, and identify which two teams they are.

(The solution appears on page 24.)

If you want to explore these data further, you can find them at *www.amstat.org/publications/stats.*



μ-sings: Statistics in the Media

The Lady Tasting Tea: A Thoroughly Delightful Read

Chris Olsen

It is only with a refined sense of presumption that a member of one generation would take up the task of recommending reading material to the next – well, OK, maybe the next next — generation of statistics students. As I write, the Harry Potter movie is out, and only today have I purchased the first Harry Potter book. Clearly my usual choice of reading material is out of touch with the current scene. However, I feel particularly confident recommending *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*, by David Salsburg, as a terrific read for all students of statistics. I feel particularly confident not because I am a statistics teacher, but because I am an experienced parent, a true believer in the mythology of parenthood, and, of course, familiar with The Speech.

As today's students know from their parents' rendition of The Speech, their parents lived in a terribly frightful world. Unlike today's soft generation, all parents had to walk to school; frequently uphill both ways, through rain, sleet, or hail, and didn't have modern conveniences like TV's, CD's, and calculators. And the worst part was trying to read about the history of mathematics! Young math majors, interested in their field of study, would quite naturally pick up histories of mathematics. These histories were abundantly supplied not merely with equations, but with equations beyond mortal comprehension. Following the story line was not unlike taking another math class! Today the situation is different. Not only is *The Lady Tasting Tea* a thoroughly delightful read, providing a fascinating escape into the recent history of statistics, there is not equation one in the whole book! (In the interest of avoiding The Speech, do *not* share this book with your parents.)

The book's title harkens back to one of the legendary incidents in the history of statistics. At a party in Cambridge in the late 1920's a lady claimed that she could tell by taste whether tea had been added to milk, or milk added to tea during the preparation of the drink. The university dons in attendance were somewhat skeptical of this ability, and a healthy discussion ensued. Who should pop up but R. A. Fisher, the father of modern statistics? Never one to pass up a chance to educate any crowd about statistics, he proposes a test to see if her claim can be verified using his theory of randomized experiments.

In the *Lady Tasting Tea*, David Salsburg is equally reluctant to pass up a chance to educate and regale readers with descriptions of the people and the personalities that shaped the direction of twentieth century statistics. Fisher, of course, is a major protagonist, and the statistical work of Karl Pearson, William Gossett, and Fisher is coherently and deftly interwoven with the story of their early disagreements. Did I say disagreements? What I meant was, serious arguments! Fisher argued with everybody and took no rhetorical prisoners. One of the most charming anecdotes in the book recalls a paper delivered by the eminent statistician Jerzy Neyman in French in the 1950's. As he went to the podium, Neyman realized that Fisher was in the audience, and during his presentation he steeled himself for the inevitable Fisherian pounce during questions from the audience. After the presentation, Salsburg writes, "… Fisher never stirred, never said a word. Later, Neyman discovered that Fisher could not speak French."

Women do not only appear in the title. Florence Nightingale (F. N.) David, daredevil on cross-country motorcycle races and author of multiple books, is shown to be a feisty pioneer in statistics. Her self-described "worm's-eye view" of the early statisticians, as well as the discussion of her own war work during the German blitz of London, is fascinating reading. And

Chris Olsen (colsen@esc.cr.k12.ia.us) *teaches mathematics and statistics at George Washington High School in Cedar Rapids, IA. He has been teaching statistics in high school for 25 years, and AP Statistics since its inception.*

then there is Gertrude Cox, assistant to George Snedecor. Snedecor, founder of the Statistical Laboratory at Iowa State University, was asked to recommend someone to head up a similar laboratory at North Carolina State University. Could he recommend a man to head such a department? Snedecor could. He made a list of ten men, and called Cox in to verify the list. She looked it over, and asked, "What about me?" Snedecor added a line to his letter: "These are the ten best men I can think of. But, if you want the best person, I would recommend Gertrude Cox." As Salsburg writes, "Since the days of Snedecor and Cox, the 'best person' has frequently been a woman."

So let's make a cross-generational deal, you and I. This weekend I will curl up with the Sorcerer's Stone; you curl up with *The Lady Tasting Tea*. We will each discover anew the pure joy of reading a well-crafted book. If you are already an accomplished student of statistics, this book will bring a new appreciation of the subject and an almost personal acquaintance with the men and women who were, as it were, present at the creation of statistics. If you are a beginning student of statistics, the *Lady* will deliver not only the personalities but also an account of the interesting statistical problems that drew their attention.

Oh, one last thing — if you do want to pass on the tea-tasting lady to your parents, there's plenty of room to scribble some well-chosen equations in the margins.

References

Salsburg, D. (2001), *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*, New York: W.H. Freeman and Company.

Two excellent, more scholarly references on the history of statistics are:

Stigler, S. M. (1986), *The History of Statistics: The Measurement of Uncertainty Before 1900*, Cambridge, MA: The Belknap Press of the Harvard University Press.

Stigler, S. M. (1999), *Statistics on the Table: The History of Statistical Concepts and Methods*, Cambridge, MA: Harvard University Press.

A more informal account of interesting statistical personalities is:

Peters, W.S. (1987), *Counting for Something: Statistical Principles and Personalities*, New York: Springer-Verlag.

Data Sleuth Solutions

Mystery 1: The Yolk's on You!

Question 1: Clearly, Fry's has something fishy going on with their egg weights.

Question 2: The weight of extra large eggs at Fry's is consistent with the other stores, so the large size has the anomalous weight.

Question 3: The students working on the project, Jeff Cummings, John Johnston, Rudy Pinon, and Jim Walewander, made a phone call to the egg department at Fry's to clear up the mystery. They found out that Fry's had trouble meeting the demand for eggs at that time because hens do not always lay the egg size that is in greatest demand around Easter. Since Fry's experienced a shortage of large eggs, they took extra large eggs and put them in the large egg containers. (It would be illegal to do the reverse.)

Mystery #2: Buy Me Some Peanuts and Cracker Jacks!

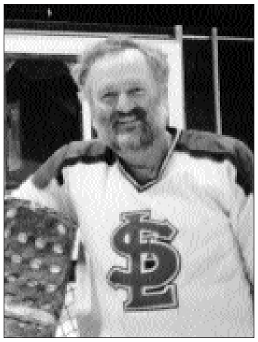
Question 1: The vast majority of the leaves have the value 0, suggesting that most of these prices are whole dollar amounts. This makes sense so that vendors do not have to worry about making change with coins.

Question 2: Leaf values of 2 and 5 also make sense, as they probably represent prices that are 25 cents or 50 cents more than a whole dollar amount.

Question 3: The two teams that have unusual leaf values in all of these variables are the Toronto Blue Jays and the Montreal Expos. The Team Marketing Report newsletter reports prices in American dollars, so the Canadian prices for those teams underwent that conversion and therefore did not appear as nice round integer values.

The Statistical Sports Fan

What's the Score in the NFL?



Robin Lock

Scoring patterns in American football games are unique since the most common methods of scoring points, a touchdown (with a kicked extra point) and a field goal, yield seven and three points, respectively. Other scoring possibilities, a two-point safety, a two-point conversion after a touchdown, or six points for a touchdown with a missed conversion, are fairly rare. Thus, a score such as 14-10 is much more likely than an 11-5 game. Participants in various sorts of football pools, in which many students and teachers of statistics may indulge, are often asked to predict individual game scores or choose game outcomes that depend on the margins of victory. Would a statistical examination of past game results help us make more reasonable forecasts of future football outcomes?

The Dataset

To investigate the characteristics of football scores, we use a dataset consisting of the scores from all regular season games played in the National Football League during the 1998, 1999, and 2000 seasons. For each of the 736 games played during this period, we know the identity and points scored by the home team and the visiting (road) team. We also have a "pointspread" assigned to each game that reflects an estimate of the relative strength of the two teams. NFL game scores can be found at a number of web sites (e.g., *www.nfl.com*). The pointspreads used here were obtained from ESPN's Pigskin Pick'em game (*games.espn.go.com*). The dataset can be downloaded as an ASCII text file from the *Journal of Statistics Education* data archive at *www.amstat.org/publications/ jse/jse_data_archive.html*.

How many points do NFL teams score in a game?

With scores for two teams in each of the 736 games, we have a total of 1472 team scores to examine. Figure 1 shows their distribution. We note that the most frequent score during this period was 24 points (102 times), followed by 20, 17, 31, and 10 points.

Scores that are multiples of seven are less common than many football fans might expect. In fact, only 19% of these game scores are divisible by seven. If we look at the frequency of game scores mod 7 (Table 1), we see that the most common result is 3 (generally some touchdowns plus a single field goal), then 6 (two field goals or a touchdown with missed conversion). These account for more than half of the game scores, while scores such as 11, 18, and 25 points are very rare. We see a similar pattern in the winning margins, with 1 in 4 games being decided by exactly one field goal or one touchdown (Table 2). The most common game scores (both teams) during this period were 16-13, 20-17, 13-10, and 23-20, each occurring about a dozen times. Note that the NFL uses an overtime period to help minimize tie games and that overtime games are frequently won with a field goal.

Can the number of points scored by one team help us to predict the points scored by its opponent?

Tennessee scored 36 points in its first game of the 1999 season against Cincinnati. What information does that give us about how many points Cincinnati scored in the game? Perhaps the game was a high scoring affair. When one team scores a lot, its opponent may be inclined to be less conservative and take more risks to match that high score, thus increasing the probability that they will also score a lot. On the other hand, a team scoring many points may be much stronger than its opponent, so perhaps we should expect a relatively low score from the other team. The first scenario would argue for a positive correlation between scores in the same game, while the second would lead to a negative correlation.

As Figure 2 demonstrates, neither viewpoint is correct. In fact, the home team's score is remarkably independent of the number of points scored by its opponent (r = 0.023, p-value = 0.540). Note: The plot

Robin Lock (rlock@stlawu.edu) *is the Jack and Sylvia Burry Professor of Statistics in the Department of Mathematics, Computer Science and Statistics and hapless goalie for the faculty/staff ice hockey team at St. Lawrence University.*

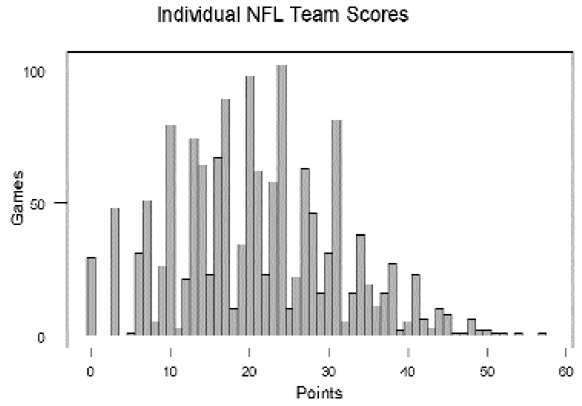


Figure 1.

uses a technique known as jittering to randomly offset points that would be plotted at the same location (i.e., identical game scores) so that we can see the multiplicity better.

One can also investigate this relationship with a contingency table. If we classify scores of each team in natural intervals of 0-6, 7-13, etc. with all scores above 35 points as a last group, a two-way table of home vs. road scores yields a chi-square value of 35.4 with 25 degrees of freedom and a p-value of 0.082. This would indicate no significant association (at a 5% level) between the home and visiting team scores.

How well do pointspreads predict game outcomes?

The pointspread is a device used to handicap games in order to make the "outcome" as uncertain as possible. In our dataset the pointspreads are applied to the road team's score (so a negative point spread indicates that the road team is favored). Thus, to determine the winner "against the spread," we add the value of the pointspread to the visiting team's score before determining the result. For example, the first game in the data set has Jacksonville playing at Chicago in the first week of the 1998 season. The pointspread for that game was –8.5 points, so Jacksonville was perceived as the stronger team by a bit more than one touchdown. The final score for that game was Jacksonville 24 – Chicago 23. Although Jacksonville won the game, the winner against the pointspread was Chicago. All pointspreads are of the form integer + 0.5 so a winner against the spread is determined for every game.

Table 1. Individual game scores (mod 7)

Score (mod 7)	Freq.	Pct.
0	279	19%
1	84	6%
2	209	14%
3	435	30%
4	31	2%
5	101	7%
6	333	23%

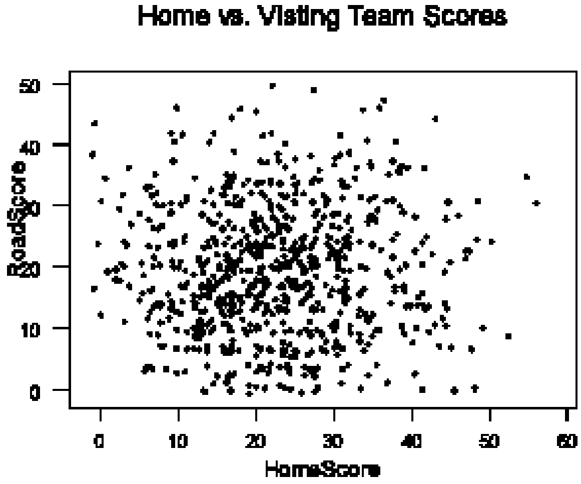


Figure 2.

Figure 3 shows a plot of the actual game difference (Home score – Away Score) vs. the pre-game pointspread. Although the plot shows a fair amount of scatter, there is a clear positive association (r = 0.44, p-value = 0.000). Thus a team that is favored to win by a lot tends to win by more points than a team with a smaller spread. Data to the right of the y-axis (the "pointspread = 0" line) are cases where the home team was favored by the pointspread (66.2%), while points above the x-axis (the "actual difference = 0" line) are games that the home team actually won (59.2%). Points in the first (upper right) and third (lower left) quadrants represent games in which the pointspread accurately predicted the winner of the game (66.8% of these games). "Upsets" appear in the second and fourth quadrants, where the actual game difference has a different sign than the pointspread.

The "$y = x$" line that's drawn on the plot indicates how teams did against the pointspread. Points above the line are games in which the home team beat the pointspread (52.3%) and those below the line are games where the winner against the spread was the visiting team (47.7%). A two-tailed test shows that this proportion is not significantly different from 50%

Table 2. Most Common Winning Margins

Winning Margin	Freq.	Pct.
3	112	15%
7	37	10%
10	42	6%
1	36	5%
17	33	4%
14	32	4%

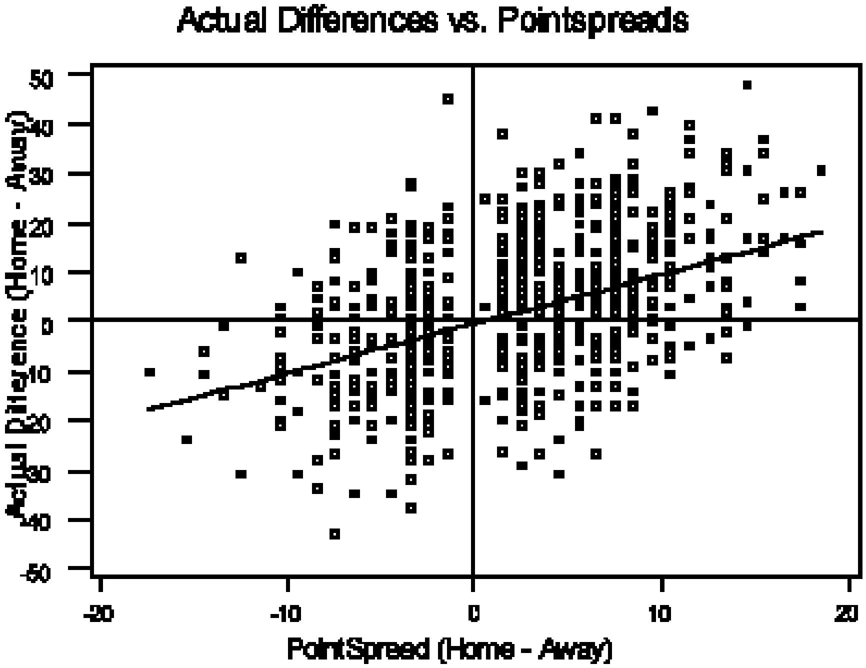


Figure 3.

(p-value = 0.210), so there would appear to be no statistical advantage to always picking for or against the home team when using a pointspread. Points between the "$y = x$" line and the x-axis represent the games (18.3%) where the winner against the spread was different than the winner of the game.

The favored team successfully beat the pointspread in 358 of the 736 games (48.6%) with the underdog either winning the other games outright or at least coming close enough to cover the spread. As with the home teams, there is no statistical advantage, when using the pointspread, to always picking for or against the favored team (p-value = 0.461). The least squares line through these data is remarkably close to this "$y = x$" line. Its equation is *Actual Difference = 0.93 + 1.02 × Pointspread*. A t-test to see if the slope is significantly different from one yields a p-value of 0.8318. Thus every increase of one point in the pointspread would yield an expected increase of about one point in the actual difference. It would appear that the oddsmakers do a pretty good job of "tilting the playing field" to even up the game results.

Suggestions for further investigations

Are the scoring patterns for other levels of football (e.g., college, high school, Canadian Football League) similar to those in the NFL? Do the patterns discovered in 1998–2000 hold in earlier (or later) seasons? How significant is the home field advantage (if one exists at all)? Some fans will always pick the underdog when the pointspread is double digits. Is this a reasonable strategy? Although the slight advantages for home teams and underdogs against the historical pointspreads were not statistically significant, would there be a clear advantage to always picking an underdog that is playing at home? Can you determine a method that uses past games scores to predict future outcomes that does better than the pointspread at picking the winning teams?

Conclusion

What score should we choose the next time we are faced with predicting the outcome of a football game? Without any other indication of the relative strength of the teams involved, the analysis above would suggest the most common margin of victory (3 points) be awarded to the home team. The median scores for home and visiting teams in the data set were 22 and 20 points, respectively, but we know that 22 = 1(mod 7) is pretty rare. The mean score for home teams was 22.5, so perhaps we should round up and predict a 23-20 victory for the home team, giving us scores congruent to 3 and 6 (mod 7) and matching one of the four most common game results. But then we also might want to consider the pointspread, look at records against common opponents, check on the status of injured players, adjust for play in recent games, factor in the phase of the moon and make a guess.

PRIZE ANNOUNCEMENT

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

The committee on Statisticians in Defense and National Security of the American Statistical Association is pleased to announce the second annual prize for the best student paper applying statistics to defense issues. The 2001 prize was won by **John Leffers** for his paper titled "Statistical Validation of Track Quality Numbers for Joint Interoperability Testing of Theater Air and Missile Defense Families of Systems".

The prize competition is open to any undergraduate or graduate student enrolled in any institution of higher education. The paper must have been written in the preceding academic year (**for this year's prize, July 1, 2001 to June 30, 2002**). The paper must be nominated and submitted by a faculty member at the institution. Papers are limited to 5,000 words or 20 pages, including graphics. Student Theses meeting these length requirements are acceptable.

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

For 2002, the prize consists of a plaque and $500

For 2002, nominations and three copies of the paper should be submitted to Professor Dave Olwell, Department of Operations Research, Code OR/OL, Naval Postgraduate School, Monterey, California, byJuly 1, 2002. Questions should also be addressed to Professor Olwell at *dholwell@nps.navy.mil*. The prize announcement will be made at the annual JSM meetings.