Enumerating Data Errors
A Survey of the Counting Literature

Elizabeth M. Pierce
MIS & Decision Sciences
Indiana University of Pennsylvania
203 Eberly College of Business
Indiana, PA 15705-1087

(Phone) 724-357-5773
(Fax) 724-357-4831
Internet: EMPIERCE@IUP.EDU

Abstract

How many errors of a given type are in a database? To answer that question, the author searched through the literature from several different disciplines: statistics, business, wildlife studies, quality control, and database management. The search revealed that the problem of enumeration is one that touches many disciplines and has produced a large number of techniques for answering the question: How Many? In this paper, the author lists the counting techniques that were found and describes how they might be applied to the problem of counting data errors.

Introduction

How many errors of a particular type are in a given data set? Omitted codes, inconsistent formats, inaccurate or out-of-date values are just a few of the problems that can handicap organizations who want to use tools such as data mining to extract as much information as possible from their data. As organizations explore new ways to make use of the data stored in their information systems, there is a growing interest in how the quality of data can be improved. For these organizations, the ability to count the incidence of erroneous datum is fundamental to the monitoring, analysis and improvement of their data's quality.

In this paper the author has tried to compile as many counting techniques as she could find that are applicable to counting erroneous data values. These techniques come from a variety of fields such as statistics, accounting, quality control, and wildlife management. Although other authors have covered these techniques extensively over the years, the objective of this paper is to list and summarize these techniques in a single source. The author hopes that this work will benefit those who need to measure data quality by providing a compact summary of available counting techniques for a wide variety of situations.

The paper organizes the counting techniques into three categories: census methods, sampling based methods, and catch data methods. Each section contains descriptions of the techniques as well as suggestions as to what kinds of data errors may be counted using those techniques. The census methods section contains techniques where it is feasible to count all the instances of erroneous data in the population. The sampling based methods section contains techniques where one uses a sample to estimate the number of erroneous values in the data population. The removal models section contain techniques whereby one uses the number of data errors that are brought to the attention of the data administrator as a basis for estimating the total number of erroneous values in the data population.

For all the sections, it is assumed one is interested in counting for a specified file, the number of records (N_e) that possess an erroneous value for one or more specified attributes. The words, "erroneous value", are meant as a general description of some specified defect (or set of defects) with the datum's value that may be of concern to the end user. In some cases, it may be preferable to report the incidence rate as a proportion (N_e / N) rather than as a raw number where N represents the total number of records in the database.

In addition, there are several issues that can complicate the counting of data errors. Because an individual record is often composed of several attributes, one may find it more useful to count the number of erroneous attributes per record rather than the total number of errors. Moreover, not all attributes are equally important. Thus an error in one attribute may be of much more concern than an equal or larger error in another attribute. In this case, one may want to restrict the counting of errors to those occurring in certain key attributes or at least differentiate the counting of critical errors compared to minor errors.

Another question that often accompanies the count of data errors is the question of magnitude. For example a database with many values that are one cent off their true value may be considered more desirable than a database with a few values that are one million dollars off their true value. Although this paper does not deal directly with measuring the magnitude of erroneous data, the techniques listed in this paper could be adapted to this question. Alternatively one could simply develop various error magnitude categories and count the number of records that apply to each.

1. Census Methods

Thanks to computerized file systems, the incidence rate of many types of errors can be counted exactly. Difficulties one encounters when trying to count objects in nature such as the lack of a complete population listing or inaccessible objects are typically not a problem in the structured world of database management systems. In addition, unlike many objects in nature, the counting task for data stored electronically can often be done by a computer. In this section, the author presents two methods for conducting a census to determine the number of errors in one's data. The first section comes from the literature on automatic data editing and imputation. The second section describes cycle counting which comes from the inventory literature.

Automatic Data Editing & Imputation Literature

The identification problem is the litmus test for whether the computer can be used to obtain a complete census of the number of erroneous records or whether another technique such as sampling or catch data must be used. Computer programs do well in identifying errors that involve duplicate records, outliers, missing values, invalid codes, inconsistent formats, or logical relationships (i.e. if medical diagnosis is pregnant then patient's gender must be female). In order to automatically detect data errors, a person must write a computer program to read the first record from the database, then compare the record's attributes against a list of editing constraints. If any violations are found, the program may choose to tag and count the erroneous record, correct it, or delete it depending on the nature of the error and the cost of correcting the record. This process is repeated until all records in the database have been checked.

There are a number of publications (Naus [1975], Naus, Johnson & Montalvo [1972], Fellegi & Holt [1976], Toney [1992]) that describe various strategies for automatically locating and correcting erroneous records. Even when the computer cannot locate the error exactly as in the case when the violation of constraints cast suspicion on several attributes, the computer can be programmed to calculate a probability for which field(s) are most likely to be in error. In addition, new techniques such as artificial neural networks (Berry & Linoff [1997]) can also be used to train the computer to recognize and correct erroneous records.

Inventory Literature

Sometimes it is desirable to check all the records for errors; however, it may not be practical or economical to check the records all at once. This is particularly true when the record must be manually verified against its original source. For this situation, one might want to consider cycle counting (Backes [1980], Wilson [1995]). Cycle counting refers to a technique that was developed in the late 1970's for use in managing inventory records. The objective of cycle counting is to maintain an acceptable level of accuracy in the inventory records. Instead of accumulating errors in the records until the next annual physical inventory, cycle counting is used to check a portion of the items (and their corresponding records) each day, week, and month until all have been counted. The idea behind cycle counting is that by checking a portion of the records at a time, the inventory records can be corrected sooner, the investigation of the cause of an error and the reconciliation occur closer to the time of the transaction, and production disruptions are minimized. Over the years, cycle counting has evolved to take into consideration different strategies for setting the cycle schedule, selecting which items to count, and incorporating statistical quality control techniques for eliminating quality problems. Although developed with inventory in mind, this technique could be adapted for any database where a manual and complete inspection of the records is required.

2. Sampling Based Methods

Sometimes error identification is only possible by comparing the datum's current value to a non-electronic source, thus preventing a complete enumeration by computer. When people rather than computers are required to do the counting, a complete census of a large data set is expensive, time consuming, and prone to measurement error. This measurement error occurs because as people are counting the errors in the database, the database's error counts are changing as records are added, deleted, and updated. Sampling techniques allow one to estimate the error count of the entire database based on a small subset of randomly selected records. The reduced size allows the count to be done quickly, with less expense and using just a few skilled enumerators to accomplish the task. In this section, the author presents four sets of sampling based techniques organized primarily by the area of literature from which they come. The four sets of literature are Statistics & Auditing, U.S. Census, Quality Control and Database Management.

Statistics & Auditing Literature

Statisticians and auditors have used sampling techniques for the purpose of count estimation (attributes sampling), as well as determining the magnitude of errors (variable sampling). In addition, sampling can be used to determine if a data set is of sufficient quality that it should be used (acceptance sampling) or for learning what types of errors exist in the data and how big a sample should be drawn (discovery sampling). The two sources referenced in this paper are Cochran [1963] and Arkin [1963]; however, one can find additional descriptions of these techniques in practically any statistics or auditing textbook.

For statisticians and auditors alike, the key to using sampling to estimate counts is the ability to draw a randomly selected set of records from the database so that every record has an equal chance of being included in the sample. With computerized record systems, it is relatively easy to conduct a simple random sample. Suppose someone wanted a sample of 100 randomly selected records from their database. Using a random number generator to select 100 numbers between 1 and N (the total number of records in the database), the person could then use those records corresponding to those positions in the database as a sample.

Alternatively some people prefer another technique known as systematic random sampling. In this method, one starts at some random position in the database and then chooses those records whose position falls on some given interval. For example, in a database of 20,000 records, one could choose every 200th record starting at some random point in the database in order to select 100 records for the sample.

One problem with simple and systematic random sampling is that there may still be problems in obtaining an accurate count, particularly if the errors themselves are not randomly distributed among the records in the database. For instance, change of address errors may be more acute among the younger adults in the population (due to job and life style changes) and older adults (due to moves to retirement homes). A better method for selecting a sample in this case would be to use a stratified sampling technique. In stratified sampling, the database records are first broken down into distinct strata or segments based on characteristics of the entity such as a person's age. Then each segment is sampled separately and independently using simple or systematic random sampling techniques.

Another way errors can be non-randomly distributed in the database is due to production run problems that may occur during a batch update of records. As a result, sections of records in the database may be more error prone than others. An appropriate sampling technique to use here is cluster sampling. In the cluster sampling method, a sample of 200 items might be obtained by drawing 10 sampling units at each of 20 different points in the database. One can use simple random or systematic sampling to obtain the points of location of the clusters, with the sample including that unit and the nine sampling units immediately following the selected point.

Sometimes the objective is to obtain an overall picture of the health of the data quality in an organization. In many businesses, there are a multitude of databases at both the corporate division level and the departmental level. Multistage sampling involves sampling on several levels. The selection of the first stage or primary sampling unit (division level files) as well as that of the secondary sampling unit (department level files) can both be accomplished, using either random number or systematic sampling techniques.

Once the sample has been obtained, the enumerators can then go to work inspecting the records and determining the number of errors found in the sample. For a simple random or systematic sampling the formulas for estimating the total number of errors in the population are straightforward. The estimated proportion of errors (p_e) in the database is the number of errors found in the sample (n_e) divided by the sample size (n). The estimated number of errors in the database (N_e) is equal to N * p_e. Using the confidence coefficients from the standard normal curve, a confidence interval for the proportion of errors in the database is given by:

$$p_e \pm Z_{\alpha/2}\sqrt{\frac{p_e(1-p_e)}{n}}$$

For example, suppose a database contains 50,000 records and in a simple random sample of 200 records taken from that database, one finds 23 records which contain a specified error. Then the estimated proportion of error in the database is 23/200 = .115 or just 11.5%. The estimated number of errors in the database is .115 * 50,000 or 5,750. A 95% confidence interval for the proportion of errors in the database would be:

$$.115 \pm 1.96\sqrt{\frac{.115(1-.115)}{200}} = .115 \pm .044$$

By multiplying through by N = 50,000, one could also obtain a 95% confidence interval for the number of errors in the database: 5,750 ± 2,200. Additional formulas are available in most statistical and auditing handbooks for estimating population counts using the stratified, cluster, and multistage sampling methods. For small databases, a correction factor can be applied to adjust the estimates for finite populations where the sample size is more than 5% of the total record population.

Although it is possible to choose a sample size arbitrarily, in most cases one wants to ensure that one is taking a sample that is big enough to yield a decent estimate, but not so big that one is wasting time and money. The choice of sample size depends on the reliability required (how close is it necessary for the auditor to estimate the population parameter), the confidence level (probability that the sampling variability will indeed be confined to the promised range) and the amount of variability in the values of the data. For example, suppose one has a database with 500,000 records and he wants to know how many of his records possess a certain type of error. Using simple random sampling, how many records should he select for his sample? To complete the problem one needs an estimate for the maximum rate of occurrence of this particular error (say at most 10% of the records could possess this error) as well as the level of confidence (say 95%) that the sample estimate is within ± 2% of the actual error level in the entire database. The classical formula in this case is given by:

$$n = \frac{(Z_{\alpha/2})^2(pq)}{e^2} = \frac{(1.96)^2(.10)(.90)}{(.02)^2} \approx 865$$

Here p refers to the belief that the maximum error rate in the database does not exceed 0.10 and q equals (1 - p) or 0.9. The Z value comes from the standard normal table and refers to the fact that if one moves 1.96 standard deviations in either direction from the center one will have covered 95% of the area under the normal curve. The error, e = 0.02, measures how much one is willing for the sample estimate to vary from the true population measure. The appropriate sample size in this case is 865 records selected randomly from the database. Note that the size of the database is not a factor in determining the appropriate sample size.

It is important to note that because one must determine the maximum error rate in the database in order to get the sample size, there is a risk that the actual reliability of the sample will depart from the desired level of reliability. If the actual error rate is well below the estimated error rate for the field then the actual reliability attained by using a sample of the indicated size is actually better than anticipated. On the other hand, if the actual error rate is greater than the estimated error rate for the field then the sample size used will provide a much poorer reliability than desired. To provide protection against either of these two events, it is important that the sample result be appraised after the sampling is completed to establish its true sampling reliability.

U.S. Census Bureau Literature

This next technique is interesting because it combines both a census count as well as a sampling count to come up with a combined estimated count which is on average, much nearer the truth when there is the possibility of undercounting with both the census count and the sampling count. The technique is called Dual System Estimation and the U.S. Census hopes to use this technique in the 2000 census to improve the U.S. Census count (Wright [1998]). The Dual System Estimation method is a variation of a statistical method called capture-recapture, that has been in use at least since 1896, when C. G. J. Petersen studied the immigration of a type of flounder into a fjord.

For a database the technique would work something like this. Suppose that there is a particular type of error in the database that is of special concern to management. Conventional census counting methods are used to get an initial estimate (i.e. the first measure) of the number of records possessing this particular error. The identity of each erroneous record is recorded so one knows exactly which records possess this error. Unfortunately, the nature of the error is such that there is concern that the count is below what it actually should be because neither computers nor people can be 100% infallible when it comes to detecting this error.

The second measure now comes into play. A random sample of records is selected from the database for a quality check. It is important that this sample be carried out independently from the census count. The selected records are carefully screened to get as accurate as possible count of the number of errors that exist in the sample (i.e. the second measure). Some of the errors detected will have already been detected in the first census measure and are therefore considered matches. Others will constitute a new detection. The combined and improved estimate can now be determined by using the following formula derived by Wright.

$$(one\_number\_census) = \frac{(first\_measure\_census)(second\_measure\_sampling)}{(number\_of\_matches)}$$

For example, suppose the database is inspected and the first measure produces a count of 100 errors in the database. For the second measure, a sample is taken and 60 errors are detected, 50 of which having already been detected in the census. The one number count would be 120 (= (100 * 60) / 50) for the estimated number of errors in the database.

Quality Control Literature

Sometimes what is of interest is not a single sample estimate of the error count in the database, but the patterns displayed by the samples over time. The control charts techniques found in the Quality Control literature may be of use when trying to use samples for monitoring and analyzing the number of data errors. References for control charts can be found in many books on statistical process control. For the purposes of this paper, the author used the book by Montgomery [1991].

Unlike auditing whose focus is on measuring the quality level of the data already in the database, control charts focus on the quality of the process by which data comes into the database. The purpose of control charts is to help those managing the data process to identify where errors occur, to prevent further errors, and to improve the performance of the data process.

The construction behind all types of control charts is essentially the same. At key parts of the data processing environment, samples are taken at periodic intervals. The results of each sample are plotted on a control chart. A number of lines separate the control chart into regions representing different levels of data quality. The centerline on the control chart represents the average measurement of the process. The upper control limit represents measurements that are 3 standard deviations above the average while the lower control limit represents measurements that are 3 standard deviations below the average. Additionally lines are drawn to represent measurements ± 1 and ± 2 standard deviations around the process average. As samples are plotted over time, the idea is to watch for patterns that might indicate that the error count is out of control due to a special cause such as a software or hardware problem. These patterns may appear as points that plot beyond the upper and lower control limits or an unusual number of points falling in one particular region of the chart.

There are several different control charts that are useful when trying to track data quality. A NP-Chart is useful when one is interested in counting the number of errors found in a sample of records. For instance, one might be interested in selecting 100 records at random from a transaction file and counting the number of records that contain an incorrect value for a given field. Its control limits and centerline are expressed by:

$$Control\_limits = n\bar{p} \pm 3\sqrt{n\bar{p}(1-\bar{p})}$$
$$Centerline = n\bar{p} = number\_of\_data\_errors\_in\_sample$$

P-Charts are analogous to NP-Charts except the P control chart is expressed in terms of the proportion of errors found in a sample of records. Its control limits and centerline are expressed by:

$$Control\_limits = \bar{p} \pm 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$
$$Centerline = \bar{p} = mean\_fraction\_of\_data\_errors\_in\_sample$$

A C-Chart is useful when one is interested in tracking the number of errors found per record. This is useful when the question is not whether or not a record contains an erroneous field value, but how many erroneous field values are there per record? Its control limits and centerline are expressed by the following formulas.

$$Control\_limits = \bar{c} \pm 3\sqrt{\bar{c}}$$
$$Centerline = \bar{c} = mean\_number\_of\_defects\_per\_record$$

Database Management Literature

The database quality literature has developed a number of models for estimating error counts. One way to model the number of errors in a database is to use Thomas Redman's idea that a database is like a lake fed by incoming streams. The cleanliness of the streams determines over time the cleanliness of the lake (Redman [1992]). Simulation experiments by the author support Redman's view that it is the quality level of the incoming transactions that predominantly dictates the quality of the database as a whole. Periodic clean up of the records in the database have only a short-term effect on the number of data errors in the database. No matter how complete the clean up, eventually the database error levels rise to reflect the error level of the incoming data. A number of techniques (duplicate performance, known errors, and simulation) have been developed to help one estimate the effect of the quality of the incoming record stream and detection efforts on the number of data errors remaining in a data set.

Duplicate Performance Method

One way to reduce the number of errors in the incoming records is to use the duplicate performance method (Strayhorn [1990], West & Winkler [1991]). In their respective articles, Strayhorn and West & Winkler developed probability models and confidence intervals for estimating the number of errors remaining in a database when it is practical to carry out the same job twice and to compare the results. For example in data entry, it may be faster to enter the data again than to check visually the entries that another person has made. Under the duplicate performance method, two data entry clerks each perform the same data operation, independently of each other, and then a computer program compares the results. Any disagreements are corrected, so that the only errors remaining in the data set are those where both data entry clerks were in error.

For example, suppose that a department uses two clerks who on a particular day, each performed 1,000 duplicate data entries. In comparing their performance, the computer uncovers 95 data entries that disagree. These 95 entries are compared with their sources and corrected. For the 905 other entries, one can surmise that either these records were correctly entered or both clerks made the same mistake when entering the data. The question is how many of these 905 entries are correct and how many are erroneous? In the case that the two clerks operate independently and both contribute equally to the number of disagreements, Strayhorn developed the following equation to estimate the number of remaining errors in the database.

Let j = Estimated Number of Errors remaining in the data set.
Let e = The number of data entries (in this case it would be 1,000).
Let d = The number of disagreements (in this case it would be 95)

Then $j = 0.5[e - d - (e^2 - 2de)^{0.5}]$ (which in this case works out to be between 2 or 3 errors left).

Known Errors Method

In addition to the duplicate performance method, Strayhorn and West & Winkler also developed probability models and confidence levels for the number of errors remaining in a database when a technique called the known errors method is used. In this technique, records with known errors are randomly placed into a data set. Then the data set, complete with the known (k_t) and any unknown errors, is given to a staff member whose job it is to check the data set. After the checking is finished, the results are tabulated as to the number of known errors found (k_f), the number of known errors missed ($k_t - k_f$), and the number of unknown errors found (u_f). Strayhorn used this information to estimate the total number of unknown errors (u_t) and the number of errors still lurking in the data set ($u_L = u_t - u_f$). Using ratios, he derived the equation: $u_t = u_f (k_t / k_f)$ to estimate the total number of unknown errors..

For example, suppose a supervisor inserts 12 known errors into a data set. In the subsequent check, 9 of these known errors are found and, in addition, 6 unknown errors are found. The total number of unknown errors originally in the data set is estimated to be 8 (= 6 * 12 / 9) and the number of unknown errors still lurking in the data set is estimated to be 2 (= 8 - 6).

The known errors method is useful when one wants to quantify the accuracy of the work of both the checker (k_f / k_t) and the original data handler (u_t). The known errors method could also be used to gather statistics on how long it takes to detect an error once it enters the database. The known errors method has the advantage over the duplicate performance method in that one does not have to enter data twice. It can also be used when the source of the error does not occur upon data entry such as an address becoming out of date. The drawback of the known errors method is that it may not be feasible to seed an working data set with known errors and a test data set might be treated in a manner that does not reflect actual working conditions with the production data.

Simulation

Once measurements have been obtained for the quality levels of the incoming transactions and the length of time a data error may exist undetected in the database, Haebich [1997] and Pierce [1997] have used mathematical equations and simulation to model the number of errors remaining in the database over time. Their work has shown that data quality levels can rise and fall as the database strives for equilibrium between the quality levels of new and updated records, record deletions, and error corrections. Their studies have shown that managers need to be aware that changes in the number of errors in the database may be due to the intrinsic movement of the equilibrium process and not due to any special external cause. In addition, such studies may be useful in gauging the impact of quality control programs or changes in the data processing operations on the long-term error rate in the database.

3. Removal Models

Sometimes the count data one has to work with is neither a census count nor a randomly drawn sample. Suppose a manager at an insurance company has been keeping track of the number of fraudulent claims found each day from a large batch of medical reimbursement forms that arrived at his office. A fraudulent claim can be considered a type of data error deliberately introduced by someone whose intent is to deceive. The manager is interested to know if the data he has gathered can tell him anything about the initial number of fraudulent claims contained in that original batch of records.

For solving the insurance manager's problem, this paper will demonstrate the use of removal models. These models were originally designed for wildlife population studies whereby undesirable animals such as rats were permanently removed from an area. This paper examines three types of removal models found frequently in the wildlife literature: Catch Effort, Maximum Likelihood, and Changes in Ratio to see how they might be applied to counting data errors. The author used Seber's [1973] book on estimating animal abundance as the main source for the description of the methods and the example data sets.

Catch Effort Methods

For the catch effort methods, the assumption is made that the size of a sample caught from a population is proportional to the effort put into taking the sample. The idea is that one unit of sampling effort is assumed to catch a fixed proportion of the population, so that if samples are permanently removed, the decline in population size will produce a decline in catch per unit effort. These methods only work if sufficient errors are removed from the data set, so that there is a significant decline in the catch per unit effort. A well-known type of catch effort method is Leslie's method (Leslie & Davis [1939]) which uses regression estimates. The notation presented here is reproduced from Seber's [1973] book.

Let N = the initial population size
Let n_i = the size of the ith sample removed from the population ($i = 1, 2, ..., s$)
Let x_i = the sum of the n_j from $j = 1$ to $j = i - 1$ ($i = 2, 3, ..., s+1$). Note $x_1 = 0$.
Let f_i = units of effort expended on the ith sample
Let F_i = the sum of the f_j from $j = 1$ to $j = i - 1$ ($i = 2, 3, ..., s+1$). Note $F_1 = 0$.

To use Leslie's method, one must be able to make a number of assumptions about the fraudulent claims in the data set. The first assumption one makes is that the population is closed. This means one must be looking at a data set where new fraudulent claims are not being added or removed for reasons other than being caught during the course of the study. The second assumption is that the sampling is a Poisson process with regard to effort. Mathematically this means that the probability of a given fraudulent claim being caught when the batch of claims is subjected to δf units of sampling effort is $k\delta f + o(\delta f)$. The symbol, k, represents the catchability coefficient and is assumed to be constant throughout the study and to be the same for each fraudulent claim. The units of effort are assumed to be independent and additive. The third assumption is that all the fraudulent claims have the same probability p_i (= $1- q_i$) of being caught in the ith sample (i.e. one cannot have inaccessible fraudulent forms that are left out of the detection process). Finally, Leslie's method assumes one is able to completely record and quantify the catch effort (number of claims inspectors, hours spent reviewing forms, types of inspection tools used, etc) into a standard unit.

Once these assumptions are reasonably satisfied, the insurance manager can then proceed to organize his data on the number of fraudulent claims found by inspectors on a given batch of insurance claims into the following table.

Date	**n_i** **number of fraudulent claims detected**	**f_i** **number of claim inspectors assigned that day**	**y_i (= n_i / f_i)** **catch per unit effort**	**x_i** **cumulative number of fraudulent claims caught**
May 23	7	8	0.875	0
May 24	6	8	0.75	7
May 25	3	3	1	13
May 26	6	8	0.75	16
May 27	3	5	0.60	22
May 30	6	8	0.75	25
May 31	6	8	0.75	31
June 1	3	5	0.60	37
June 2	5	7	0.71	40
June 3	5	9	0.56	45
June 6	4	7	0.57	50
June 7	0	1	0	54
June 8	1	2	0.5	54
June 9	3	6	0.5	55
June 10	2	5	0.4	58
June 13	3	5	0.6	60
June 14	4	8	0.5	63
This data is a modified version from one found in Seber [1973], contributed by DeLury [1947]				

The next step is to plot the catch per unit effort against the cumulative number of fraudulent claims found to see if the plot is reasonably linear. If a linear relationship is revealed, one then uses the usual regression techniques to obtain the estimate for k (i.e. $y_i = kx_i + e_i$).

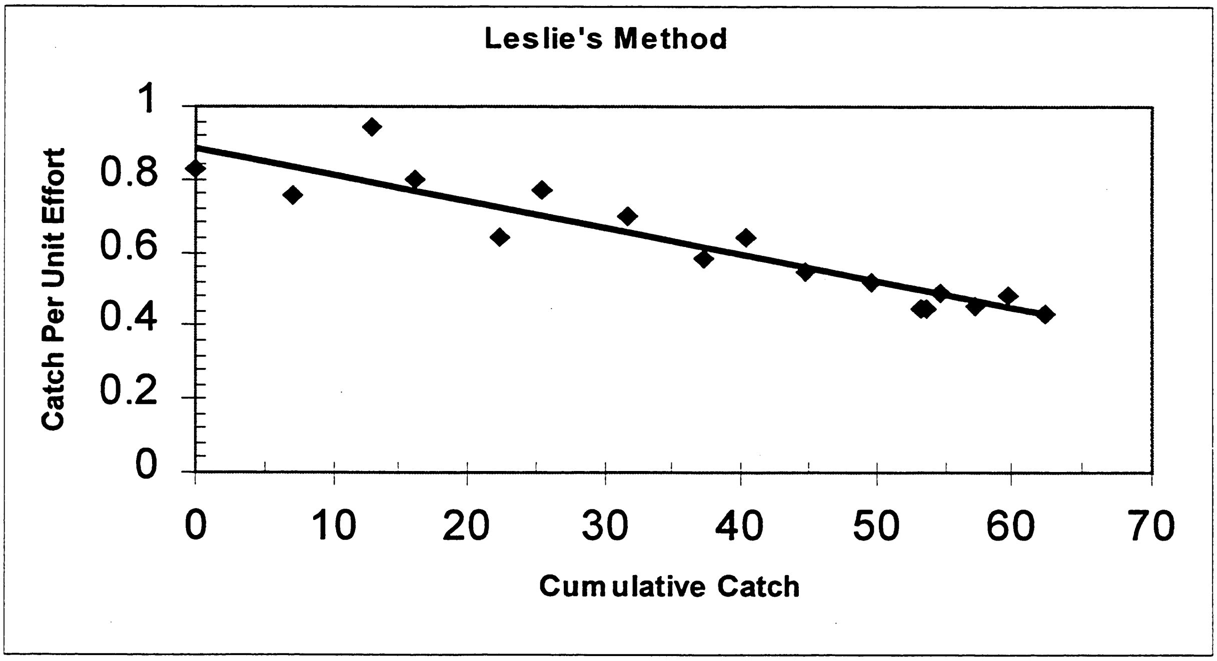
Note: k can be defined as the average probability that a fraudulent claim is discovered with one unit of effort. To estimate N, the initial population size, one uses the formula.

$$\tilde{N} = \bar{x} + \left(\frac{\bar{y}}{\tilde{k}}\right)$$

When s (the number of samples) is sufficiently large, the estimate for N is approximately normally distributed and the usual 100 (1-α) percent confidence interval can be applied using confidence coefficients from the t distribution. The formula for the variance of N is.

$$V[\tilde{N}] = \frac{\sigma^2}{k^2}\left[\frac{1}{s} + \frac{(N - \bar{x})^2}{\sum_{i=1}^{s}(x_i - \bar{x})^2}\right]$$

Using the fictitious insurance manager's data, the plot of the catch per unit effort against the cumulative number of fraudulent claims found does appear to be reasonably linear. In this example, the linear regression coefficient k equals .0079. The linear regression line has the form: $y_i = .9045 - .0079x_i$. The average catch effort per unit (y_i) is 0.613 and the average cumulative catch (x_i) is 37.06. Plugging this into the formula, one obtains an estimate for N of approximately 115 initial fraudulent claims [= 37.06 + (0.613 / .0079)]. The variance of N is approximately 390 and an approximate 95% confidence interval for N is (73, 157) (d.f. = 15).



In some situations, it is reasonable to assume that the same effort is used for each sample under similar conditions. In that case, p_i is constant and equal to some value p. Leslie's model can be adapted to this situation by setting each f_i equal to 1 unit and then solving using the usual formulas.

Maximum Likelihood Methods

Although maximum likelihood methods have been developed for the case of variable sampling effort (Seber [1973]), the regression methods are the best known. However in the case of constant sampling effort, Zippin [1956, 1958] obtained maximum likelihood estimates of N provided the following assumptions hold. (1) The population is closed. (2) The probability of capture in the ith sample is the same for each fraudulent claim exposed to capture. (3) The probability of capture, p, remains constant from sample to sample. Zippin calculated that under these assumptions, the joint probability function of the {n_i} and the maximum likelihood estimates for N and p are given by the following equations:

$$\frac{N!}{(N-x_{s+1})!\prod_{i=1}^{s} n_i!}\,p^{x_{s+1}}q^{sN-\sum_{i=1}^{s+1} x_i} \qquad \hat{N}=\frac{x_{s+1}}{(1-\hat{q}^s)} \qquad \frac{\hat{q}}{\hat{p}}-\frac{s\hat{q}^s}{(1-\hat{q}^s)}=\frac{\sum_{i=1}^{s}(i-1)n_i}{x_{s+1}}(=R\_say)$$

Graphs can be used to help facilitate finding a solution to the above equation. Seber has found that this equation has a unique solution for q in the range [0, 1] for 0 <= R <= (s - 1) / 2. When R > (s - 1) / 2 the above method is not applicable, but according to Seber, the probability of this happening will generally be small for large N and will decrease as the number of samples, s, increases.

For large enough N, one can assume asymptotic normality for the estimates and the confidence limits for N can be calculated in the usual manner using a variance for N based on the formula below.

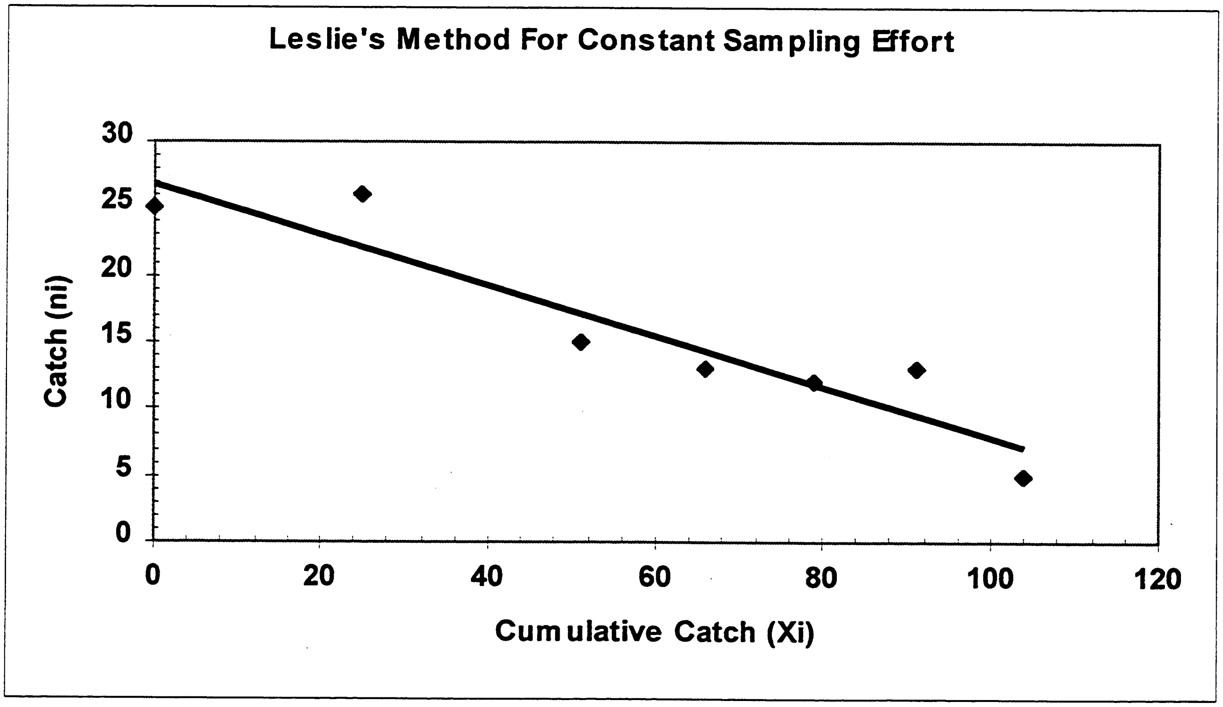
$$V[\hat{N}]=\frac{N(1-q^s)q^s}{(1-q^s)^2-(ps)^2q^{s-1}}$$

To demonstrate this method, suppose the insurance manager has another batch of insurance claims and a new set of catch data. In this scenario, the insurance manager feels that the sampling effort is constant over each trapping occasion for fraudulent claims.

Week (i)	i - 1	Catch (n_i)	Cumulative catch (x_i)	(i-1)n_i
1	0	25	0	0
2	1	26	25	26
3	2	15	51	30
4	3	13	66	39
5	4	12	79	48
6	5	13	91	65
7	6	5	104	30
Totals		**109**		**238**
This data is a modified version from one found in Seber [1973], contributed by Ricker [1958]				

Using Zippin's method, the estimate for R is found by dividing 238 by 109 to get 2.18. Using the graph in the appendix, one can see for s = 7, this works out to a value for p of 0.19 (q = 1 - p = 0.81 and (1 - q^7) = 0.77).

N is found by dividing the cumulative catch (109) by (1 - q^7) to obtain an initial fraudulent claims estimate of 142 (= 109 / 0.77), with a variance of 268 so an approximate 95 per cent confidence interval for N would be (110, 174). Applying Leslie's method to this data using (f_i = 1), one plots the linear regression line between $y_i = n_i$ (catch) and x_i (cumulative catch) and calculates for this data, k = 0.1895. Using Leslie's method, the formula also estimates an initial population of 142 fraudulent claims.



Open Population Removal Models

All the removal models discussed so far have assumed that the population is closed. However, one may face a situation when one wants to estimate the number of errors based on catch data in a database that is subject to the addition, updating, and deletion of records which in turn is affecting the number of errors existing in the data. For this situation one must consider the use of open population removal models. These models are complicated by the fact that one must be able to estimate the rate at which errors are being added to the system as well as the rate at which errors are being removed for reasons other than detection.

Seber [1973] and Fischler [1965] describe several open removal population models. One that may be of interest to the problem of counting data errors is a variation on Leslie's regression model. Leslie's regression model can be modified to handle the case where the addition of new errors and the removal of errors through detection are the only processes affecting the number of errors. To accomplish this, one needs to track r_i (the recruitment rate) which is the number of new errors that entered the database in the i^{th} interval. The basic formulas for estimating the initial population and its variance are the same. What changes are the formulas for calculating x_i and y_i. For period 1, the formulas for calculating x_1 and y_1 are:

$$x_1 = \frac{n_1}{2}$$

$$y_1 = \left[\frac{(2 - r_1)}{2}\right]\frac{n_1}{f_1}$$

For period 2, the formulas for calculating x_2 and y_2 are:

$$x_2 = 2\left(\frac{n_1}{2 + r_i}\right) + \frac{(2 - r_1)n_2}{2(2 + r_1)}$$

$$y_2 = \left[\frac{(2 - r_1)(2 - r_2)}{2(2 + r_1)}\right]\frac{n_2}{f_2}$$

For any period > 2, the formulas for calculating x_i and y_i are:

$$x_i = \left[\frac{2n_1}{(2+r_1)}\right] + \left[\frac{2n_2(2-r_1)}{(2+r_1)(2+r_2)}\right] + \ldots + \left[\frac{2n_{i-1}(2-r_{i-2})(2-r_{i-3})\ldots(2-r_1)}{(2+r_{i-1})(2+r_{i-2})\ldots(2+r_1)}\right] + \left[\frac{n_i(2-r_{i-1})(2-r_{i-2})\ldots(2-r_1)}{2(2+r_{i-1})(2+r_{i-2})\ldots(2+r_1)}\right]$$

$$y_i = \left[\frac{(2-r_i)(2-r_{i-1)}\ldots(2-r_1)}{2(2+r_{i-1})\ldots(2+r_1)}\right]\frac{n_i}{f_i}$$

Changes in Ratio Methods

Wild life studies have used for years a method that takes advantage of changes in observed sex ratios, age ratios, or marked-to-unmarked ratios to estimate population abundance. This method, known as Changes in Ratio, comes in a variety of forms depending on the sampling and population assumptions. For more information on the wide variety of applications using this techniques, see Seber [1973]. The basic idea behind this method can be demonstrated by the following simplified example.

Suppose a closed population of data contains two types of errors: records with out-of-date information (type 1 error) and records with information that has always been inaccurate (type 2 error). Due to the nature of these errors, it is difficult to find all instances of these errors in the database. The best the database administrator can do is to look over the records at periodic intervals and try to count as many errors as she can find. Suppose there is a differential change in the ratio of type 1 to type 2 errors between time t_1 and time t_2. An inspection of the database prior to time t_1 showed that there were 83 type 2 errors found for every 100 type 1 errors. After a clean up of the database takes place, a second record inspection is conducted and the new count at time t_2 reveals that there were now 53 type 2 errors found for every 100 type 1 errors. During the clean up, 248 type 2 errors were removed and 60 type 1 errors were removed. The question that the data administrator would like to know is what was the total population count of type 1 and type 2 errors in the database at time t_1.

Let X_i = number of type 1 errors in the population at time t_i
Y_i = number of type 2 errors in the population at time t_i
$N_i = (X_i + Y_i)$ (total population of errors at time t_i)
$P_i = X_i / N_i = (1 - Q_i)$. In this case, $P_1 = 83 / 183 = 0.4536$ and $P_2 = 53 / 153 = 0.3464$.
$R_x = X_1 - X_2$ (number of type 1 errors removed from the population, i.e. 248)
$R_y = Y_1 - Y_2$ (number of type 2 errors removed from the population, i.e. 60)
$R = R_x + R_y$ ($= N_1 - N_2$) (total number of errors removed from the population, i.e. 308)
Note: An addition can be expressed as a negative removal to the population.

Using this information, one can express P_2 = the number of type 1 errors in the population at time t_2 as

$$P_2 = \frac{X_2}{N_2} = \frac{X_1 - R_x}{N_1 - R} = \frac{P_1 N_1 - R_x}{N_1 - R}$$

Solving for N, one obtains the formula.

$$N_1 = \frac{R_x - RP_2}{P_1 - P_2} = \frac{248 - 308(0.3464)}{0.4536 - 0.3464} = 1318$$

If there were an estimated 1,318 total errors in the database at time t_1, then the number of type 1 errors at time t_1 must have been $X_1 = P_1 * N_1 = 598$ and the remaining 720 errors were type 2 errors.

Summary

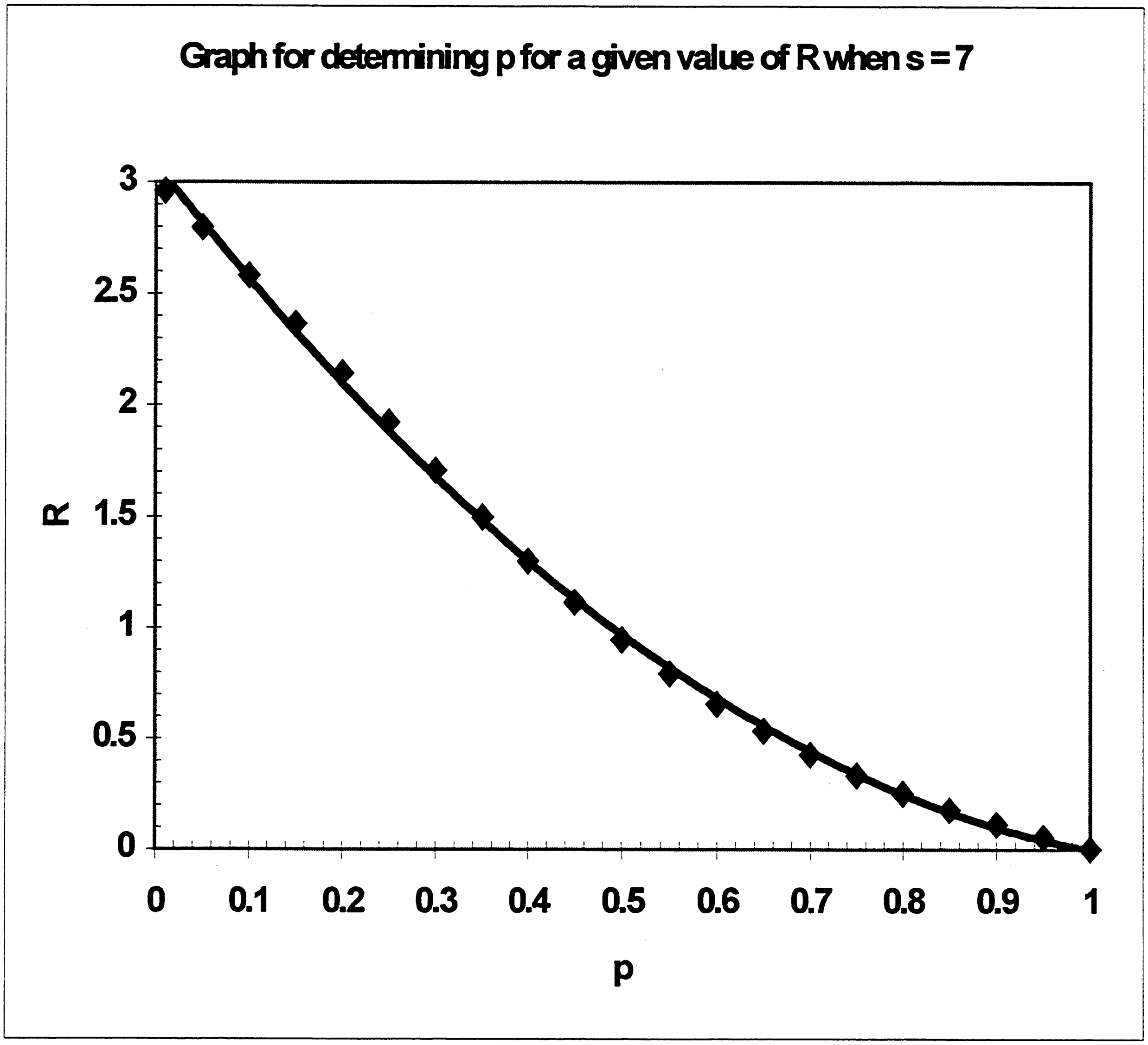
The counting literature spans across many disciplines and includes a wide variety of techniques. When trying to choose an appropriate technique for counting data errors, the author feels that there are three main criteria to consider: the nature of the errors, the nature of the database, and the nature of the desired count.

The nature of the error includes three sub-criteria: how easy is it to identify the error, the source of the error, and characteristics of the error. If the error is easy to identify, then the computer can be used to automate the detection, counting, and possibly the correction of the errors. Under these circumstances, census techniques should be used to get an accurate count of the errors. If the identification of the error requires human intervention, then the sampling or removal methods are a better choice than the census techniques. The source of the error is also valuable information. If the errors are occurring during the input process, sampling of the incoming data using either classical methods or the duplicate performance method may be useful in identifying the incoming error rate. However, if the data errors are occurring after data entry or one wants to verify the quality of the error detection efforts then using classical sampling methods or the known errors methods against the database itself may be appropriate. Finally the characteristics of the error affects the choice of counting technique. For instance if the error occurs randomly throughout the database, simple random or systematic random sampling techniques are appropriate. If the errors exhibit some underlying pattern, then cluster or stratified random sampling that take advantage of this pattern will yield a more accurate count.

The nature of the database itself will also affect the choice of counting technique. Here the criteria to consider are the size of the database, the amount of change taking place in the database, and the level of monitoring taking place in the database. If the database is large, sampling is generally more cost effective for counting errors than using a census technique. If the database is constantly changing, sampling (because it is completed faster than census) or removal methods that assume open populations may produce a more accurate count. The level of quality control for the records in the database will determine the amount and type of data available for analysis. If records are periodically sampled, then classical sampling techniques can be used to provide count estimates. If the number of errors captured is documented then removal techniques such as the changes in ratio method can be used to estimate the initial error count. If the level of detection (catch) effort over time is documented then catch effort (for varying levels of catch effort) or maximum likelihood methods (for constant levels of catch effort) may be appropriate.

The last set of criteria involves the nature of the count itself. The three main items to consider are the type of count, the time frame of the count, and the accuracy level of the count. In deciding what type of count is desired, one should consider if the metric should reflect the number of some given type of error in the database (or some subset of the database), or the number of errors found per record. In addition, one must decide if the metric should be expressed as a single raw number, proportion or percentage or as a confidence interval. Classical statistics provides several different methods for expressing each of these types of count. Another criteria to consider are the time frames of the count. If the goal is to track errors over time, then the quality control techniques are most appropriate. If the desire is to forecast future levels of errors in the database then a simulation technique may be useful. The last criterion involves how much accuracy is required for the count. If accuracy is extremely important and every record must be checked then the one number estimate from the census bureau may be useful. Cycle counting techniques from the inventory literature may also be appropriate.

For those whose responsibility it is to quantify the error levels in their databases, it is hoped that some of the techniques presented here prove useful for a given situation. It is important to note that many of the techniques that were described are merely representative examples of the available methods in a given category. Time and space prohibit a full listing of all the variations of the methods available. In closing, the author welcomes comments on any of the counting techniques presented, as well as news of additional models that were omitted. In terms of future work, the author would like to investigate how successful the implementation of these techniques are in practice. In particular, it would be interesting to know when one has a choice between 3 or 4 different techniques, which one is the most accurate or reliable in its estimation of the actual count of errors in the database.


Graph for determining p for a given value of R when s = 7
3
2.5
2
1.5
1
0.5
0
R
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
p

Bibliography

Arkin, H., *Handbook of Sampling For Auditing and Accounting*, McGraw-Hill Co., 1963.

Backes R.W., "Cycle Counting - a Better Method for Achieving Accurate Inventory Records", *Management Accounting*, January 1980, pp. 42 - 46.

Berry, M. J. A, Linoff, G., *Data Mining Techniques for Marketing, Sales, and Customer Support*, John Wiley & Sons, New York, 1997.

Cochran, W.G., *Sampling Techniques*, 2nd Edition, John Wiley & Sons, Inc, 1953.

DeLury, D.B., "On the Estimation of Biological Populations", *Biometrics*, 1947, Vol. 3, pp. 145 - 167.

Fellegi, I. P., Holt, D., "A Systematic Approach to Automatic Edit and Imputation", *Journal of the American Statistical Association*, March 1976, Vol. 71, No. 353, pp. 17 - 35.

Fischler, J., "The use of catch-effort, catch-sampling, and tagging data to estimate a population of blue crabs", *Transactions of the American Fisheries Society*, 1965, Vol. 94, pp. 287 - 310.

Haebich, W., "A Quantitative Model to Support Data Quality Improvement", *Proceedings of the 1997 Conference on Information Quality*, October 1997, pp. 194 - 208.

Leslie, P. Davis, D., "An attempt to determine the absolute number of rats on a given area", *Journal of Animal Ecology*, 1939, Vol 8, pp. 94 - 113.

Montgomery, D.C., *Introduction to Statistical Quality Control (2nd Edition),* John Wiley & Sons, New York, 1991.

Naus, J. I., *Data Quality Control and Editing*, Marcel Dekker Inc., New York, 1975.

Naus, J. I, Johnson, T. G., Montalvo, R., "A Probabilistic Model for Identifying Errors in Data Editing", *Journal of the American Statistical Association*, December 1972, Vol. 67, No. 340, pp. 943 - 950.

Pierce, E., "Using P-Charts to Track Data Quality", *Proceedings of the 1997 Conference on Information Quality*, October 1997, pp. 170 - 186.

Redman T.C., *Data Quality - Management and Technology*, Bantam Books, 1992.

Ricker, W.E., "Handbook of Computations for Biological Statistics of Fish Populations", *Bulletin of Fisheries Research Board of Canada*, 1958, Vol. 119.

Seber, G. A. F., *The Estimation of Animal Abundance and Related Parameters*, Hafner Press, New York, 1973.

Strayhorn, J. M., "Estimating the Errors Remaining in a Data Set: Techniques for Quality Control", *The American Statistician*, February 1990, Vol. 44, No. 1, pp. 14 - 18.

Toney, S. R., "Cleanup and Deduplication of an International Bibliographic Database", *Information Technology and Libraries*, March 1992, pp. 19 - 28.

West, M., Winkler, R. L., "Data Base Error Trapping and Prediction", *Journal of the American Statistical Association*, December 1991, Vol. 86, No. 416, pp. 987 - 996.

Wilson, J. M., "Quality control methods in cycle counting for record accuracy management", *International Journal of Operations & Production Management*, July 1995, Vol. 15, No. 7, pp. 27 - 40.

Wright, T., "Sampling and Census 2000: The Concepts", *American Scientist*, May-June 1998, Vol. 86, pp. 245 - 253.

Zippin, C., "An Evaluation of the Removal Method of Estimating Animal Populations", *Biometrics*, June 1956, Vol. 12, pp.163 - 189.

Zippin, C., "The Removal Method of Population Estimation", *Journal of Wildlife Management*, 1958, Vol. 22, pp. 82 - 90.