METHODOLOGY CORNER

Approaches to Cleaning Data Sets: A Technical Comment

DIANA Y. BARHYTE • LYND D. BACON

Although extensive research design and methodology literature (Polit & Hungler, 1983) is available to nurse researchers, little material exists on some of the more technical aspects of research, which are important in determining the quality of research findings. Beginning researchers usually learn the "tricks of the trade" through an apprenticeship. One of these activities is euphemistically called "cleaning the data." The goal is to obtain a set of data that contains a minimum of errors resulting from human factors in coding and data entry.

There are several methods commonly used by experienced researchers in cleaning data. Several assumptions are made in the following discussion. It is assumed that data are coded from some type of form and are to be entered into a computer data file. For example, the data might be responses to a structured questionnaire that had been sent for keypunching for entry onto computer tape. A second assumption is that the format for each case (record) is fixed; that is, the location and width of each variable does not vary across cases. For example, variable X is always in columns 5 through 7 and is a whole number.

The first method used by researchers for cleaning data consists of examination of listings of the records. In fixed format the field containing a variable can be inspected for alignment across the cases, assuming there is a minimum of one blank between fields. For example, if variable A is contained within the file columns 15 through 20, then an inappropriate entry in column 21 would result in the right margin jutting out. The misalignment easily can be found when the data are visually inspected. It is helpful to draw vertical lines at the beginning and end of the fields, or a computer ruler can be used. A computer ruler is an inflexible straight edge that has markings for columns rather than parts of inches: in addition, the center portion magnifies the numbers being read.

Visual inspection of the fields insures that each field either does or does not contain appropriate data. It does not, however, answer the question of whether the value listed for a variable is correct: for example, the correct value for variable A is 555, but it is listed as 333. The value of visual inspection is that a researcher can see if there are alignment difficulties prior to proceeding with the other methods of cleaning data. An illustration of the visual inspection method is given in Figure 1.

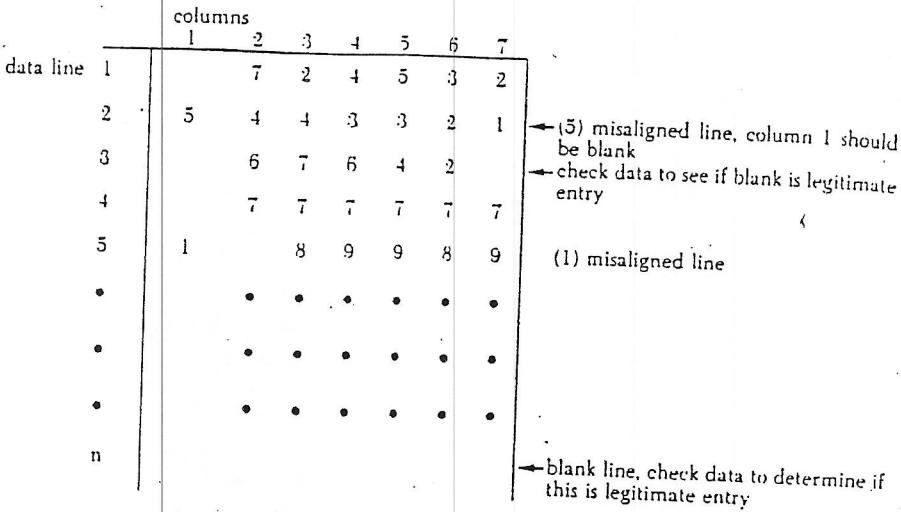
The second method for cleaning is to do frequency distributions for

DIANA Y. BARHYTE, PhD, RN, is director of Nursing Systems Management Program, Rush-Presbyterian-St. Luke's Medical Center, Chicago, IL.

LYND D. BACON, PhD, is the project director, Nursing Administration, Northwestern Memorial Hospital, Chicago, IL.

Accepted for publication March 19, 1984.

Figure 1. Illustration of visual inspection for misaligned data



venient since all variables have a legitimate range of values. Prior to doing the frequency distributions, the researcher determines the expected upper and lower limits for the variables. For example, the expected age limits for employed nurses would be set at 20 and 65 years respectively, since it would be unexpected for an employed nurse to be less than 20 years of age or more than 65 years old. In addition, some data are discrete in the sense that the precision of measurement limits the values that may occur. For example, time measured in whole hours should not yield data values that are in fractional portions of hours. The illegitimate values will be detected by the frequency counts for each variable.

Cross tabulations of measures also may be used when certain combinations of values for two variables are illegitimate. An example of this is when days of the week and calendar date are encoded. Another example would be that in a sample of females and males, only females should be coded as mothers and males as fathers. There is a unique one-to-one correspondence that defines legitimate pairs of these data, and cross tabulation can detect violations of this correspondence. As in the case of visual inspection, where a value of a variable can be recorded in a legitimate field yet be incorrect, neither frequencies, cross tabulations, nor visual inspection reveal an incorrect value for a given variable.

The preceding two methods focus on the configuration of the values assigned to the variables. The third method focuses on the entire record for each subject. It involves examining a sample of records, and is therefore a compromise between the effectiveness of verifying each value encoded for a variable and the efficiency involved in doing so. Before beginning this method of verification, the researcher usually makes two decisions. The first decision is to determine the number of error-containing records that will be accepted before all records would be verified. The second decision is to determine the size of a sample of records that should be drawn and verified.

The approach is essentially a lot sampling plan for quality control. The records of a data set may be considered a population and the researcher wishes to make an inference concerning the number of "defective" (i.e., error-containing) records it contains. Assuming that the errors are both randomly and independently distributed across records, each record randomly drawn can be thought of as a trial in a binomial experiment, that is, the records are samples from a Bernoulli process (Hays, 1973).

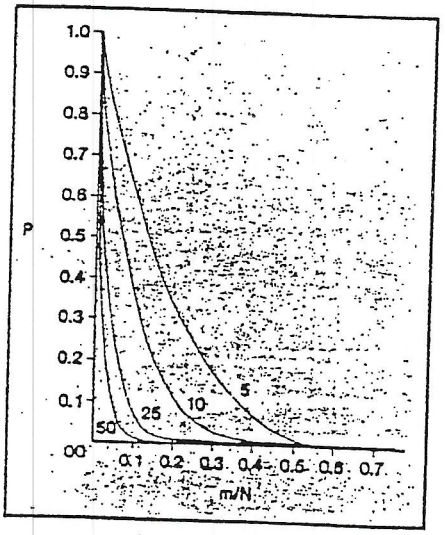
For discussion purposes, assume that the researcher decides not to accept any incorrect records. Thus, the first decision is that the acceptance number is zero. This may seem to be a rather conservative decision. However, considering that in scientific measurement error variance is usually assumed not to originate in measurement instruments, it seems reasonable.

The second decision using this approach is the size of the sample to draw. The probability of drawing an error-free sample may be calculated using the binomial probability distribution (Hays, 1973). In the present context, the probability of drawing a sample of size n containing only error-free records from a data set of N records, m of which contain errors, is $p=(1-(m/N))^n$.

In other words, the two decisions made by the researcher can be represented in a formula. The first decision, number of error-containing records that the researcher will accept before verifying every record against the raw data, is reflected in the ratio m/N. That is, the ratio m/N is the proportion of error-containing records in the entire data set or population of records. The numerator m is the number of records that have one or more encoding errors and can have a value range of zero to the total number of records in the data set. The denominator N is the total number of records in the data set; that is, the size of the population.

The equation shows that the probability of obtaining an error-free sample depends on the size the sample drawn. The exponent n stands for the sample size and reflects the second decision that the researcher needs to

Figure 2. Operating characteristic curves for sampling plans having an acceptance number of zero. p is the probability of drawing an error-free sample of records, and is plotted as a function of the ratio m/N, the proportion of error-containing records in the data set. The parameter is n, the size of the error-free sample drawn.



make. Before continuing, note that regardless of the sample size when $m/N=0$, $p=1$, since an error-free data set (population) can only yield error-free samples. To help decide the appropriate sample size, a researcher can plot the probabilities that would result for various sample sizes. The plot could be a graphic representation of a family of operating characteristic curves (OCCs) for various sample sizes. Figure 2 is an example of OCCs for sample sizes of 5, 10, 25, and 50 records.

Figure 2 shows that the probability of a researcher making a correct decision about whether there are errors increases with sample size. However, it is only when p equals one or zero that a correct decision is guaranteed by lot sampling plans.

Each of the three preceding methods may be used in examining a data set for encoding errors. The methods vary in their effectiveness of uncovering errors and the amount of time needed by the researcher to complete the task. In addition, there is a fourth method that is effective in detecting errors and efficient of the researchers' time. It is called multiple entry.

The method of multiple entry consists of constructing two duplicate data sets for comparison. Each data set is keyed by a different individual so that the problem of individual bias toward certain types of keystroke errors is minimal. Once the data sets have been created, the values for each variable of the corresponding records are compared for discrepancies. Discrepant records are then verified against the original forms and the appropriate corrections made. The advantage of this technique is that it is economical in terms of the researcher's time but still allows a value-by-value examination.

The comparisons of the separately keyed data sets can be done several ways. Two computer printouts of the listed cases can be aligned and then held up to a light source. By reading through the two printouts, discrepancies can be detected. Another approach is to use a computer program to compare the corresponding values of the two data sets.

Summary

Since each of the methods described has its costs and benefits, more than one method should be used. The combination of methods helps the researcher to obtain an error-free data set. Ideally, value-to-value verification for all data sets is preferred but the multiple entry method is more efficient for large data sets. Each researcher has a definition of large and small data sets. One definition of a large data set is one containing more than 250 records. Such a definition will not be appropriate for all investigators. Therefore, each researcher should balance the allocation of resources needed for a method against the confidence required for the data. NR

References

HAYS, W. L. (1973). *Statistics for the social sciences*. New York: Holt, Rinehart & Winston.

POLIT, D., & B. HUNGLER. (1983). *Nursing research: Principles & methods*. (2nd ed.). New York: J. B. Lippincott.

Calendar of Events

(Continued from p. 61)

June 16–21, 1985
ICN 18th Quadrennial Congress, "Nurses as a Social Force," will be held in Tel Aviv, Israel. For information contact: Greta Green, Kenness International, One Park Ave., New York, NY 10016; (212) 684-2010 or (800) 235-6400.

June 17–27, 1985
Cornell University will present its 25th Annual Health Executives Development Program in Ithaca, NY. For information contact: Health Executives Development Program, N222 Martha Van Rensselaer Hall, Cornell University, Ithaca, NY 14853; (607) 256-7770.

June 18–20, 1985
The 4th Conference on Cancer Nursing Research will be held in Honolulu, HI. For information contact: Ruby H. Rutherford, RN, National Representative, Medical Affairs, American Cancer Society, Western Area Office, 5660 S. Syracuse Circle, Englewood, CO 80111; (303) 773-1502.

June 30–August 22, 1985
"Human Sexuality," Summer Study Abroad Program—1985 in The Netherlands, sponsored by the Department of Health Education, New York University. Application deadline is May 10, 1985. For information contact: Professor Deryck Calderwood, PhD, Human Sexuality Program, 715 Broadway, New York, NY 10003; (212) 598-3925.

July 14–17, 1985
"High Tech/Intensive Caring," a special meeting celebrating the 10th anniversary of *MCN, The American Journal of Maternal/Child Nursing*, will be held in Baltimore, MD. For information contact: 10th Anniversary of *MCN*, G & T Management, 211 E. 43rd St., Suite 1601, New York, NY 10017; (212) 867-4480.