BIOMETRICS 35, 683–686
September, 1979

Detecting Systematic Errors in Multi-Clinic Observational Data

NANNY WERMUTH

Johannes Gutenberg Universität, Mainz, Federal Republic of Germany

W. G. COCHRAN

Statistics Department, Harvard University, Cambridge, Massachusetts 02138, U. S. A.

Summary

In multi-clinic studies it is hard to maintain a uniformly high quality of measurement and coding. Systematic errors almost always occur, in spite of the best of intentions and the most rigid protocols. It is the statistician's responsibility to plan for the detection of these errors, as well as to try to avoid them and not be misled by them. The practice of examining the univariate and multivariate sample frequency distributions of the variables under study, with an eye open for anything that looks puzzling, can be very helpful in detecting and trying to correct systematic errors that would bias the analysis. Examples are given from a 21-clinic study on pregnancy and child development.

1. Systematic Errors

In research on variables that may be related to morbidity rates, cooperative observational studies among a number of clinics are often necessary in order to attain a desired sample size and to sample a more extensive population. A constant problem in multi-clinic studies is that of maintaining a consistently high quality of measurement and coding from clinic to clinic and of avoiding systematic errors. Misconceptions by the statistical analyst about what has been observed, measured, or defined, can have serious effects on the interpretation of the results.

2. Use of Frequency Distributions

As a precaution against misunderstandings of the data and undetected systematic errors, examination of the frequency distributions of all variables relevant to the relationships or hypotheses under study is often recommended before any analysis is undertaken. In particular we recommend the following steps:

(1) Examine all univariate distributions for the sample data.

(2) If results are available also for a broader population close to the target population, compare the frequency distributions in the sample and the broader population. Discrepancies may suggest something wrong with the sample measurements that needs investigation, or may give clues as to how well the sampled and target populations agree.

Key Words: Frequency distributions; Missing observations; Multi-clinic studies; Observational studies; Systematic errors.

(3) Compare the sample frequencies for each clinic, looking for clinics that appear to be outliers.

(4) If a substantial proportion of the measurements are missing for any variable x_i, investigate the reasons why. In this connection it may help to compare the distributions of other variables in the subsample in which x_i was measured and in the whole sample, as will be illustrated (Example 7).

(5) Proceed similarly for two-way distributions for the main variables under study.

In examining these frequency distributions, look particularly for anything puzzling or surprising. Is the explanation a systematic error in measurement, or something wrong with a definition? Can this error be corrected for the data, or will new measurements be necessary? It is at this stage that a pilot study can be most useful. Some pilot work is always advisable in multi-clinic studies, especially in checking the clarity and understanding of questions, but its size and extent are matters of judgment. A small study may not reveal discrepancies and a large one is costly and slows up the main study.

3. Examples

The following examples illustrate the use of frequency distributions in seeking out systematic errors. The examples were taken from a collaborative study, in 21 clinics, of the relations between events in pregnancy and child development (DFG-Forschungsbericht 1976). The study began in 1964 in the Federal Republic of Germany. All women entered the study during their first trimester of pregnancy, returned for repeated examinations, and kept diaries on the course of their pregnancies and on the development of their children up to the age of three years whenever possible. For a report of the study, 478 classified variables for 7,870 women were available for analyses. The frequency distribution checks in the following examples are classified into univariate checks, bivariate checks, and checks involving missing values.

3.1. Univariate Checks

Example 1. In the case of a (0,1) variable, the sample frequency distribution reduces to a single proportion unless we are interested in the order in which 0's and 1's appear. Two things were puzzling about the reported frequency of women who had nephritis during the year prior to the onset of pregnancy. First, there were no missing values, an answer being recorded for every woman in the sample. Did all women in fact know whether they had nephritis during the past year? Secondly, the proportion of nephritis cases in the sample was much lower than the prevalence rates reported in other studies.

The explanation here lay in the question from which the data were coded. This question did not ask directly about nephritis, being instead: "Which illnesses did you have during the past year?" Direct questions about the specific diseases, with possible answers "Yes," "No," "Don't know," are necessary in attempts to estimate prevalence rates. This is the type of issue that can be picked up in an initial pilot study before the form of the questions is fixed.

Example 2. The sample frequency distribution of the weights of the mothers prior to pregnancy, recorded in kg, was bimodal, showing one region of high frequencies between 50 and 70 kg, and another between 100 and 140 kg. It was clear that many women had stated their weights in lbs, contrary to instructions. Detection and correction of the weights that were recorded in lbs was not difficult, because in doubtful cases there were other weight records available from the repeated examinations during pregnancy.

Example 3. Each mother was asked at what age her child started to speak two-word sentences. This question, if posed correctly, meant "At what age did the child not only speak single words but combine at least two words, like 'Mama, drink,' meaning 'Mama, I want a drink.' " When the responses by clinics were compared, one of the 21 clinics appeared to be an outlier. In this clinic the ages for two-word sentences were two years or more, as against values below two years in all other clinics. Investigation showed that in this clinic the question had been interpreted as asking for the age at which the child began to speak grammatically correct subject-predicate sentences. The answers of this clinic were subsequently omitted.

3.2. Bivariate Checks

Example 4. One part of this study was a comparison of congenital defects in the child for mothers who had a subclinical infection of rubella (German measles) during the first trimester and mothers who had not. To obtain a criterion for the rubella status, titer values the rubella at the beginning and end of the first trimester were recorded on a scale: $0, 2^2, 2^3, 2^4, \ldots, 2^{12}$. A mother was coded as having had rubella if the titer value had increased by at least two steps from the beginning to the end of the first trimester.

However, a bivariate classification of the mother's initial titer value against her rubella status (yes, no) caused this definition to be reconsidered. A substantial number of women had initial rubella titer values so high that an increase by two or more units on the scale was either impossible or very unlikely. Moreover, a high initial titer was an indication of a very recent rubella infection. Mothers with such high initial titers were treated in some analyses as a third comparison group.

Example 5. Another bivariate table revealed a puzzling error in coding. Pregnancies were classified by duration and also by whether the pregnancy ended in an abortion or not. For a number of women with more than 260 days pregnancy, the outcome was coded as an abortion. If the code for duration was correct, such cases should have been coded as stillbirths: after 190 days of pregnancy, a fetus is generally regarded as capable of living. Since only a few cases were involved, correct codes were obtained by going back to the individual case records.

3.3. Checks Related to Missing Data

Example 6. Starting in 1968 all newborns in the study were to be examined for antibodies to rubella, the results being stated as titer-values. Examination of the titer-values for individual clinics showed that (1) some clinics had high missing-value rates for this variable, and (2) in these clinics the recorded values were in general greater than in other clinics. Why? Investigation showed that in these clinics the antibody status of the newborn was determined only when the physician saw a specific reason for it, such as a typical malformation in the body, or a high titer-value or observed infection in the mother. Inclusion of the incomplete data from these clinics in studies of factors related to the child's antibody status might result in biases. Consequently, the data were not used for this purpose but instead for investigating when a high titer-value of the mother were transferred to the child. For this analysis it was judged that the biases would be smaller.

Example 7. One statistical analysis dealt with the relation between the still birth rate and presence or absence of the symptoms edema, hypertension, and proteinuria in the mother at the time of birth.

However, no proteinuria was recorded for a number of mothers. On comparing the subsample in which all three study variables were measured with the complete sample, the relative frequencies of edema and hypertension, both singly and together, were found to be much lower in the subsample. Evidently, proteinuria had not been measured in a number of mothers with definite edema and hypertension symptoms. What had happened was this: When a woman in poor condition came to the clinic close to the time of birth showing either or both edema and hypertension, some physicians decide to induce labor immediately and not waste time by determining the amount of protein in the urine. In order to avoid these missing values, the symptom value for a variable at the last examination prior to birth was used if the value at birth was not available.

In conclusion, the preceding errors are all elementary, and might scarcely be worth writing about except that they can easily be overlooked by the consulting statistician who believes without checking what he or she reads or is told about the way in which the data have been measured. Moreover, visiting each clinic to inspect the measurement process in action, though highly desirable as a check, is often not feasible. Scrutiny and simple analyses of frequency distributions, as illustrated, will help to detect and correct some systematic errors in the original data.

Résumé

Dans la recherche collaborative entre cliniques, il est difficile de s'assurer que les mesures qu'on va analyser sont de bonne qualité et ont le même sens dans toutes les cliniques. A ce but, il est utile à examiner, avant l'analyse, les distributions de fréquence des variables principes, en cherchant des erreurs de mesurage ou de classification qui pourraient biaiser les résultats. Le papier donne des exemples de ce pratique, priées d'une recherche entre vingt et un cliniques, sur la grosseuse et le développement de l'enfant.

Reference

DFG-Forschungsbericht (1976): *Schwangerschaftsverlauf und Kindesentwicklung*, Harald Boldt Verlag, Boppard.

Received August 1978; *Revised April* 1979