QUALITY CONTROL OF STATISTICAL REPORTS

Paul S. Levy and Monroe G. Sirken, National Center for Health Statistics

1. Introduction

The National Center for Health Statistics (NCHS) develops and maintains systems capable of providing reliable, general purpose, national, descriptive health statistics on a continuing basis and publishes these statistics for the use of the health industry and related industries, both public and private [1]. Examples of such data systems are the national vital statistics of births, deaths, fetal deaths, marriages and divorces; sample surveys linked to birth and death records; a continuing nationwide survey of households by means of interviews; a series of national surveys based on physical examinations of samples of the population; surveys of hospitals and other health care facilities; as well as other periodic and ongoing surveys.

The principal form of output of the Center's work is published statistical reports. These come out in several series, one of which is entitled Vital and Health Statistics. This series often is referred to as the "rainbow series" since each data system has its own series of reports with distinctively colored jackets. In this series, some 35 to 40 substantive statistical reports are produced every year covering various aspects of the data collected in the systems mentioned above. Each substantive report contains a text which analyzes the data presented in a set of statistical summary tables, which are subject to sampling and measurement errors, estimates of which are present in the appendix. In order to ensure that the technical material presented in these reports meet specified standards and that the statistical statements made in the text are valid, a surveillance program in the form of a standardized procedure for reviewing such reports has been developed within the Office of Statistical Methods (OSM), which is the primary statistical support group within NCHS. The purpose of this article is to describe some of the main features of the quality control program for published statistical reports.

2. General Format of the Program

It should be mentioned that this surveillance program is not yet fully operational, and our experience is derived primarily from pilot projects. We have begun this program on a limited basis and expect it to become fully operational within the next several months.

At present, most of the reviews are being performed ex post facto, i.e., after the reports have been published. While this is not optimal, publication schedules do not allow these reports to be delayed too long for a statistical review. It is hoped, however, that as we gain experience and familiarity with the review procedures, and as we recruit additional personnel to do the reviewing, we can undertake to perform speedy reviews of each report prior to its publication.

As it stands now, however, a published report would be assigned for review to a junior or mid-level mathematical statistician within OSM soon after it was published. The review procedure which will be described below in more detail consists of two general parts. The first part deals with the review of the technical material, tables and figures in the report, while the second part deals with the review of statistical statements made in the report. An instruction manual has been drafted which guides the reviewer, step by step, in the procedure and ensures that all reviews are performed according to protocol [2].

When the procedure is completed, the findings are first reviewed by the senior statistician responsible for this surveillance program, and then they are discussed with the author of the report and/or the director of the subject matter division responsible for the report.

3. Review of the Technical Material, Tables and Figures

The purpose of this phase of the procedure is to ensure that (1) the report describes the essential design and estimation features of the data collection system, and (2) all statistics presented in the substantive tables meet prescribed standards of accuracy and precision, and have sampling errors available for them in the report.

In NCHS reports, the technical material describing the design and estimation features of the data collection system often is presented in a technical appendix but also may appear in the main part of the text. Specific items comprising this technical material which we feel should be adequately described in the report are the universe, frame, number of primary sampling units (PSU's), stratification, clustering, data collection and processing procedures, quality control procedures, estimation methods, methods of obtaining sampling and measurement errors, etc.

Standards for these items are given in the instruction manual, and the reviewer using a checklist systematically goes through the report and notes whether each technical item on the checklist meets the standards given in the instruction manual. For example, the description of the sampling frame in a substantive report would meet the standard set for it if it clearly states what are the enumeration units and elements, and if it makes reference to the NCHS publication describing the frame in detail (should such a publication exist).

Each substantive table and figure is checked for purposes of determining whether the statistics meet specified standards of precision and accuracy, whether the reader has sufficient

information in the report to determine the sampling error of every statistic presented in the substantive table and whether the technical terms used in the table titles are defined clearly and accurately in the report.

4. Review of Statistical Statements

4.1 Overall Objective

The second major component of the review procedure is the review of the statistical statements made in the text of the report. Its purpose is to ensure that the inferences made in the statistical statements are based on sound statistical methodology and judgment.

The primary objective is to estimate the proportion of statistical statements made in the report that are valid, invalid or untestable. In order to achieve this objective, it was necessary to develop a methodology which will be described in the remainder of this section.

4.2 Statistical Statements

4.2.1 Definitions

We use the following as a working definition of a statistical statement:

A statistical statement is any phrase, clause, sentence or combination of words that makes an inference from sample observations about characteristics of a population. A statistical statement is valid if the inference made in the statement is justified statistically. If the inference is not justified, the statement is said to be invalid. If neither of these decisions can be made, it is said to be untestable.

A statement that is untestable may be adequate or inadequate. It is considered adequate if the inference made in it is clear but there is no usable statistical procedure for validating the statement. It is considered inadequate if the inference made in the statement is unclear.

For these definitions to be applied with any reproducibility, some objective criteria were decided upon for identifying statistical statements within the text of a report and for declaring a specific statistical statement valid, invalid, or untestable.

4.2.2 The identification and classification in a text

The main criterion for deciding whether a sentence, phrase, clause or group of words constitutes a statistical statement is whether an inference is drawn from a sample to a population. If more than one such inference is drawn, then the group of words would constitute more than one statistical statement. Using this criterion, a reviewer with a little instruction and practice can identify statistical statements in a report with reasonable reliability.

Since the basic characteristic of a statistical statement is that it draws a statistical inference, and since the tools that the reviewer uses to judge the validity of the inference depend on the type of inference implied in the statement, we have attempted to classify statements according to type of inference and to specify methods for testing each type of statement. After some effort, we have arrived at the following taxonomy of types of statistical statements:

Type 1. Quotation of Estimates. These are statements which characteristically involve only one subdomain and are generally, as the name implies, simple quotations of estimates. For example, the statement "On the basis of examinations, approximately 24 million U.S. children aged 6-11 years averaged an estimated 1.4 DMF per child" [3]. A Type 1 statistical statement is considered valid if its coefficient of variation is below the tolerance set by the appropriate subject matter division.

Type 2. Simple Comparisons. These are statistical statements in which comparisons are made between two domains. These statements are divided into two subtypes given below:

Type 2A. Simple Comparisons of Equality. These are statements which draw the inference that the level of a characteristic in one domain is different from that in another domain. For example, "White children had somewhat better levels of oral hygiene than Negro children. As a result, the average OHI-S (Simplified Oral Hygiene Index) was 1.41 for the former and 1.66 for the latter"[4]. A statement of this type is considered valid for our purposes if the difference between the two domains with respect to the quoted statistics is significant at the 5 percent level of significance as determined by the usual test of a difference between two means or proportions.

Type 2B. Simple Comparisons Involving Magnitude. This type of statistical statement is of the form "X_1'" is r times as large (or as small) as "X_2'" where X_1' is the estimated level of a characteristic in one domain, X_2' is the estimated level of the same characteristic in another domain and r the estimated ratio of X_1' to X_2'. This type of statement is considered valid if (1) r′ meets the standards of precision given for Type 1 statistical statements and (2) the difference between X_1' and X_2' is significantly different from zero, as defined in the discussion of Type 2A statements.

Type 3. Comparisons Involving More Than Two Domains.

Type 3A. Comparisons Among Several Subdomains Within a Domain. These statements have the general form "within domain A, there were differences in the level of characteristic X among the subdomains A_1, A_2, and A_k." In judging the validity of such statements, a multiple comparison test based on the Bonferroni inequality is used [5].

Type 3B. Comparisons Between the Corresponding Subdomains of Two Domains. The inference in this type of statement is of the form "within each subdomain, the level of characteristic X is higher in domain A than in domain B. To judge the validity of this type of statement, differences between domain A and B in the level of characteristic X are tested for each of the k subdomains implied in the statement. If all k of these differences are statistically significant, the statement is considered valid. If r (< k) of these differences are significant, the number, r, is examined in a sign test table based on k signs. If r is greater than or equal to the critical value of the sign test statistic (for the 5 percent level of significance), then the statement is considered valid. Otherwise, it is considered invalid.

Type 4. Statements of Statistical Relationship.

Type 4A. Statements of "Trend". This type of statement makes an inference that there is an association between two variables whose domains of definition are on interval or ordinal scales. For example, "the relationship is an inverse one with the proportion of men and women of all races who need to see their dentist at an early date decreasing sharply with rising levels of yearly income [6]. The validity of this type of statement is judged by testing whether the linear component representing the relationship is significantly different from zero. Because of the complexities of data collected from complex surveys, a modified estimate of the linear relationship and a modified significance test was devised for this purpose [3].

Type 4B. Statements of Association Other Than Trend. This category would include any statement implying a relationship between two or more characteristics that cannot be interpreted as a Type 4A statement. Most of these statements imply a relationship between two variables, one or both of which are attribute or categorical variables. Other statements of this type express rather complex types of association. An example of such a statement is "The difference in the direction of the trend (in need for dental care) between Negro men and the other sex-race groups is not due to differences in the age composition of the various educational attainment groups." [6] This type of statement is often difficult to interpret and we have no standard protocol for testing statements of this type. Often what appears to be a Type 4B statement can be interpreted as another type or broken up into one or more types for which we have standard tests. However, many of these Type 4B statements elude interpretation or validation.

These are the main categories of statements that we have classified. From preliminary studies, we found that these four types account for almost 75 percent of the statements found in NCHS statistical reports, and that approximately two-thirds of all statistical statements made are testable. As we gain experience with this procedure, we expect to develop a more refined taxonomy of statistical statements and to develop further methodology for testing these statements.

4.2.3 Sampling of statistical statements. A main objective of the review of statistical statements is to obtain an unbiased estimate of the proportion of valid, invalid, and untestable statistical statements made in the report. Since resources do not permit the testing of every statistical statement, our estimates are based on a probability sample. The choice of what kind of sampling design to use presented us with an interesting statistical problem which is the subject of a separate article [7]. In summary, after experimentation with several types of sampling designs, we have chosen for reasons of logistics and cost efficiency, one which uses lines of text as the enumeration units, statistical statements as the elementary units and a conventional enumeration rule to link the enumeration units with the elementary units.

Once the required sample size, ℓ, of enumeration units is specified, a systematic sample of approximately ℓ lines of text is taken using a random start. After these are chosen, the reviewer examines each sample line and reads the entire paragraph overlapping the line. If one or more statistical statements appear on the line, each entire statistical statement is enclosed in brackets. A statistical statement may begin on a previous line and/or continue on subsequent lines. In such a case, the enclosed statement may overlap several lines in addition to the sample line. Once the statistical statements overlapping the sample line are identified all statistical statements which begin on the sample line are included in the sample of statements to be tested.

4.3 Estimation of the proportion of valid, invalid, and untestable statistical statements.

After each statistical statement has been tested and found to be either valid, invalid or untestable, inflation estimates are obtained of the total number of statistical statements in the report, and the total number of valid, invalid, and untestable statements. Finally, ratio estimates are obtained for the proportion of statistical statements which are valid, invalid, and untestable; and the estimated variances of these estimated proportions determined.

5. Comments

In the management of large government statistical systems, a considerable effort in the way of time and money goes into the quality control of the data collection, data preparation and data processing operations for the system. Effort, however, is needed also to assess the quality of the end product of the system, namely the substantive report. Since the responsibility for preparing substantive statistical reports generally lies in the hands of subject matter persons who are not professional mathematical

statisticians, we feel that a surveillance of the technical items and statistical statements in these reports is necessary to ensure a high quality product. The procedure described above is a first attempt to formulate such a systematic quality control program.

Such a program, even done ex post facto, gives us information on the types of errors which analysts are making in their inferences and on inadequacies in their presentations of technical material. This information can serve as a useful resource in our planning of intramural training programs which would have as an objective the teaching of analysts to make clear, accurate and testable statistical statements in their reports. We feel that ultimately this would result in improved statistical reports.

One of the spinoffs which we hope to obtain from this program is in the analysis of the statements found by reviewers to be untestable. While some of these are untestable because of the lack of clarity, others are untestable because no methodology exists for making a test which is appropriate for data collected from complex surveys. It is hoped that as we catalogue these statements, we can encourage research in the development of methodology for testing hypotheses in data collected from complex sample surveys.

References

[1] U.S. Department of Health, Education and Welfare; The Mission and Policies of the National Center for Health Statistics.

[2] Levy, Paul S. and Sirken, M. G.; A Manual for the Review of Statistical Reports. Unpublished document.

[3] National Center for Health Statistics; Decayed, Missing and Filled Teeth among Children, United States. Vital and Health Statistics PHS Pub. No. 1000 - Series 11 No. 106. Public Health Service, Washington. U.S. Government Printing Office, August, 1971.

[4] National Center for Health Statistics; Periodontal Disease and Oral Hygiene Among Children - United States - 1963-65. (A forthcoming NCHS report).

[5] Miller, Rubert G. (1966). Simultaneous Statistical Inference. McGraw-Hill, New York.

[6] National Center for Health Statistics; Need for Dental Care Among Adults, United States, 1960-62. Vital and Health Statistics PHS Pub. No. 1000 - Series 11 - No. 36. Public Health Service, Washington. U.S. Government Printing Office, March, 1970.

[7] Sirken, M. G. and Levy, P.S.; Multiplicity estimation of proportions based on ratios of random variables. Unpublished manuscript (1972).