STATISTICAL STANDARDS FOR ANALYZING DATA BASED ON COMPLEX SAMPLE SURVEYS

Monroe G. Sirken, B. Iris Shimizu, Dwight B. Brock, and Dwight K. French
National Center for Health Statistics

Introduction

We are preparing a manual of standardized procedures that will be used by mathematical statisticians in reviewing statistical reports that are prepared by analysts in the National Center for Health Statistics (NCHS). The rationale and procedures of the quality control program for reviewing the Center's published reports were described by Levy and Sirken [1972]. After the manual has been tested, it will be used to train the analysts. In this paper, we describe the standards and protocols for preparing the texts of statistical reports that are based on complex sample surveys and we summarize our preliminary efforts in applying these standards in the review process.

Vital and Health Statistics is one of the principal publications of the Center. The publication contains eleven series of reports, the combined series being often referred to as the "rainbow reports" since each series of reports has its own distinctively colored jackets. There are four series of reports which are based on data collected in complex sample surveys including the Health Interview Survey, the Health Examination Survey, the Institutional Population Surveys and the Hospital Discharge Survey. Some 20 to 25 substantive statistical reports are published annually in these four series.

Although the styles differ somewhat depending on the series and the analysts, basically, the format and composition of the four series of reports are quite similar. There are essentially three parts to a report. (1) Summary tables present the basic findings of the survey. (2) Appendices describe the statistical limitations of findings including the estimates of sampling error and, when available, estimates of measurement error. (3) The text, which is usually descriptive rather than analytical, summarizes and highlights the findings presented in the summary tables. Standards and protocols for each part of the report will be presented in the manual for reviewing statistical reports. We limit the scope of this paper, however, to a consideration of the statistical standards for reviewing the text which has presented, by far, the greatest difficulty and challenge.

We are well aware of the limitations of these initial efforts, and some qualifiers need to be made. In preparing these statistical standards, we limited our concern to developing statistical tests that would justify the kinds of statistical statements that are being made in the texts of statistical reports of the National Center for Health Statistics. Although we believe the standards might be applicable to descriptive reports prepared by other Federal statistical agencies, we cannot vouch for this possibility. These standards provide a basis for accepting or rejecting the statements that appear in the texts, but only from the viewpoint of sampling errors. We assume that estimates of sampling errors are available for all statistics presented in the summary tables, and we also assume that large sample normality assumptions apply. The standards do not at present consider the effects of nonsampling errors. Nor do these standards provide a means for determining whether statistical statements that are justified by the data and should have been made, in fact, are made in the texts of statistical reports.

Data based on complex sample surveys present difficult problems for analysts in the Center as well as in other Federal statistical agencies. Recently, Kruskal [1973] and Moore [1973] have commented on some of these common core problems. We are not satisfied with the tests we have developed, and hopefully better tests will evolve. In the meantime, we believe there is some virtue in establishing some standards if for no other reasons than to increase the comparability of reports prepared by different analysts and to improve communication between the analysts and the data consumers.

Statistical Standards

In reviewing the text of a report, the basic unit of analysis is the statistical statement. This is defined as a phrase, clause, sentence, or group of sentences which make inference(s) about population parameter(s) from statistics that are subject to sampling and/or measurement errors. A statistical statement may either contain an estimate of a parameter, or it may report the outcome of a test of a hypothesis.

The principal function of statistical statements is to describe and summarize the estimates that are presented in the summary tables of the report. Usually, the text will devote at least a paragraph to each summary table. Exhibit A is a model of the typical summary table. The stub of the table presents demographic variables, in this example, age and sex. The spread variable is usually a health variable, in this case hospital bedsize. Other examples of health variables include: type of acute conditions, cause of death, birthweight (of newborns), type of aid used by residents in nursing homes, etc. The cell entries in the table are estimates of morbidity rates for which the denominators are usually the sizes of the exposed-to-risk populations. Some examples of population morbidity rates are: percentages of the population with a specified health attribute, incidence and prevalence rates, etc. But the denominator is not necessarily the size of a population of persons. In the model table presented in Exhibit A, for example, the denominator of the average length of stay is the number of discharges.

EXHIBIT A. Average Length of Stay, by Sex and Age of Patient and Bed Size of Hospital: United States, 1967

Sex and Age	Hospital Bedsize					
	All Sizes	< 100 Beds	100-199 Beds	200-299 Beds	300-499	500 + Beds
Both Sexes	Average Length of Stay (in days)					
All Ages	8.4	7.5	7.6	8.7	8.9	10.3
< 15	5.5	4.1	4.8	5.2	5.9	8.4
15-44	6.2	5.1	5.5	6.4	6.7	7.7
45-64	10.1	8.5	9.2	10.2	10.7	13.1
65+	14.1	13.1	13.6	14.4	14.9	16.0
Male						
All Ages	9.0	7.5	7.8	9.3	9.7	11.4
< 15	5.5	4.1	4.8	4.9	6.0	8.2
15-44	7.3	5.4	6.1	7.9	8.3	9.8
45-64	10.2	8.2	9.0	10.3	11.3	13.1
65+	13.5	12.2	12.4	14.3	14.2	15.6
Female						
All Ages	8.1	7.5	7.5	8.2	8.3	9.5
< 15	5.5	4.2	4.7	5.5	5.8	8.5
15-44	5.8	5.0	5.3	5.9	6.1	6.9
45-64	10.1	8.8	9.3	10.1	10.2	13.0
65+	14.7	13.8	14.5	14.4	15.4	16.3

The application of statistical standards in reviewing the texts of statistical reports involves three distinct operations: (1) identifying statistical statements, (2) classifying statistical statements and (3) testing the validity of statistical statements.

The reviewer identifies statistical statements by bracketing sets of contiguous words in the text. Actually this involves two steps. First, the reviewer decides whether a set of words is a statistical statement because it makes an inference. Second, he identifies the words that begin and end every statement, such that none of the statements overlap.

After the reviewer identifies the statistical statements, he classifies them. According to our current classification scheme, which is based entirely on the Center's statistical reports, there are essentially five type of statistical statements. These types are listed and defined in Exhibit B. The illustrations of the types of statements presented in Exhibit B refer to morbidity rates that are displayed in Exhibit A.

Finally, the reviewer judges the validity of the testable statistical statements. Tests have been devised for each type of statement except for type 5. This type of statement is untestable either because it is ambiguous or it is not amenable to an existing test. Parsimony was a guiding principle in developing the typology of statements because we opted for a few general tests in preference to many specific tests. We were hopeful that type 5 statements would comprise a small portion of all statistical statements.

Simple statements do not involve testing hypotheses. Hence, the morbidity rates contained in these statements are subject only to reliability checks. We require that the coefficient of variation of the estimated rate be less than or equal to 25 percent. If fewer significant digits are given in the statistical statement than are shown in the summary table, we also require that the difference between the text figure and the estimate in the table either is less than one standard error of the estimate or less than five percent of the estimated value, whichever is smaller.

Single, multiple, and joint comparison statements make comparisons between two or more morbidity rates and hence involve tests of statistical hypotheses. The tests are made at the five percent level of significance and they are carried out as two-tailed tests except for those statements in which the analyst has specifically stated an interest in one-sided alternatives.

The test for a single comparison statement, which compares morbidity rates for a population containing two subdomains, is the usual test for significant differences between normal deviates. It is noteworthy that <u>not</u> all statements that compare morbidity rates for two subdomains are

EXHIBIT B. Definitions and Illustrations of Types of Statistical Statements

Type of Statement	Definition	Examples
1. Simple Statement	Infers the morbidity rate of a population.	1. The average length of stay for females was 8.1 days. 2. The average length of stay for persons under 15 in hospitals with 300-499 beds was about 6 days.
2. Single Comparison	Compares the morbidity rates for a population domain containing two subdomains.	1. The average length of stay for males was higher than for females. 2. Males had an average length of stay in the largest hospitals that was 1.9 days longer than the corresponding average for females.
3. Multiple Comparison	Compares morbidity rates for a population domain containing more than two subdomains.	1. The age group with longest average length of stay was the group 65 and over. 2. The average length of stay in the largest hospitals ranged from a low of 7.7 days in the 15-44 year age group to 16.0 days for those persons 65 and over. 3. Average length of stay for males increased for each successive age group. 4. Average length of stay tended to increase as hospital bedsize increased. 5. Average length of stay for males was higher in the largest hospitals than in the smallest.
4. Joint Comparison	Compares the morbidity rates between the subdomains for two or more population domains.	1. Males had a longer average length of stay than females in each of the three largest bedsize categories. 2. The average length of stay for males increased with each increase in age for all categories of hospital bedsize. 3. Average length of stay tended to increase within each age group as hospital bedsize increased.
5. Untestable Statement	An ambiguous statement or a statement for which a statistical test does not exist.	1. There exists a significant difference in the age distributions of average length of stay between males and females. 2. Comparable percentages among females are somewhat more variable across the age range than those for males.

classified as single comparison statements. Unless the analyst presents evidence in the text that he conjectured the hypothesis before viewing the data, these statements are classified as single comparison statements only if the two domains, such as male or female, comprise the entire population domain. Otherwise statements comparing the morbidity rates of two subdomains are classified as multiple comparison statements. For example, a statement comparing the average length of hospital stay for two bed size groups of hospitals is a multiple comparison statement because hospitals are classified into more than two bed size groups (see Exhibit A).

Multiple and joint comparison statements imply comparisons between more than two morbidity rates. With very few exceptions, which we will not describe here, statements of these types are tested by the Bonferroni method of multiple comparisons. This test is applied to the hypotheses implicated by the statistical statement. For instance, suppose that there are H possible pairwise comparisons between the morbidity rates comprising the population subdomains. Then the null hypothesis is that each of the H differences is equal to zero. The alternative hypothesis, however, could take a variety of forms: perhaps one, or two or even all of the H differences are nonzero. The alternate hypothesis, whichever form it takes, is determined by the statistical statement.

As indicated in Miller [1966], the Bonferroni method consists of a family of tests. Using the notation of Dayton and Schafer [1973],

let us define the probability of a nonzero family error rate by P(F); that is, P(F) is the simultaneous significance level for the defined family of tests. To avoid the additional assumption of independence of all component tests in the family, we use the Bonferroni inequality as an upper bound on the simultaneous significance level. Specifically, $P(F) \leq \sum_{i=1}^{H} \alpha_i$, where α_i is the significance level of the ith component test. This bound can be derived from Boole's inequality, a well-known result in probability theory--see, for example, Feller, [1968].

In our application of the Bonferroni technique to multiple comparisons, each component test is the usual test for significant differences between normal deviates, but with the α_i (as above) adjusted so that the simultaneous significance level is $\sum_{i=1}^{H} \alpha_i = .05$. For simplicity, this adjustment is made by setting each $\alpha_i = .05/H$, where, again, H is the total number of possible comparisons implicated by the statement.

Joint comparison statements may be viewed as combinations of single comparison statements and/or multiple comparison statements. Joint comparisons are tested by the Bonferroni method of multiple comparisons with $\alpha_i = \frac{.05}{\sum_{j=1}^{S} H_j}$, where H_j is the total number of possible comparisons among subdomains in the jth population domain, and S is the number of population domains covered by the statistical statement.

Experimental Test of Statistical Standards

It is one thing to devise a set of standards for statistical statements and another to apply these standards with a degree of reliability. Therefore, we designed and conducted an experiment in order to obtain a measure of the reliability of applying the standards. Another objective of the experiment was to estimate how often "untestable" and each of the other four types of statistical statements appear in the NCHS reports. A final goal will be to obtain preliminary estimates of the proportion of statistical statements in NCHS reports that are valid on the basis of these standards. However, the final part of the experiment has not yet been completed.

The experiment was based on eight recently published NCHS rainbow series reports. Two reports were selected from each of four series of reports that are based on data collected in complex sample surveys. A compact section containing about 50 to 70 statistical statements was randomly selected within each report, making a total of more than 400 statistical statements in the experiment. It should be noted that these eight reports were prepared by the analysts prior to the development of the standards that were being applied to them.

Six mathematical statisticians served as reviewers in the experiment. They represented a variety of statistical backgrounds ranging from a junior statistician to mid-level statisticians with Ph.D.'s in mathematical statistics. Each reviewer independently read the sample texts of the eight reports, identifying and classifying the statistical statements according to the protocols described earlier. The experiment, thus, is based on six independent reviews of the texts of eight reports. In addition, a seventh measurement was made. This seventh measurement, referred to as the "true" measurement in the following analysis, represents the majority opinion of the six reviewer observations. In those cases for which there was no majority opinion among the reviewers, the statements were referred for adjudication.

According to the "true" measurement, the text covered in the eight reports of the experiment contained 457 statistical statements. These statements are distributed by classification type in Table 1. About 10 percent of the statements are untestable, about 20 percent are simple statements and about two-thirds are comparative statements. The six reviewers identified 2684 statements, an average of 447.3 statements per reviewer. Table 1 also distributes these 2684 statements according to the types classified by the reviewers. The two distributions are in close agreement. Nevertheless, there were substantial differences among the reviewers as to the number of statements identified and as to classification of statements.

Table 1: Percent Distribution of Statement Types by "True" Classification and Reviewers' Classification of the Type of Statistical Statement

Type of Statistical Statement	"True" Classification	Reviewers' Classification
Number of Statements	457	2684
Total	100%	100%
1. Simple	22%	23%
2. Single Comparison	9%	9%
3. Multiple Comparison	31%	30%
4. Joint Comparison	26%	27%
5. Untestable	12%	11%

Let us note here that even though a reviewer may disagree with the "true" measurement on the classification of a statement, this does not necessarily imply that there would be disagreement on the validity of that statement. For example, while a reviewer may misclassify a

multiple comparative statement and call it a single comparative statement, it is possible that the difference implied by the comparison is not significant under either classification. However, since we have not yet tested the statements in the experiment, we make no further comments in this paper concerning agreements with respect to statement validity.

We turn our attention to the differences between the reviewers in identifying and classifying statements. The statements identified and classified by each reviewer were compared with the "true" measurements. The total of these comparisons, summed over all reviewers, are summarized in Table 2. From this table, we are justified in inferring differences among the reviewers, since, as we noted earlier, the "true" measurement represents the majority opinion of the six reviewers. Table 2 indicates that reviewers were subject to two kinds of errors; identification errors and classification errors.

Identification errors were committed by reviewers whenever statistical statements were entirely missed or when nonstatistical statements were erroneously identified as being statistical statements. Identification errors were also committed when reviewers merged two or more separate statistical statements into a single statement or divided a single statistical statement into two or more statements. In the former case, we counted two statistical statements as missed and one as erroneously identified, and in the latter case, we counted one statistical statement as missed and two statements as erroneously identified. If none of the reviewers had erroneously identified any statements, a total of six times the number of "true" statements, that is $6 \times 457 = 2742$ statements, would have been counted in the experiment. Henceforth, we shall refer to this number as the total actual number of statistical statements. According to Table 2, however, a total of 3206 statistical statements were counted in the experiment because the six reviewers identified only 2220 statements that were actual statistical statements, according to the "true" measurement. In addition, they erroneously identified 464 statements which were not statistical statements and failed to identify, or missed, 522 statistical statements. The difference between the missed and the erroneously identified statements represents a net identification error of -2.1 percent, or a net undercount of 2.1 percent of the total actual number of statistical statements. The sum of the missed and erroneously identified statements represents a gross identification error of about 36 percent. That is, the number of statements incorrectly identified by reviewers represents more than a third of the total actual number of statements contained in the experiment.

Classification errors were committed by reviewers when they misclassified the 2220 statements that they correctly identified. Thus, according to Table 2 the reviewers incorrectly classified about 17 percent of the 2220 statements which were correctly identified or about 12 percent of the 2742 total actual number of statistical statements in the experiment.

Next we consider the net and gross errors in identifying and classifying statements by type of statement. These errors are derived from Table 2 and they are presented in Table 3. For example, of the 612 statements that should have been identified and classified by reviewers as simple statements, 81 were missed entirely and 5 were identified but erroneously classified. On the other hand, the reviewers classified 87 nonstatistical statements as simple statements, and in addition, they misclassified 10 statistical statements as simple statements. Thus, the net identification error and the net classification error are $(87 - 81)/612 = 1.0$ percent and $(10 - 5)/612 = 0.8$ percent, respectively. Their algebraic sum, which we will refer to as the combined net error, is 1.8 percent. Similarly the gross identification and classification errors are $(81 + 87)/612 = 27.5$ percent and $(5 + 10)/612 = 2.5$ percent, respectively, and their sum, 29.9 percent, is the combined gross error. The errors presented in Table 3 for the remaining types of statements are calculated in a similar manner.

The absolute values of the combined net errors in Table 3 exceed 5 percent for two of the five statement types. The combined net error for the multiple comparative type is -5.3 percent, and for the untestable statements, it is -10.1 percent. With respect to the components of the combined net errors none of the absolute values of identification errors or classification errors is greater than 5 percent except for the net identification error for untestable statements, which is -12.6 percent.

The combined gross errors in Table 3 are large for every type of statement; they range from about 30 percent for simple statements to 100 percent for untestable statements. All of the gross identification and classification errors are about 25 to 50 percent, except for the gross classification error for simple statements, which is 2.5 percent.

For every type of statement, identification errors contributed more than classification errors to both the combined net errors and the combined gross errors. Entirely missing statistical statements and erroneously enumerating nonstatistical statements were relatively minor identification problems compared to the problem of setting the bounds for the statements. Merging two or more statistical statements into a single statement and dividing a single statistical statement into two or more statements were both rather common types of identification errors for testable statements. On the other hand, errors due to dividing statements were much more common than those due to merging statements in identifying untestable statements because the reviewers often identified each testable part of an untestable statement as a separate statistical statement.

In closing we feel obliged to note the ways in which limitations in the execution of the experiment may have contributed to the large gross errors that were detected. The manual was

in a draft form and it became necessary to make some changes in the protocols during the experimental period. Also, two of the six reviewers had virtually no experience with the standards prior to the experiment. Since we view the experiment as a pretest of the standards, we anticipate that if we were to repeat the experiment the measurement errors would be substantially smaller than those presented in this paper.

Table 2: Comparison of Reviewers' Classification with the "True" Classification of Statistical Statements.

"True" Classification of Type of Statistical Statement	Reviewers' Classification						
	Total	Statistical Statements Missed by Reviewers	Type of Statistical Statement				
			1 Simple	2 Single Comparison	3 Multiple Comparison	4 Joint Comparison	5 Untestable
Total	3206	522	623	253	807	715	286
Non-Statistical Statements Identified by Reviewers	464		87	42	129	136	70
1. Simple	612	81	526	2	1	1	1
2. Single Comparison	258	48	1	171	25	8	5
3. Multiple Comparison	852	168	5	14	583	60	22
4. Joint Comparison	702	115	2	15	58	465	47
5. Untestable	318	110	2	9	11	45	141

Table 3. Net and Gross Errors in Identifying and Classifying Statements by Type of Statistical Statement

Type of Statement	"True" Number of Statements	Percent Net Errors			Percent Gross Errors		
		Combined	Identification	Classification	Combined	Identification	Classification
1. Simple	612	1.8	1.0	0.8	29.9	27.5	2.5
2. Single Comparison	258	-1.9	-2.3	0.4	65.5	34.9	30.6
3. Multiple Comparison	852	-5.3	-4.6	-0.7	57.9	34.9	23.0
4. Joint Comparison	702	1.9	3.0	-1.1	69.4	35.8	33.6
5. Untestable	318	-10.1	-12.6	2.5	101.3	56.6	44.7

References

Miller, R. G., [1966]. Simultaneous Statistical Inference. McGraw-Hill Book Co., Inc., New York.

Feller, W., [1968]. An Introduction to Probability Theory and Its Applications, Volume I (3rd Editn.). John Wiley and Sons, Inc., New York.

Levy, P. S. and Sirken, M. G., [1972]. "Quality Control of Statistical Reports." Proc. Am. Statis. Assoc. Social Statis. Sec., 356-359.

Dayton, C. M. and Schafer, W. D., [1973]. "Extended Tables of t and Chi Square for Bonferroni Tests with Unequal Error Allocation." Journal of the American Statistical Association, 68, 78-83.

Kruskal, W., [1973]. "The Committee on National Statistics." Science, 180, 1256-1258.

Moore, G. M., [1973]. "On the 'Statistical Significance' of Changes in Employment and Unemployment." Statistical Reporter, 137-139.