A BRIEF ORIGINAL CONTRIBUTION

Quantitative Evaluation of Multiplicity in Epidemiology and Public Health Research

Kenneth J. Ottenbacher

Epidemiologic and public health researchers frequently include several dependent variables, repeated assessments, or subgroup analyses in their investigations. These factors result in multiple tests of statistical significance and may produce type 1 experimental errors. This study examined the type 1 error rate in a sample of public health and epidemiologic research. A total of 173 articles chosen at random from 1996 issues of the American Journal of Public Health and the American Journal of Epidemiology were examined to determine the incidence of type 1 errors. Three different methods of computing type 1 error rates were used: experiment-wise error rate, error rate per experiment, and percent error rate. The results indicate a type 1 error rate substantially higher than the traditionally assumed level of 5% (p < 0.05). No practical or statistically significant difference was found between type 1 error rates across the two journals. Methods to determine and correct type 1 errors should be reported in epidemiologic and public health research investigations that include multiple statistical tests. Am J Epidemiol 1998;147:615-19.

Problems involving multiple statistical testing of hypotheses in health care and medical research arise for the following reasons: 1) the repeated analysis of accumulating data; 2) the use of multiple dependent measures; and 3) the analysis of data from subgroups (4). All three of these practices are common in public health and epidemiologic research. For example, Godfrey (5) demonstrated that researchers frequently present and analyze means from several groups within the same study. She found that the most common method of statistically comparing several means involved the use of multiple t tests. Godfrey correctly argued that the use of univariate statistical procedures to analyze the results of studies containing multiple contrasts was inappropriate. Her analysis revealed that of 50 articles examined from the New England Journal of Medicine, a majority (54 percent) used improper univariate statistical procedures to analyze differences between subgroup means.
The use of several dependent variables in the analysis of data from a single sample also results in multiple statistical tests being reported. The complex nature of epidemiologic and public health research has led investigators to routinely include multiple dependent variables in their investigations (6). An epidemiologic researcher may be interested in the effect of a particular intervention on dependent variables such as weight, blood pressure, hematocrit, and serum cholesterol values in a sample of patients. As the number of dependent variables increases, so does the number of statistical tests. When this occurs, the researcher may obtain positive results on the basis of sampling error (7).

Numerous clinical researchers have suggested that multiple hypothesis testing without adjusting for inflated type 1 error rates is a common problem in medical and public health research (8–10). The purposes of this investigation were: 1) to examine the extent of the multiple testing in epidemiologic and public health research, and 2) to determine the prevalence of type 1 errors in a sample of published research.

METHODS

Five issues of both the American Journal of Public Health and the American Journal of Epidemiology were randomly selected from the journal issues published in 1996. Each individual article was examined to determine the experiment-wise error rate, the error rate per experiment, and the percent error rate (see descriptions of error rates below). All articles that reported tests of statistical significance were included in the investigation. Articles that summarized the results of previously published research and articles that did not report statistical significance tests were not included in the analysis.

Experiment-wise error

The overall experiment-wise error rate (EW) is the probability of making at least one type 1 error for the collection of tests performed in the investigation. The experiment-wise error rate can never be smaller than the error rate per comparison. The relation of per-comparison and experiment-wise error rates depends on the degree of statistical dependence of the tests. For totally independent tests, the experiment-wise error rate is equal to \(1 - (1 - \alpha)^c\), where \(c\) is the number of independent tests and \(\alpha\) is the error rate per test (traditionally 0.05 or 0.01). From this equation, it is apparent that experiment-wise error rate increases rapidly with the number of hypotheses statistically examined. For example, in a study for which five statistical tests are conducted at the 0.05 level of significance, the EW is \(1 - (1 - 0.05)^5\) or 0.23.

Error rate per experiment

The error rate per experiment (EP) is the expected number of type 1 errors in a particular group of statistical significance tests and is computed using the formula \(EP = c(\alpha)\), where \(c\) represents the number of comparisons, and \(\alpha\) is the significance level and remains constant across all tests. For example, given 20 independent statistical comparisons at the \(p = 0.05\) confidence level, \(EP = 20(0.05) = 1\). This means that at the 0.05 level we would expect one type 1 error in 20 tests of statistical significance. It is important to note that the error rate per experiment (EP) is an expected value, while the experiment-wise error rate (EW), as defined above, is a probability. The experiment-wise error rate for 20 comparisons at the 0.05 significance level is \(1 - (1 - 0.05)^{20}\) or 0.64, indicating that the probability of at least one type 1 error occurring among these tests reported as significant at the 0.05 level is 0.64.

Percent error rate

The formula for computing the percent error rate (PE) is \(PE = \frac{100c\alpha}{M}\), where \(c\) is the total number of comparisons, \(\alpha\) is the alpha level for a set of comparisons, and \(M\) is the number of statistical tests less than the designated alpha level. The percent error rate reflects the proportion of results labeled as statistically significant that are likely to be chance results. As the ratio approaches 1.00 (100 percent), it indicates that the number of tests found to be statistically significant approximates the number of tests one would expect to find to be significant purely by chance. As the ratio decreases and approaches the individual alpha level for a set of comparisons, it reflects the percent of results that are attributable to chance. The percent of results likely to be caused by non-chance factors is equal to \(100 - PE\). For example, if 1 out of 20 comparisons evaluated at the 0.05 level is statistically significant, the PE = 100(20)(0.05)/1 = 100 percent, suggesting that the number of tests found to be significant, that is 1, is the number expected by chance. On the other hand, if 4 out of 20 comparisons conducted at the 0.05 significance level are found to be statistically significant, then PE = 100(20)(0.05)/4 = 25 percent, indicating that about 25 percent of the results are expected as the result of chance, while the remaining 75 percent (three tests) are likely to be due to non-chance factors.
Rating process

The reporting style in some of the articles made the determination of the exact number of statistical tests conducted and the number found statistically significant a difficult task. Two independent raters with research degrees (PhDs) reviewed all articles and identified both the total number of tests conducted and the number reported as statistically significant. When the two raters did not agree, a third rater reviewed the article in question and the value agreed upon by at least two raters was used in the analysis. In spite of the high agreement between the raters (see below), the results reported in this investigation should be viewed as approximations of the various error rates rather than as exact values. A post hoc analysis is necessarily somewhat arbitrary in determining the number of tests conducted because the actual number cannot be precisely determined without direct access to the original data.

The relation between error rates per comparison and error rates per experiment is complex with dependent tests, a condition which may be assumed to always hold to some degree when multiple statistical tests are conducted using subjects from the same sample. Strahan (11) has argued that, although it may be difficult to estimate the exact experiment-wise error rate due to correlation among the variables, it should be clear that it is greater than 5 percent. When discussing the impact of non-independence on error rates, it is important to distinguish types of non-independence that may exist. Ryan (12) originally identified the following four instances where non-independence may occur. The first includes all those situations where several groups or subgroups are statistically compared within the context of one study. The second case is referred to as "multiple tests with intercorrelated variables." This most commonly occurs when researchers compute multiple correlation coefficients for a single sample. The third instance of multiple testing is the use of multiple factors in the analysis of variance. The $F$ ratios obtained from a factorial analysis of variance may not be independent if a common error estimate is used across the tests. Similar problems arise if other statistical procedures such as multiple $t$ tests are used to analyze data in what is essentially a factorial design. The final type of multiple testing situation is what Ryan (12) referred to as "replicated tests of a single hypothesis." This classification includes studies for which several different methods of assessing the same dependent variable are employed.

The situations described by Ryan (12) are not mutually exclusive. They do serve, however, to make it clear that interdependence between multiple statistical tests is complex and produced by numerous factors. Although the lack of independence may influence error rates, Ryan argues that it is not the main problem in interpreting error rates. He states that "The error rate per comparison and per experiment are completely unaffected by independence or lack of it. The only important factor in these rates is the number of comparisons to be made. Only the experiment-wise error rate is affected by lack of independence" (12, p. 34). In the case of the experiment-wise error rate, the more highly related the tests, the closer the experiment-wise error rate is to the error rate specified for an individual comparison.

In this examination, multivariate statistical tests that included procedures to control for type 1 error rates were considered as a single statistical test. This included analysis of variance (ANOVA) involving tests of interaction and accompanying post-hoc procedures using Scheffe, Tukey, Duncan, Newman-Keuls, or other appropriate methods of post-hoc analysis. Each ANOVA, including the post hoc analysis, was counted as one statistical procedure.

RESULTS

The 71 articles in five issues of volume 86 of the American Journal of Public Health and the 102 articles in five issues of volume 141 of the American Journal of Epidemiology contained sufficient statistical information to be included in the analysis. The intrarater agreement for all information coded from each of the articles was examined using the intraclass correlation coefficient (ICC) (13). The ICC values for all recorded information ranged from 0.91 to 1.00. Descriptive information for the experiment-wise error rate, the error rate per experiment, and the percent error rate for the articles published in the two journals appear in table 1.

A comparison of the values for different error rates illustrates that experiment-wise error rate (EW) and the percent error rate (PE) have an easier interpretation than the error rate per experiment (EP), since EW and PE are essentially bounded while EP has no upper limit. The tabled values indicate that the EW in many articles is high, revealing a likelihood of type 1 errors in the reports. This is not surprising given the stochastic nature of the quantitative analysis of public health research. The prospect that many articles which report large numbers of statistical significance tests also report occasional type 1 errors does not seem alarming. What is of more concern is the percent error rate. The average individual alpha level used in a given study provides a lower bound for the percent error rate. Thus, for most of the investigations included in the analysis, 5 percent is the lowest value PE can achieve given the 0.05 significance level. Yet, in many of the
studies, the PE indicated that approximately 20 percent or more of the findings may be erroneous. The average PE for the studies in the American Journal of Public Health was 19.16 percent, while the average mean PE for articles in the American Journal of Epidemiology was 18.73 percent (table 1).

In a majority of the 173 articles (n = 156), the error rate per experiment (EP) was greater than 5 percent. The analysis also suggests that the percent error rate provides information not specifically contained in the EW and EP. The correlation between EW and EP rates for the articles included in table 1 was $r = 0.47$. The correlation of PE with EP was $r = 0.41$ and the correlation of PE with EW was $r = 0.32$.

DISCUSSION AND CONCLUSIONS

The problem of multiple hypothesis testing has implications regarding the interpretation and implementation of epidemiologic research. For example, more than a decade ago the Food and Drug Administration refused to approve sulfinpyrazone (Anturane®, CIBA, Summit, New Jersey) as a medication to reduce mortality in the first 6 months following myocardial infarction (14). The refusal was based in part on the results of a clinical trial that included the repeated analysis of accumulated data. No procedure was used to control for the effect of multiplicity and the validity of the results was open to question.

The probability of obtaining statistically significant results from two independent tests that address the same research question can be obtained by multiplying the individual probabilities that each test will produce a significant result. For $p = 0.05$, the probability that both tests will be statistically significant is $0.05 \times 0.05 = 0.0025$. The probability that neither result will be significant is $0.95 \times 0.95 = 0.9025$. The probability that at least one of the two test results will be statistically significant is $1 - 0.9025$, or 0.0975. Thus, the probability of incorrectly deciding that the members of either one or both pairs of means are unequal using just two tests is nearly twice the probability of making the same error for a single test (0.0975 vs. 0.05). If we add a third comparison, the probability that none of the three tests will be significant is $0.95 \times 0.95 \times 0.95 = 0.8574$, so the probability that at least one test will be significant is about 14 percent or nearly three times the 0.05 level. As the number of independent statistical tests increases, the probability becomes much larger than 0.05, the original alpha (see table 1).

In trials where multiple dependent variables are used, the obvious solution to control or reduce experiment-wise error is to use some form of multivariate analysis. Multivariate procedures such as Hotelling’s $T^2$, discriminant function analysis, and logistic regression offer viable alternatives to traditional univariate approaches when multiple dependent variables are present. These procedures have been described by public health and epidemiologic researchers and are beyond the scope of this paper (15, 16).

In some instances, the best solution may be to reduce the per comparison significance level to a more stringent criterion. The Bonferroni adjustment provides a widely advocated procedure to achieve this goal. The Bonferroni inequality involves dividing the alpha level desired for the overall family of statistical tests (usually 0.05) by the number of statistical comparisons to be conducted. If two groups are compared on five separate dependent measures, each statistical comparison would be evaluated at 0.05/5 = 0.01. The Bonferroni method controls the type 1 error rate for each decision and maintains the selected alpha level (e.g., 0.05) for all the tests conducted in the investigation. The limitation of the Bonferroni method is that as the probability of making a type 1 error is decreased, the chance of committing a type 2 error is increased. Silverstein (17) demonstrated that when more than a small number of comparisons (say, five to eight) are included in a study, the Bonferroni procedure results in a dramatic loss in statistical power. Benjamini and Hochberg (18) have recently described alternatives to the Bonferroni adjustment that do not result in substantial reduction in statistical sensitivity. The Bonferroni and other p value adjustment methods, however, are viewed as too conservative by some

<table>
<thead>
<tr>
<th>Journal</th>
<th>No. of articles</th>
<th>Experiment-wise error rate</th>
<th>Error rate per experiment</th>
<th>Percent error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Am J Public Health, Vol. 86</td>
<td>71</td>
<td>0.68 0.24</td>
<td>0.90 0.57</td>
<td>19.16 9.01</td>
</tr>
<tr>
<td>(nos. 3, 4, 7, 9, 12)</td>
<td></td>
<td>0.70 0.29</td>
<td>0.87 0.51</td>
<td>18.73 9.32</td>
</tr>
<tr>
<td>Am J Epidemiol, Vol. 141</td>
<td>102</td>
<td>0.87 0.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(nos. 2, 5, 6, 9, 10)</td>
<td></td>
<td>0.90 0.57</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*SD, standard deviation.

Vol. 147, No. 7, 1998

Am J Epidemiol
Public health and epidemiologic researchers must appropriately define research questions and hypotheses as succinctly as possible and interpret the results using an alpha level appropriate to the extent of multiple testing. Procedures such as the percent error rate do not directly control type 1 error, but they do provide the investigator (and reader) with valuable information concerning the possible presence of a type 1 error in a family of statistical tests.

Determining the experiment-wise error rate for a "family" of statistical procedures can be a complex task. In this study, the statistical test was the unit of analysis and no distinction was made among statistical procedures within a study versus those between studies. Statistical tests conducted within a study generally use data from the same sample and are, therefore, assumed to be more related than statistical tests from different investigations (or samples). It is possible, however, that two different samples may be included in one research report, or that a single research article might include the results of more than one investigation. An argument could be made that the family-wise error rate should be determined based on statistical tests that address the same research question across multiple investigations, or even across the lifetime of an investigator working in a particular area (12). The individual statistical test was the unit for determining the experiment-wise error rate in this study. Other units are possible, for example, the study sample, the research report, the research question, or even the investigator. How the different units of analysis affect the experiment-wise error rate for a "family" of statistical tests is a question that can only be answered by additional research.

Technical or statistical solutions to the problem of multiplicity in epidemiologic research should not obscure a more fundamental scientific principle. There is a continuing need in health-related research to formulate concise research questions and hypotheses before the collection and analysis of data. Statistical hypothesis testing is necessarily an empirical compromise between claiming too much and suggesting too little. Public health and epidemiologic researchers must prospectively define research questions and hypotheses as succinctly as possible and interpret the results using an alpha level appropriate to the extent of multiple testing. Knowledge of experiment-wise error procedures can help achieve this goal.

ACKNOWLEDGMENTS

This research was partially supported by grant no. MCH-360646-010 from the US Department of Health and Human Services, Bureau of Maternal and Child Health.

REFERENCES