No Adjustments Are Needed for Multiple Comparisons

Kenneth J. Rothman

Adjustments for making multiple comparisons in large bodies of data are recommended to avoid rejecting the null hypothesis too readily. Unfortunately, reducing the type I error for null associations increases the type II error for those associations that are not null. The theoretical basis for advocating a routine adjustment for multiple comparisons is the "universal null hypothesis" that "chance" serves as the first-order explanation for observed phenomena. This hypothesis undermines the basic premises of empirical research, which holds that nature follows regular laws that may be studied through observations. A policy of not making adjustments for multiple comparisons is preferable because it will lead to fewer errors of interpretation when the data under evaluation are not random numbers but actual observations on nature. Furthermore, scientists should not be so reluctant to explore leads that may turn out to be wrong that they penalize themselves by missing possibly important findings. (Epidemiology 1990;1:43–46)

Keywords: multiple comparisons, null hypothesis, significance testing, statistics.

Scientists have always had a special problem in interpreting the odd or unanticipated finding. The problem is exacerbated when the unusual result does not pertain to the central focus of study, but is either incidental to the main focus or is one of many relations that a study examines. In many instances an unexpected result can be ascribed to measurement error. In other situations an odd finding may be judged to be real but inexplicable, becoming a problem that might eventually lead to revolutionary developments in understanding. But how is a researcher to know whether to ignore an unanticipated result or to conjure up an entirely new line of thinking because of it? A common practice in the biomedical and social sciences has been the half-hearted adoption of a statistical principle to cope with this problem of interpretation. This statistical principle is the procedure of adjustment for multiple comparisons. Unfortunately, this principle mechanizes and thereby trivializes the interpretive problem, and it negates the value of much of the information in large bodies of data.

In its most common guise, the multiple-comparison problem is closely linked with statistical significance testing. Under a hypothesis that two factors are unrelated and that any apparent relation in the data is attributable to chance (the *null* hypothesis), a significance test will indicate a "statistically significant" association between the factors with a probability of α, where α is the arbitrary cutoff value for significance. If n independent associations are examined for statistical significance, the probability that at least one of them will be found statistically significant is $1 - (1 - \alpha)^n$, if all n of the individual null hypotheses are true. If n is large, the probability of some statistically significant findings is great even when all null hypotheses are true and therefore any significant departure from them is the result of chance. For example, with $\alpha = .05$ and $n = 20$, the probability of at least one statistically significant finding is 0.64, assuming that all 20 of the null hypotheses are true. In practice, n is often much larger: for example, Gardner (1) examined 5000 separate associations relating sociodemographic, environmental, and mortality characteristics of English towns; even if all the null hypotheses were true, about 250 of these associations would be statistically significant at the 0.05 level for α, and the probability of at least one statistically significant finding is near 100%.

The purported problem with all these "significant" *P*-values is that many null hypotheses will be rejected even if they are correct. Of course, there is nothing peculiar about conducting a multitude of comparisons, as opposed to a single comparison, that increases the probability of rejecting a specific null hypothesis when it is correct. If $\alpha = 0.05$, there is a five percent probability of rejecting a correct null hypothesis, whether one or one billion are examined. The core of the supposed problem is that with many comparisons the *number* of potentially incorrect statements regarding null hypotheses will be large, simply because of the large number of comparisons.

Adjustment for multiple comparison is an insurance policy against mistakenly rejecting a null hypothesis re-

Editor, Epidemiology

garding any given pair of variables if in reality that null hypothesis is correct. These adjustments typically involve increasing the *P*-value, with a consequently smaller probability that the *P*-value will be less than α and thus statistically significant. Unfortunately, the cost of the insurance policy is to increase the frequency of incorrect statements that assert no relation between two factors, an error that can occur when an association in the data is not the result of chance.

The issues are analogous to those in setting a cutoff for a screening or diagnostic test. In screening, the predictive value of a test is known to be dependent on the prevalence of the disease condition in the population being tested. Thus, the relative frequency of false-positive and false-negative statements depends on how many of the individual null hypotheses are actually true, that is, on the prevalence of true null hypotheses among those relations being examined. If random numbers are analyzed, all the null hypotheses are true and there can only be false-positive statements; in this case the adjustment may be a good idea. On the other hand, if such an adjustment is made when at least some of the relations studied are not null, the net result is to weaken the information in the data on those associations. An association that would have been interesting to explore if examined alone can thus be converted to one that is worth much less attention if judged by the criteria based on adjustments. Since other associations in the set of comparisons may have no bearing on the one in question, the upshot is that irrelevant information from the data can diminish the informativeness of an association of possible interest.

The motivating concern with multiple comparisons boils down to this: chance alone can cause the unusual finding. This statement does not carry any obvious statistical implications, but it does have philosophic implications about the definition and importance of the concept *chance*. The conventional statistical doctrine that is designed to "correct" the "problem" of multiple comparisons is built on two presumptions:

1. Chance not only can cause the unusual finding in principle, but it does cause many or most such findings.
2. No one would want to earmark for further investigation something caused by chance.

Without presumption 1, as already demonstrated, there would be no need for corrective statistical action, and therefore this presumption is fundamental to the theory of adjustments for multiple comparisons. Presumption 2 is inherent in the understanding of chance that underlies much of statistical theory. The statistical solutions to the multiple-comparison problem follow from these presumptions. If either one of the presumptions is wrong, statistical adjustments for multiple comparisons cannot easily be defended. I contend that both are wrong, and with the exception of contrived settings no adjustment for multiple comparisons is appropriate.

Presumption 1: Chance Not Only Can Cause the Unusual Finding in Principle, but It Does Cause Many or Most Such Findings

A *P*-value is sometimes misinterpreted as the probability that the null hypothesis is true, that is, the probability that chance alone accounts for the degree of association observed between two variables. Because the *P*-value is in fact calculated assuming the truth of the null hypothesis, it only indirectly reflects on the validity of the assumption. Whether the null hypothesis is correct cannot be calculated as an objective probability. The tenability of the null hyothesis needs to be viewed with respect to both the evidence in the data and the tenability of other explanations. Even if the *P*-value is low, the null hypothesis may be the most reasonable explanation, in the absence of other explanations. If the *P*-value is high, it is widely appreciated that the null hypothesis may nevertheless be wrong. The *P*-value is an indicator of the relative compatibility between the data and the null hypothesis, but it does not indicate whether the null hypothesis is a correct explanation for the data.

The isolated null hypothesis between two variables serves as a useful statistical contrivance for postulating probability models. It is possible, of course, to imagine many scientific situations in which two variables, say gum chewing and the occurrence of brain cancer, would plausibly be unassociated—that is, one can imagine that many individual null hypotheses are correct. Any argument in favor of adjustments for multiple comparisons, however, requires an extension of the concept of the isolated null hypothesis. The formal premise for such adjustments is the much broader hypothesis that there is no association between any pair of variables under observation, and that only purely random processes govern the variability of all the observations in hand. Stated simply, this "universal" null hypothesis presumes that all associations that we observe (in a given body of data) reflect only random variation.

This extension of the ordinary null hypothesis is not necessary for any statistical analysis, since it is always possible to rely on a separate null hypothesis for each pair of variables. Yet, the generalization to a universal null hypothesis has profound implications for empirical science. Whereas we can imagine individual pairs of

variables that may not be related to one another, no empiricist could comfortably presume that randomness underlies the variability of all observations. Scientists presume instead that the universe is governed by natural laws, and that underlying the variability that we observe is a network of factors related to one another through causal connections. To entertain the universal null hypothesis is, in effect, to suspend belief in the real world and thereby to question the premises of empiricism.

For the large bodies of data for which adjustments for multiple comparisons are most enthusiastically recommended, the tenability of a universal null hypothesis is most farfetched. In a body of data replete with associations, it may be that some are explained by what we call "chance," but there is no empirical justification for a hypothesis that all the associations are unpredictable manifestations of random processes. The null hypothesis relating a specific pair of variables may be only a statistical contrivance, but at least it can have a scientific counterpart that might be true. A universal null hypothesis implies not only that variable number six is unrelated to variable number 13 for the data in hand, but also that observed phenomena exhibit a general disconnectivity that contradicts everything we know.

The untenability of this universal null hypothesis is nearly always skipped over in the presentation of procedures to deal with multiple comparisons. Teachers of statistics sometimes even lapse into a tacit acceptance of this hypothesis. Consider, for example, this incorrect statement from an article on adjustments for multiple comparisons in the recently published book, *Medical Uses of Statistics* (2):

> In general, if we make n tests, the probability of finding at least one spuriously significant result can be calculated as follows: Prob(at least one spurious test result) $= 1 - (1 - \alpha)^n$.

This statement is false because the "significant" results are spurious only if the universal null hypothesis is indeed correct, an essential qualification that the author omitted.

Are there any settings for which a universal null hypothesis might be applicable? The burden of answering this question should be put to those who advocate that multiple comparisons constitute a problem in need of correction. (One might pose a type of universal null hypothesis to evaluate the results of studies on extrasensory perception (3), but even for this area of study it is easy to theorize how nonrandom associations might arise from biases even if extrasensory perception does not exist.) Without a firm basis for posing a universal null hypothesis, the adjustments based on it are counterproductive. Instead, it is always reasonable to consider each association on its own for the information it conveys. This is not to say that the setting in which the observations are made should be ignored, but only to emphasize that there is no formula that can substitute for critical evaluation of each association or observation that comes to attention.

Presumption 2: No One Would Want to Earmark for Further Investigation Something Caused by Chance

Chance is a term often used as if its meaning were well understood. Commonly it is taken to denote a mysterious force that introduces random variation into observable phenomena, and, indeed, I have used the term in this conventional sense up to this point. Nevertheless, it is important to scrutinize the concept. The Oxford English Dictionary gives 13 definitions for the noun *chance* (4). The first is "the falling out or happening of events." The sixth definition comes closest to the statistical and scientific usage: "absence of design or assignable cause; often itself spoken of as the cause or determiner of events, which appear to happen without the intervention of law, ordinary causation or providence." Despite common reference in statistics and science to "chance" as an "explanation" for observed associations, the term *chance* explains nothing. The randomness usually associated with chance is a mathematical assumption that is typically not logically related to an "absence of design" and does not enhance the vacuous explanation that "chance" provides in understanding how the observations occurred. The most one might say for the explanatory value of the term is that it implies that other explanations are obscure. Nevertheless, these other explanations may be discoverable and meaningful, and should not necessarily be ignored.

The inherent unpredictability of chance phenomena would seem to preclude meaningful research on such phenomena. Randomness, however, is only a theoretical idealization. Most, perhaps all, of the events that routinely are classified as "chance" occurrences have causal explanations (5). What we refer to as a "chance" encounter may be unexpected or unusual, but it is caused and usually could have been prevented. Dice rolls, coin tosses, and random-number generators behave according to known physical laws that account for the outcome. We describe the outcome as a chance result because the causal explanations are too intricate, the outcome is too complicated a function of the initial conditions, or the initial conditions are not known sufficiently well. As Poincaré explained (6),

> . . . it may happen that slight differences in the initial con-

ditions produce very great differences in the final phenomena; a slight error in the former would make an enormous error in the latter. Prediction becomes impossible and we have the fortuitous phenomenon.

We tend to ascribe to chance the variability in observations that we cannot predict. In doing so, we use the term *chance* to connote variability that might be accounted for with greater knowledge. Whereas the occurrence of lung cancer may once have been viewed entirely as a chance phenomenon, we can now explain a great deal of the variability in its occurrence. What variability we still cannot explain we consider to be due to chance, but this degree of ignorance need not be taken to be a permanent state, since advancing knowledge can reduce the unpredictable variability further. Thus, much of what is now viewed as chance will, upon further research, be explainable and no longer be considered chance. To the extent that adjustment for multiple comparisons shields some observed associations from more intensive scrutiny by labeling them as chance findings, it defeats the purpose of scientists.

In a recent paper about multiple comparisons, the author stated (7):

Thus, unless account is taken of multiplicity, the investigator may be mistakenly impressed by the seemingly extreme (and thus seemingly rare) result.

By claiming that to be impressed by the extreme result is a mistake, the writer accepts the universal null hypothesis as true, at least as a starting point. A scientist, however, should be posing theories to explain natural phenomena (8). Since an empirical scientist presumes that nature follows regular laws, the scientist confronted with an extreme observation or association should grasp at every opportunity to understand it rather than to ignore it. Being impressed by an extreme result should not be considered a mistake in a universe brimming with interrelated phenomena. The possibility that we may be misled is inherent to the trial-and-error process of science; we might avoid all such errors by eschewing science completely, but then we learn nothing.

Those who subscribe to adjustments for multiple comparisons face what has been whimsically described as a "penalty for peeking" (9). The more that one observes, the stiffer the penalty exacted for the privilege of observing. If this premise is allowed, many logical inconsistencies arise. Imagine an investigator who studies the contrast between drugs A and B. Assume that drug C is studied also, and because of a multiple-comparison adjustment the contrast between A and B is considered not worth pursuing. But perhaps data on C were late in coming; if A and B had been compared more hastily, before the data on C arrived, the contrast between A and B would have seemed more important. The "penalty for peeking" at the information on drug C reduces the apparent importance of the contrast between drug A and B. Suppose that drug C differs considerably in its effect from drug B. Will this difference be less worthy of attention when, sometime in the future, information on drug D comes along as part of the same research program? Should an investigator estimate on the first day of data analysis how many contrasts ultimately will come along before making adjustments for multiple comparisons? Where do the boundaries of a specific study lie, or a specific investigator's frame of reference?

The paradox of paying a penalty for having more information is a concept that is commonly accepted. The paradox arises only if we are willing to assume the truth of the universal null hypothesis; however, the premise of a universal null hypothesis is one that empirical science constantly refutes. It lacks any apparent heuristic value. Therefore the "penalty for peeking" at the data should be unacceptable to any empiricist. Science comprises a multitude of comparisons, and this simple fact in itself is no cause for alarm.

References

1. Gardner MJ. Using the environment to explain and predict mortality. J R Stat Soc A 1973;136:421–40.
2. Godfrey K. Comparing the means of several groups. In: Bailar JC, Mosteller F, eds. Medical uses of statistics. Waltham, Massachusetts: New England Journal of Medicine Books, 1986.
3. Diaconis P. Theories of data analysis: from magical thinking through classical statistics. In: Hoaglin DC, Mosteller F, Tukey JW, eds. Exploring data tables, trends, and shapes. New York: John Wiley and Sons, 1985.
4. The Oxford English dictionary. Oxford: Oxford University Press, 1971.
5. Kolata G. What does it mean to be random? Science 1986; 231;1068–70.
6. Poincaré JH: Chance. In: Newman JR, ed. The world of mathematics. Vol 2. New York: Simon and Schuster, 1956:1380–94.
7. Godfrey K. Comparing the means of several groups. N Engl J Med 1985;313:1450–6.
8. Burke M. The scientific method. Science 1986;231:659.
9. Light RJ, Pillemer DB. Summing up. The science of reviewing research. Cambridge: Harvard University Press, 1984.