Multiplicity and statistical significance

Gilbert Berdine MD

The *New England Journal of Medicine* (NEJM) has recently announced new guidelines for statistical reporting of significant findings.[1] "Some Journal readers may have noticed more parsimonious reporting of P values in our research articles over the past year." *The New England Journal of Medicine* is concerned that P values may be misused in some situations to misrepresent Type 1 statistical errors as statistically significant outcomes. In particular, *the New England Journal of Medicine* is concerned about issues of multiplicity. The purpose of this article is to discuss the issue of multiplicity in terms understandable by researchers who are not statisticians.

Consider a measurement which is normally distributed about a mean value of μ with a standard deviation of σ (Figure 1). For normally distributed measurements, the mean, median, and mode averages are equal. The standard deviation represents an average difference between values for each element of the sample and the sample mean. A small standard deviation means that individual element values are tightly grouped about the mean value for the sample and a large standard deviation means that individual element values are loosely grouped about the mean value for the sample.

The standard deviation is explicitly related to confidence limits or P values. From Figure 1, one can see that 95% of elements in the sample have values within 2 standard deviations from the sample mean. The 95% confidence limit for this sample is approximately the mean value plus or minus 2 standard deviations. P values are used to determine the likelihood that another sample whose mean value differs from the mean value of the control sample is statistically

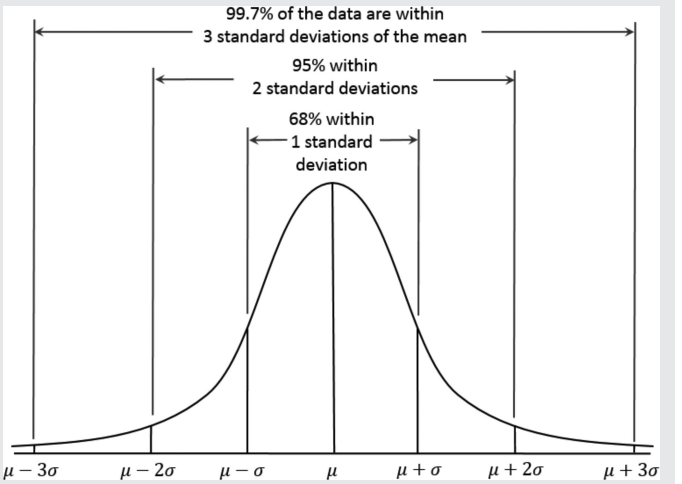

Figure 1. Normal Distribution.[2]

different from the control sample. The larger the difference between the means of the test sample and the control sample, the more likely the two samples are statistically different. The standard deviations of the two samples become the scale for measuring the differences between group means. A P value of 0.05 means that there is a 5% chance that the difference between two means is attributable to random factors or luck rather than the hypothesis under question. The null hypothesis is that the two samples are statistically the same. A P value less than 0.05 is arbitrarily used to discriminate significant differences and non-significant differences. However, one should not view a P value of 0.049 as fundamentally different from a P value of 0.051; both results should be considered to have a Type 1 error rate of about 5%. Type 1 error is when we believe that two samples are from different sources when they are, in fact, both from the same source and any observed differences are attributable to luck rather than meaningful effect.

Multiple measurements or multiple comparisons increase the likelihood that an individual value will be outside a 95% confidence limit or that the P value

Corresponding author: **Gilbert Berdine**
Contact Information: Gilbert.Berdine@ttuhsc.edu
DOI: 10.12746/swrccc.v8i34.681

for a single comparison will be less than 0.05. This problem is called multiplicity. Multiplicity can be simply illustrated by considering a chemistry panel of 20 different tests. Each test has a 95% confidence limit meaning that 95% of healthy individuals will have test values within the confidence limit. The probability that a normal healthy individual will have N tests all within a confidence limit of α is given by α^N. For a panel of 20 tests and a confidence limit of 95%, the probability of all 20 tests being within the confidence limit or normal range is only 35%. More tests mean a lower probability. One can design studies with a large number of independent measurements and, with a large enough number of variables, expect to find so-called statistically significant results even with pure noise.

Multiplicity issues can arise when multiple measurements of the same variable are made longitudinally over time. Consider a Kaplan-Meier analysis of survival following a procedure. If one analyzes the differences at each time increment separately, one can expect to see individual differences reach a P value less than 0.05 due to the accrual of Type 1 error.

New England Journal of Medicine guidelines provide some suggestions for dealing with multiplicity.[3] One can avoid multiplicity by having a single primary outcome. Results for secondary outcomes can be used to suggest further studies but not be the basis for treatment recommendations. Study design can include multiple measures if the statistical threshold for significance is adjusted properly. The Food and Drug Administration (FDA) has published guidelines for controlling the Type 1 error rate with multiple primary endpoints.[4]

Data mining or data dredging is inappropriate. These practices are what made these new guidelines necessary. Data mining has two basic forms. The first form takes large amounts of data, performs many tests of statistical significance and reports the positive results. The second form of data mining is to try different statistical methods on a set of data until one gets the desired result. Both forms have led the NEJM and FDA to require the primary endpoints and the statistical methods used to test the primary endpoints be specified prior to conduct of the clinical trial. This is straightforward for the FDA since studies must be filed before they are conducted, but not so straightforward for academic journals since there is no requirement of public filing of study plans prior to conducting a trial.

Control of Type 1 error with multiple primary endpoints can be single step or multi-step. Single step methods are easier to perform, and credibility is unquestioned when the single step method is satisfied. The downside of single step methods includes loss of statistical power; a study using single step correction will generally require more subjects to succeed. The Bonferroni Correction is probably the most well-known single step method. The Type 1 error is equally divided among all the endpoints being tested. One divides the desired P value threshold (say 0.05) by the number of measurements.[4] For example, a study comparing conservative oxygen supplement with liberal oxygen supplement with primary endpoints of number of days on ventilator, number of hospital days, survival at 30 days, survival at 90 days, and survival at 1 year would use a P value threshold of 0.01 for each of the 5 primary endpoints.

It is possible to divide the total Type 1 error unequally among the multiple primary endpoints. This is done by assigning different weights or percentages of the total Type 1 error. The weights must add up to 1, the weights must be specified before the study is performed, and there can be no modifications to the weights once the study has begun.

The Holm procedure is a multi-step control of Type 1 error that starts with the most significant endpoint and steps down in order of most significant to the least significant endpoint.[4] The P value threshold for the most significant result is the same as for the unweighted Bonferroni Correction. For our above example of a total Type 1 error of 0.05 and 5 endpoints, this P value would be 0.01. The next most significant endpoint is tested against a P value divided by the number of endpoints remaining. For our above example, this would be 0.05/4 or 0.0125. This process continues until the least significant endpoint is tested against the total Type 1 error. For our example above, the least significant endpoint would be tested against a P value of 0.05. The Holm procedure terminates whenever a result is not significant. No further

tests are performed, and the remaining endpoints are considered to be not statistically significant.

The Hochberg procedure is like an inverse Holm procedure. The Hochberg procedure is a multi-step control that starts with the least significant endpoint and continues in order of least significant endpoint to most significant endpoint. The least significant endpoint is tested against a P value equal to the Bonferroni Correction. For our above example, the least significant endpoint would be tested against a P value of 0.01. If the test fails, the next least significant endpoint is tested against the total Type 1 error divided by the number of remaining endpoints. For the 2nd test, this would be 0.05/4 = 0.0125. Whereas testing continued for the Holm procedure until a test fails, testing continues for the Hochberg procedure until a test succeeds. All subsequent endpoints are deemed significant. For the case where the Hochberg procedure succeeds on the first test, it becomes identical to the Bonferroni Correction.

Needless to say, the FDA and NEJM require the method to be specified before the study is performed. One cannot perform the study, try the Bonferroni Correction, then try the Hochberg procedure, and then try the Holm procedure and use whatever generates the most significant results.

P values are tools helpful to interpreting results. P values should not be considered as ends in themselves. Rather P values should be considered as means to achieve ends. There is nothing magical about a P value less than 0.05. Problems of multiplicity cannot be solved by using confidence limits or odds ratios in place of P values. Confidence limits and odds ratios have their own limitations. Confidence limits are just another expression of the standard deviation. Odds ratios replace arithmetic differences between sample means with geometric differences between sample means. Odds ratios may be statistically significant but practically irrelevant. Doubling the frequency of a rare event remains a rare event. Odds ratios can hide the number to treat to achieve a single benefit or the cost of achieving a single benefit. Researchers need to consider what statistical methods are most appropriate for the questions being asked. Single primary endpoints should be used whenever practical. If multiple primary endpoints are necessary, then methodology to control Type 1 error must be chosen prior to performing the study.

Keywords: statistics, p value, multiplicity, type 1 error

Article citation: Berdine G. Multiplicity and statistical significance. The Southwest Respiratory and Critical Care Chronicles 2020;8(34):61–63
From: Department of Internal Medicine, Texas Tech University Health Sciences Center, Lubbock, Texas
Submitted: 3/28/2020
Accepted: 4/7/2020
Conflicts of interest: none

References

1. Harrington D, D'Agostino RB, Gatsonis C, et al. New guidelines for statistical reporting in the journal. N Engl J Med 2019; 381:285–6.
2. Empirical rule. Wikimedia Commons https://upload.wikimedia.org/wikipedia/commons/a/a9/Empirical_Rule.PNG. Accessed 4/7/2020.
3. Dmitrienko A, D'Agostino RB. Multiplicity considerations in clinical trials. N Engl J Med 2018;378:2115–22.
4. Multiple endpoints in clinical trials. Food and Drug Administration https://www.fda.gov/media/102657/download. Accessed 4/7/2020.