Bonferroni and Šidák Corrections for Multiplicity Effects with Subtotal Scores of Comparison Question Polygraph Tests

Raymond Nelson

Abstract

The problem of multiple statistical comparisons is discussed as it applies to the use of subtotal scores of comparison question polygraph tests. Multiplicity phenomena, including inflation of alpha when any of a set of multiple subtotal scores are used to make deceptive classifications, and deflation of alpha when all of a set of multiple subtotals are used to make truthful classifications of test results. Common statistical corrections, including the Bonferroni correction and Šidák correction are described. Mathematical examples are provided to illustrate the application of these statistical corrections to the comparison question polygraph test.

Keywords: *scoring, test data analysis, polygraph techniques, normative data, reference distributions, statistical significance, Bonferroni, Šidák, multiple comparisons, inflated alpha, deflated-alpha, multiplicity*

Introduction

Multiplicity effects, also known as the problem of multiple comparisons (McDonald, 1996; Miller, 1981), are well known to scientists, researchers, statisticians and other professionals whose work involves the evaluation of data as a basis for classification and inference. These effects have also been referred to as the "look elsewhere" effect (White, 2011), because of the impulse or desire to continue to look elsewhere when we do not initially find what we are looking for. In the context of scientific research and testing, multiplicity effects, and the impulse to keep looking elsewhere until we find what we are looking for, can be thought of as a manifestation of a confirmation bias described by Nickerson (1998). We ignore results when we are unsatisfied and continue searching until we find a result with which we are satisfied.

A card-playing analogy can be useful to better understand the practical implications: image a poker player who deals himself a hand of cards with the goal of doing so repeatedly until he gets a Royal Flush. Probability theory tells us that with a sufficient number of trials, the odds will accumulate to a sufficiently high level that we are likely to eventually observe its occurrence. But, assuming a fair and unbiased deck of cards, it will be a mistake to attempt to infer that the deck of cards has any special characteristics or that the player has any unique attributes that caused the Royal Flush to occur. Instead, the occurrence of the Royal Flush is simply a function of continuing to look elsewhere (in subsequent hands of cards) for its occurrence. Similarly, looking repeatedly at any scientific dataset can confound our attempts to make realistic and accurate inferences about significance or meaning when we eventually observe what we are looking for. More specifically, multiplicity effects are the compounding of error probabilities. They can result in a loss of accuracy or precision and corresponding increase in classification error.

Discussion

Multiplicity effects play a role in comparison question polygraph examinations when using subtotal scores to classify the results as deceptive or truthful. Subtotal scores for individual relevant questions have been shown to be an effective basis for deceptive classifications when the grand total score is inconclusive (Senter, 2002; Senter & Dollins, 2003; 2008). But polygraph techniques that make use of grand total scores have consistently produced higher accuracy rates than techniques for which decisions are based solely on subtotal scores (APA, 2011). Subtotal scores have been the traditional basis with which to classify the results of multiple issue screening exams (Department of Defense, 2006a, 2006b) when hand scoring. The grand total hand scores are not traditionally used in multiple issue screening tests.

Use of polygraph subtotal scores as a

basis for statistical classification and inference will introduce known and predictable mathematical and statistical increases to the probability of error unless corrections are applied. These effects occur because every imperfect and non-deterministic test result is a probabilistic result. There is always some associated probability that the result is correct and incorrect. The test error estimate will be an aggregation of the errors for all probabilistic results used to classify the test result.

Inflation of alpha for deceptive results of event-specific diagnostic polygraphs.

Use of subtotal scores in event-specific examinations, for which one classification will be made at the level of the test as a whole, introduces multiplicity into the statistical model. It amounts to the practice of making multiple statistical decisions regarding a single classification. When making multiple probabilistic judgements regarding a single target incident or allegation, for which any deceptive subtotal result will result in the classification of the examination as deceptive, the resulting probability of error is the cumulative or additive probabilities of error for all subtotal probability scores. In the case of an event-specific diagnostic polygraph with three relevant questions (RQs) and alpha = .05, the total error probability can be determined by summing the alpha levels for all RQs (.05 + .05 + .05 = .15). Calculations indicate a potential for a 15% error rate even though the test is conducted with alpha = .05, with the goal of constraining errors to a rate less than 5%. This has sometimes been referred to as the problem of *inflated alpha* because of the predictable increase in test errors. Left unmanaged, inflated alpha can result in a false positive error rate that is potentially several times greater than that which was intended or anticipated. So while the goal was to constrain false positive errors to 5%, the practice of using subtotals increased that false positive error rate to about 15%.

Bonferroni correction.

Fortunately, the problem of inflated alpha is only mildly vexing and is quite easily rectified through the use of a simple statistical correction - the *Bonferroni correction* (Abdi, 2007), named for famous Italian statistician Carlo Emilio Bonferroni (1892-1960). The Bonferroni correction is calculated by dividing the desired alpha level by the number of statistical decisions. The number of statistical decisions is equal to the number of subtotal scores which is the same as the number of RQs. The resulting corrected alpha level is referred to as the *Bonferroni corrected alpha*.

For event-specific diagnostic polygraph with three RQs and a desired alpha of .05 we divide .05 by three (alpha = .05 / 3 RQs = alpha .0167 per RQ). It will be necessary to use the Bonferroni corrected alpha = .0167 for each of the three subtotal scores. When these per question error probabilities accumulate (.0167 + .0167 + .0167 = .05) the total cumulative margin of error for the test will be alpha = .05. The error estimate will be constrained to within the desired range of less than 5%. Event-specific diagnostic exams with two RQs will require the use of Bonferroni corrected alpha = .025. This is because alpha = .05 / 2 RQs = .025 per RQ, and this will accumulate to .025 + .025 = .05. Similarly, event-specific diagnostic exams with four RQs will use Bonferroni corrected alpha = .05 / 4 = .0125 per RQ, which will accumulate to .0125 + .0125 + .0125 + .0125 = .05. Because subtotal scores are not used to make truthful classifications for event-specific diagnostic exams, no statistical correction is needed for truthful classification for these type of exams.

Deflation of alpha for truthful results of multiple-issue screening exams.

Multiple-issue screening exams make use of subtotal scores for both deceptive and truthful classifications. This is accomplished with the "any or all" rubric which states that *any* subtotal result that is deceptive will be sufficient to classify the exam result as deceptive, whereas *all* subtotal results must indicate truth-telling in order to classify the overall exam result as truthful. Results are inconclusive whenever one or more of the subtotal scores are not statistically significant for truth-telling and none of the subtotal scores is statistically significant for deception. As with event-specific diagnostic exams the test error statistic for a multiple-issue screening polygraph is a function of the number of subtotal scores (the number of relevant questions). This phenomena applies to all forms of testing that involve multiple statistical comparisons.

With event-specific diagnostic exams all relevant questions describe details related to a single allegation or incident. Relevant questions for multiple-issue screening polygraphs will describe different behavioral issues with a strong assumption of independence. The independence assumption is not premised

solely on the use of different action verbs or semantic content for each relevant question. It also involves the assumption that an examinee could engage in one or more behaviors while conceivably remaining completely uninvolved in other behaviors. (This independence assumption is not used with event-specific diagnostic polygraph for which the all relevant questions describe aspects of a single allegation or incident.)

The independence assumption is said to be a strong assumption because, in reality, although the target *behaviors* might be assumed to be independent or unaffected by one another, even though the examinee´s *responses* to multiple issue polygraph stimulus questions are *not* completely independent. Responses to different target stimuli can affect one another within an exam. This is because all responses to multi-issue screening stimuli have an important source of shared variance - the examinee. The fact that responses are not completely independent appears to be the basis of the need for the "any or all" rubric and for traditional prohibitions against attempting to make both deceptive and truthful classifications within a single examination. Because the any or all rubric does not allow both truthful and deceptive results, it will eliminate the potential to observe both false positive and false negative errors within a single examination. Instead, observed testing errors will be in the form of either false-positive or false-negative errors, for which we can constrain their occurrence to desired levels.

Because the target issues for multiple-issue screening exams are treated independently, there is no great concern that we are subjecting a single target issue to multiple statistical decisions. Screening tests are intended to identify possible problems that can be subsequently evaluated in more thorough detail, and test sensitivity is therefore an important concern. Statistical corrections are not used when there is a potentially costly loss of sensitivity that would reduce the test effectiveness (McDonald, 2009). For these reasons, Bonferroni correction is not used to make deceptive classifications for multiple-issue screening polygraphs. Deceptive classifications of multiple-issue screening polygraphs are made with the uncorrected alpha boundary.

Multiplicity plays an important role in *truthful* classifications for multiple-issue screening polygraphs, but in a slightly different way. Truthful classifications are made when the observed data differ at a statistically significant level from the statistical reference distributions for deceptive cases. Alpha for truthful classifications therefore represents the tolerance for risk or error that a deceptive person may be classified as truthful in a multiple issue screening test (a false-negative error). Quite obviously, most deceptive persons can be expected to produce deceptive test scores, and the proportion of deceptive persons that produce a test question score that is statistically significant for truth-telling (i.e., differs at a statistically significant level from the normative reference distributions for deceptive cases) is expected to be observed at the defined alpha level (.05). Perhaps equally obvious is the fact that the proportion of deceptive persons that produce *two* statistically significant truthful scores in a test with two relevant questions will be lower than the proportion of deceptive persons who produce only one statistically significant truthful score. Similarly, the proportion of deceptive persons who produce three out of three truthful scores, or four out of four truthful scores, can be expected to be even lower. This phenomena can be thought of as the *deflation of alpha* that occurs as a result of the requirement that the examinee pass *all* questions in order to pass the test. Deflation of alpha will result in a reduction of the observed false-negative error rate to something predictably lower than the established alpha tolerance for error.

Deflation of alpha will reduce testing errors for deceptive classifications, but will also have an effect on truthful classifications. The requirement that all subtotal scores are statistically significant for truth-telling will effectively provide the truthful examinee with multiple opportunities to not produce a statistically significant truthful score. This is a simple feature of the fact that all tests are probabilistic and not deterministic, and that probabilities can be cumulative under these circumstances. For this reason, the requirement for statistically significant truthful scores for *all* subtotals can be expected to cause a substantial *inflation of inconclusive* results for truthful persons, along with a corresponding substantial reduction of test specificity for truth-telling - unless a statistical correction is used.

Šidák correction.

The preferred statistical correction for truthful classifications of multiple-issue screening polygraphs is not the Bonferroni

correction but is instead a related procedure called the *Šidák correction* (Abdi, 2007, Šidák, 1967). The Šidák correction is named for Zbyněk Šidák (1933-1999), a renowned Czech statistician. It is an exact version of the simple Bonferroni correction that is better suited to the context of multiple independent classifications. Calculation of the Šidák correction is thus: $1-(1-\text{alpha})^{\text{number-of-decisions}}$. The Šidák correction is the mathematical compliment of the compliment of the alpha raised to number of decisions. As with the previously described Bonferroni correction, the number of decisions is equal to the number of subtotal scores which is also equal to the number of relevant questions.

The normal form of the Šidák correction is used to calculate the inflation of alpha. But we are concerned with the deflation of alpha, so it will be the inverse of the Šidák correction that is used calculate this deflation. The inverse of the Šidák correction is calculated using the following equation: $1-(1-\text{alpha})^{1/\text{number-of-decisions}}$. The inverse Šidák is the mathematical compliment of the compliment of the alpha raised to the inverse of the number of decisions.

To demonstrate the application of the Šidák correction to adjust or correct the alpha boundary for the number of relevant questions, consider the following example: a multiple issue polygraph with 4 relevant questions for which alpha = .05 will give the following uncorrected, deflated, alpha level: $1-(1-.05)^{1/4} = .0127$. That means that instead of constraining false negatives to our desired 5% we actually constrain them to 1.27%. This will result in a corresponding increase in inconclusive truthful cases. Correcting for this will involve first calculating the corrected alpha boundary using the normal form: $1-(1-.05)^4 = .1854$. Use of the Šidák corrected alpha = .1854, will give the following: $1-(1-.1854)^{1/4} = .05$. This will preserve the test specificity to truth-telling for multiple-issue screening polygraphs at acceptably high levels, while reducing the occurrence of inconclusive results for truthful persons. It will also constrain the occurrence of false-negative test errors to rates that are within the tolerance level expressed by the alpha = .05 level.

In practice, statistical corrections can be applied to either the alpha boundary or to the p-values using either the normal or inverse forms. However, correction of p-values can only be accomplished after conducting and scoring an examination, whereas correction of alpha boundaries can be accomplished prior to the conduct of an examination. This is accomplished by using the subtotal score with a p-value at or below the corrected alpha as a decision threshold.

Conclusion

All scientific testing is a process of classification and inference. Classification, in this case, refers to the formulation of a simple categorical test result. Inference is the process of calculating a statistical or probabilistic estimate of the likelihood that an error has occurred. In a more abstract sense, the purpose of scientific testing is to evaluate and quantify an amorphous phenomena that cannot be subjected to simple and perfect deterministic observation or to direct physical/linear measurement. Deterministic observation requires the existence of some phenomena that is uniquely and perfectly associated with the thing we want to evaluate. This would be theoretically perfect, and would also obviate the need for testing. Physical measurement, in contrast, is near perfect, though still subject to mechanical measurement error, and would require two things: 1) a physical substance to measure, and 2) a well-defined unit of measurement. Scientific tests are inherently probabilistic - they are neither deterministic nor an actual physical measurement. Scientific tests are not expected to be perfect. They are expected to quantify the probabilistic margin of uncertainty surrounding a conclusion. Good scientific tests will do this in manner such that the predicted proportions of testing errors concurs reasonably with the observed evidence of testing errors. Multiplicity effects have a potentially serious impact on the accuracy of test error estimates. The use of statistical corrections can be an important part of the validity and effectiveness of a test method.

Two core ideas underlie all scientific tests and experiments. The first core idea is that all scientific conclusions or hypotheses are relative to some alternative. Professionals who make scientific conclusions are expected to articulate the alternatives and to use probability theory to weigh the evidence. The second core idea is that all conclusions and hypotheses must be stated as statistical or probabilistic hypotheses in order to be quantifiable. Conclusions or hypotheses that cannot be stated as statistical hypotheses cannot be measured or tested, and are therefore not sci-

entific. Unscientific ideas that portend to be scientific can be said to be pseudoscience.

Related to the need for testable hypotheses is the need to make *a priori* declarations about the tolerance for error and required alpha level for statistical significance. Field practitioners generally do not themselves decide on alpha boundaries or numerical cutscores - these are most often a matter of agency policy and are developed around the needs specific to the risk management context. Field practitioners themselves are also not expected to calculate statistical formulae themselves. Instead, they commonly use published statistical reference tables for which calculations have been previously computed for all possible test results.

If the polygraph test is merely a tool to amplify or enhance an interrogation or interview, then examiners need not ever account for or explain the test results. If this were the case they need not even score the test, and certainly need not learn about probability theory and statistical phenomena. Similarly, polygraph examiners will never be expected to account for or explain a test result if a confession is obtained for every deceptive test result without fail. If the information from the pretest and posttest discussions are the sole purpose of the polygraph test then there would be no need to ever provide a test result. If, however, there is ever a need to explain a test result or account for the level of certainty or uncertainty that should be attributed to a test result, examiners might be obligated to numerically score and statistically quantify the test result. Examiners who are unprepared to do this will be vulnerable to professional embarrassment, either due to an inability to provide evidence based computations of the expected test precision and error rate, or due to frustration when it is eventually discovered that polygraph results are probabilistic and imperfect despite a feigned attitude of certainty.

Examiners who are prepared to account for test results using the basic principles and concepts of statistics and probability and theory will be better prepared to make favorable professional impressions while discussing test results without the sense of insecurity that stems from naive expectations for deterministic perfection from a probabilistic test. Although there will always be practical value in in the information that can be obtained from the polygraph pretest and posttest interviews, test results without realistic computations of statistical error estimates will, in the end, be of no real value.

Ultimately, all test scores, including both grand total and subtotal scores of comparison question polygraph tests, will have an associated probability of error. Probability theory informs us that error rates are predictably cumulative whenever we attempt to make multiple statistical comparisons within a single test or experiment. While mildly concerning, the predictability of multiplicity phenomena means that we can also apply the principles of probability theory to statistically correct for multiplicity effects - if we understand the principles of probability. While very simple calculations such as the Bonferroni correction can be easily managed in field settings, field practitioners should be relieved of complex calculations such as the Šidák correction through the inclusion of statistically corrected information in published normative reference tables. Use of computer algorithms can also accomplish the application of these statistical corrections with automated reliability. Although many researchers, statisticians and scientists will prefer to use omnibus statistical methods such as ANOVA and other methods to simultaneously test multiple statistical hypothesis without the introduction of multiplicity effects, Bonferroni correction and Šidák correction are two classical solutions to the well know problems of multiplicity. They are well suited to the analysis and interpretation of comparison question polygraph test results.

References

Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. In N.J. Salkind (Ed.), *Encyclopedia of Measurement and Statistics.* Sage.

Department of Defense (2006a). *Federal psychophysiological detection of deception examiner handbook.* Reprinted in Polygraph, 40 (1), 2-66.

Department of Defense (2006b). *Test data analysis: DoDPI numerical evaluation scoring system.* [Retrieved from http://www.antipolygraph.org on 3-31-2007].

McDonald, J. H. 2009. *Handbook of Biological Statistics 2nd ed..* Sparky House Publishing, Baltimore, Maryland.

Miller, R. G. (1981). *Simultaneous Statistical Inference 2nd Ed.* Springer Verlag New York.

Nickerson, R. S. (June 1998). "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises". Review of General Psychology 2 (2): 175–220.

Šidàk, Z. (1967). Rectangular confidence region for the means of multivariate normal distributions. *Journal of the American Statistical Association, 62,* 626–633.

White, L. A. (August 12, 2011). *Word of the Week: Look Elsewhere Effect.* Sanford National Accelerator Laboratory.