Behavior and Social Issues, 24, 164-186 (2015). doi: 10.5210/bsi.v.24i0.6048

PROBABILITY PYRAMIDING REVISITED: UNIVARIATE, MULTIVARIATE, AND NEURAL NETWORK ANALYSES OF COMPLEX DATA

Chris Ninness[1]
Behavioral Software Systems

Robert Henderson
Stephen F. Austin State University

Sharon K. Ninness
Texas A&M University–Commerce

Sarah Halle
Behavioral Software Systems

ABSTRACT: Historically, behavior analytic research and practice have been grounded in the single subject examination of behavior change. Within the behavior analytic community, there remains little doubt that the graphing of behavior is a powerful strategy for demonstrating functional control by the independent variable; however, during the past thirty years, various statistical techniques have become a popular alternative form of evidence for demonstrating a treatment effect. Concurrently, a mounting number of behavior analytic investigators are measuring multiple dependent variables when conducting statistical analyses. Without employing strategies that protect the experimentwise error rate, evaluation of multiple dependent variables within a single experiment is likely to inflate the Type I error rate. In fact, with each additional dependent variable examined in univariate fashion, the probability of "incorrectly" identifying statistical significance increases exponentially as a function of chance. Multivariate analysis of variance (MANOVA) and several other statistical techniques can preclude this common error. We provide an overview of the procedural complications arising from methodologies that might inflate the Type I error rate. Additionally, we provide a sample of reviewer comments and suggestions, and an enrichment section focusing on this somewhat contentious issue, as well as a number of statistical and neural network techniques that enhance power and preclude the inflation of Type I error rates.
KEYWORDS: Type I error rate, multivariate analysis, univariate analysis, inflation, experimentwise error rate, neural network, external validity

Behavior analytic research is grounded in the systematic observation of the single participant and an experimental preparation in which graphing the participant's baseline and treatment behavior provides confirmation of findings. Contrasting baseline behavior with behavior change occurring during various treatment conditions (and following treatment conditions) demonstrates functional control by the independent variable and provides evidence

[1] Correspondence should be addressed to Chris Ninness, Behavioral Software Systems, 2207 Pinecrest Dr., Nacogdoches, Texas 75961. E-mail: cninness@suddenlink.net

for experimental effect. This special concentration on the individual participant's moment-by-moment behavior across conditions stands in contrast with group designs employed in most of psychology where the emphasis is on assessment of the "average effect" of a treatment (or treatments) within or between experimental and control groups (Roscoe, 1975). As forwarded by Ninness et al., (2002), "Sidman (1960) points out that in traditional psychological research, subject variability is considered a source of experimental *error*, while in behavioral research it is a source of experimental *interest*" (p. 64).

Although behavior analytic research remains firmly grounded in single-subject design, an increasingly common practice for behavior analysts is to conduct studies in which the evidence for experimental effect lies in determining the level of statistical significance (Ninness et al., 2002). Indeed, it has become fairly common for behavior analytic researchers (as well as researchers in many related disciplines) to examine more than one dependent measure and to analyze data by way of traditional parametric and nonparametric inferential statistics (Leary & Altmaier, 1980). As described by Stevens (2009), a researcher might be comparing two methods of teaching reading to second-grade students. Upon completing the intervention protocol, the researcher measures performance relating to basic foundations of "…syllabication, blending, sound discrimination, vocabulary, comprehension, and reading rate" (p. 145). Such a study entails the assessment of multiple dependent variables analyzed within the same experiment.

Concepts and Terminology

Before discussing the above hypothetical experiment and the manner in which its findings might be analyzed, a few concepts and terms should be addressed. Tukey (1949) first used the term "experimentwise error rate" as a reference to the level of risk for obtaining a false positive error when examining all the hypotheses within a group of potentially accurate hypotheses. The term "familywise error rate" refers to the same type of risk for arriving at one or more erroneous statistical conclusions regarding statistical significance existing somewhere among all of the hypotheses being tested within a given experiment (see Šidák, 1967, for a discussion). Statistical techniques focused on controlling the experimentwise/familywise error rate are said to be "conservative" in guarding against the possibility of inadvertently rejecting the null hypothesis (a Type I error). However, employing these conservative statistical techniques reduces the risk of a false positive determination at the risk of failing to identify statistically significant findings. The term "false discovery rate" refers to a group of statistical techniques that are more liberal in allowing researchers to identify statistically significant outcomes, but these techniques exercise less rigorous control over the experimentwise error rate and thus increase the likelihood of a Type I error (Benjamini & Hochberg, 1995). These strategies are said to have greater statistical power in the search for outcomes that could not have occurred simply by chance. Hochberg and Tamhane (1987) provide a classic guide to multiple comparison techniques that improve statistical power while maintaining reasonable safeguards against inflating the Type I error rate.

Analyzing Multiple Variables within a Single Experiment

In the hypothetical experiment described above by Stevens (2009), the multivariate analysis of variance (MANOVA) is robust with regard to determining the experimentwise error rate. Subsequent to identifying statistical significance that addresses the experimentwise error rate, a series of multiple comparison tests are conducted to identify the particular variables showing significant differences among all possible hypotheses being tested. Beyond protecting the

experimentwise error rate, Stevens (2009) provides additional reasons for employing multivariate techniques when analyzing several dependent variables within an experiment:

> 1. Any worthwhile treatment will affect the subjects in more than one way. Hence, the problem for the investigator is to determine in which specific ways the subjects will be affected, and then find the sensitive measurement techniques for those variables.
>
> 2. Through the use of multiple criterion measures we can obtain a more complete and detailed description of the phenomena under investigation, whether it is teacher method effectiveness, counselor effectiveness, diet effectiveness, stress management technique effectiveness, and so on.
>
> 3. Treatments can be expensive to implement, while the cost of obtaining data on several dependent measures is relatively small, and maximizes information gain. (p. 2)

Big Trouble in Small-*n*

The fact that in many experimental settings, behavior analysts elect to utilize inferential statistics rather than single-subject methodology is not at issue here. Research methodology is often a matter of funding authority requirements, personal design preference, or a researcher's particular ambitions for demonstrating external validity. However, it has become common for researchers across the behavioral and social sciences to employ univariate statistical tests when examining multiple concurrent dependent variables within group designs (Cafri, Kromrey, & Brannick, 2010). This practice becomes particularly problematic when the researcher employs a limited number of scores within an experiment that is measuring several dependent variables concurrently. As discussed in Ninness, Rumph, Vasquez, & Bradfield (2002), "…when small-*n* measures are taken and several dependent variables are analyzed using traditional univariate statistics, the internal validity of the study is seriously compromised" (p. 65). That is, while the researcher may find a number of dependent variables that appear to be most impressive in terms of their apparent levels of statistical significance (*P*-values), the actual probabilities associated with these outcomes are not at all what they appear. Neher (1967) notes that statistical findings calculated with multiple dependent variables but analyzed by way of a series of univariate tests within the same experiment are, in fact, inflated as a function of "probability pyramiding." Neher states:

> Reporting the 5 percent level for a finding means that there is only a 5 percent chance that it is a spurious finding resulting solely from chance variations. If, however, two independent analyses are done, the probability that at least one such analysis will yield a spurious, significant finding at this level is greater than 5 percent. (The assumption of independence of the two analyses, while not always true, simplifies the discussion without introducing serious error.) To determine the new probability level, one may calculate the probability that a significant result would not be obtained in either of the two tries (.95 X .95) and then subtract this from 1. Thus, $1-(.95)^2 = 1-.902 = .098$. If three independent analyses are done, the real level becomes $1-(.95)^3 = 1-.857 = .143$. (Each individual analysis increases the probability pyramiding, even though it may be part of one large 'analysis', such as stepwise multiple regression, item analysis, etc.). (p. 259)

Pyramiding Probabilities

Consistent with the example provided by Stevens (2009), consider a circumstance in which five dependent measures are employed in the same study, and separate univariate *t*-tests are performed on each dependent variable. Here, multiple dependent variables are analyzed separately in univariate fashion, and each analysis fails to take the other into account. In such a cavalier statistical analysis, the use of univariate *t*-tests inflates the overall Type I error rate, and the investigator/s are likely to conclude that treatments are effective in conditions where pure chance is operating. In an example described by Ninness et al., (2002), "Employing 5 concurrent *t*-tests, the probability of no Type I errors is: (.95)(.95)(.95)(.95)(.95) ≈ .774 since the chances of not making a Type I error for each test is .95. In this example, the likelihood of making at least one Type I error is 1 -- .774 ≈ 0.226" (p. 65). When this probability pyramid is developed from the top down, each dependent variable adds yet another layer of potential inaccuracy to the experimentwise sampling error and increases the researcher's likelihood of obtaining one or more false positives. With each dependent variable added to the analysis, we see layers of the pyramid expanding exponentially as the Type I error rate expands. With five dependent variables, the researcher's likelihood of inflating $\boldsymbol{\alpha}$ has expanded from 0.05 to 0.226 (i.e., $1 - .95^{5} \approx 0.226$).

Table 1 shows the gradual progression of false positive *P-values* for a series of concurrent univariate tests. Examining a researcher's chances of obtaining a false positive when employing one dependent variable, the actual probability is, as it should be, at the 0.05 level. However, as the researcher pursues additional univariate tests within the same experiment, his/her likelihood of finding a significant difference somewhere among all the existing possibilities begins to inflate exponentially. If, for example, the researcher were willing to entertain the idea of examining ten concurrent univariate tests within the same experiment, which is not an unusual occurrence, the process of probability pyramiding becomes all too apparent. As shown in Table 1, the researcher has a better than 40 percent chance of finding at least one false positive when examining all ten possible univariate tests within the same experimental preparation (Stevens, 2009).

Monte Carlo Experiments and Type I Errors

Randomization tests (also described as permutation tests) are computationally intensive resampling techniques in which the *P*-values are repeatedly calculated for all possible levels of significance that exist within a specific data distribution (Edgington, 1995; Good, 1994; Ninness et al., 2002). The "obtained" *P*-values are the proportion of all possible data arrangements greater than the actual/obtained value for a particular experimental finding.

Previous to the early 1980's, randomization tests (and related resampling techniques) were not commonly employed since the computations often required more processing speed than the existing technology could provide. The mid-eighties witnessed a dramatic increase in computer power, and with this computing power came calculation-intensive applications for previously unimaginable large and convoluted datasets. Many of these datasets were composed of non-normal, nonlinear, and nonindependent variables. By 2000, the analyses of computer simulated datasets became a highly specialized area of research in its own right. In 2002, we, (Ninness et al.), developed several Monte Carlo datasets with two dependent variables in each set. Significance tests were conducted by way of a traditional multivariate procedure (Hotelling's T^{2})

Table 1. *Increasing Number Dependent Variables with Inflation of the Experimentwise Error Rate*

Dependent Variables	Experimentwise (Type I) Error Rate
1	$1 - .95^{1} \approx 0.0500$
2	$1 - .95^{2} \approx 0.0975$
3	$1 - .95^{3} \approx 0.1426$
4	$1 - .95^{4} \approx 0.1855$
5	$1 - .95^{5} \approx 0.2262$
6	$1 - .95^{6} \approx 0.2649$
7	$1 - .95^{7} \approx 0.3016$
8	$1 - .95^{8} \approx 0.3365$
9	$1 - .95^{9} \approx 0.3697$
10	$1 - .95^{10} \approx 0.4013$

and by way of randomization tests. Had we not adjusted their obtained *P*-values by way of a Bonferroni algorithm within the randomization test procedure, all of the *P*-values would have shown seriously inflated Type I error rates. However, when we implemented the Bonferroni adjustment within our randomization algorithm, the "…correlations between Hotelling's T^2 and randomization tests were at .95, .95, .94, and .93, for group sizes of 12, 10, 8, and 6, respectively" (p. 71). Importantly, since the Bonferroni adjustment procedure was incorporated within our randomization test algorithm, we found very little difference in obtained *P*-values when employing Hotelling's T^2 versus randomization tests.

The Empirical Pyramid

Beyond the findings in the above Monte Carlo experiments, an extremely large body of empirical evidence exists demonstrating that with continuing univariate analyses of dependent variables, there is a conspicuous inflation of the Type I error rate (cf. Austin & Brunner, 2004). Each successive univariate test employed in an attempt to locate yet another statistically significant difference among variables *within a single experiment* makes it more likely that a finding of significance is a function of sampling error rather than a function of treatment (see Table 1). Just as described by Neher (1967), Stevens (2009), and Edgington (1995), the more univariate tests conducted within a single study, the more inevitable it becomes that we will see inflation of the probability pyramid (cf. Leary & Altmaier, 1980).

As ironic as it is exasperating, we now have a virtual library of empirical studies demonstrating that the inflation of Type I error rates has become an insidiously pervasive problem throughout much of the current behavioral science literature. For example, Baldwin, Murray, and Shadish (2005) reviewed 33 publications employing large group experimental designs and found that, as a result of violating key assumptions concerning independence of

observations, researchers of the 33 studies examined from peer-reviewed journals had conducted their statistical methodologies incorrectly. Armstrong and Henson (2005) reviewed 54 articles published in the *International Journal of Play Therapy* and found a continuous stream of statistical inaccuracies inflating Type I error rates. Subsequent to recommendations disseminated by the American Psychological Association's Task Force on Statistical Inference (APA, 1996), Schatz, Jay, McComb and McLaughlin (2005) reviewed the accuracy of statistical analyses in neuropsychology investigations published within the *Archives of Clinical Neuropsychology.* The authors found inflated Type I error rates in 275 of 406 scrutinized publications. They describe the incorrect use of null hypothesis testing and associated incorrect determination of *P*-values as a major factor contributing to investigators' likelihood of inflating the Type I error rate. Schatz et al. (2005) state:

> Such error usually occurs because of incorrect statistical procedures and inappropriate emphasis on *P*-values. In most cases, researchers treat each statistical test individually, instead of examining the results as a whole, demonstrating a lack of control for Type I error. Performing multiple ANOVAs or following a MANOVA with univariate ANOVAs without adjusting the alpha level… . (p. 1054)

Schatz et al. (2005) forward the view that neuropsychologists who conduct empirical investigations and publish results in refereed journals must become much more attentive to the critical assumptions and theories within the area of inferential statistics. While this study revealed an unacceptable level of experimentwise error, the *Archives of Clinical Neuropsychology* should be commended on its willingness to audit its studies and bring this pervasive academic complication to the attention of the behavioral science community.

Threats to External Validity

As described above, the mathematical logic for attempting to obtain significant *P*-values revolves around generating a body of evidence showing that particular outcomes could not have occurred simply by chance. If significant *P*-values are obtained, this supports the notion that subsequent interventions conducted in accordance with the same protocol are very likely to obtain the same (or very similar) levels of statistical significance. Unfortunately, it appears that researchers who employ multiple dependent variables within the same experiment fail to recognize that analyzing concurrent univariate findings inevitably inflates the experimentwise Type I error rate (Tatsuoka, 1973). These researchers overlook the critically important point that any variable taken in isolation may affect the criterion differently from the way it will act in the company of other variables. To reiterate, when one or more of the multiple dependent variables, analyzed univariately, is identified as reaching statistical significance, the external validity is compromised (Fish, 1988). This may be one of the strongest arguments against conducting multiple univariates tests within a single experiment without a strategy that protects the familywise error rate. One must ask, "What is the point of claiming statistical significance if subsequent investigations that attempt to replicate a researcher's published protocols are unlikely to obtain similar findings of significance?"

Alternative Perspectives on Type I Error

Despite the extensive evidence described above, there remains some disagreement regarding the inflation of alpha within and across disciplines, and researchers with differing research perspectives have made salient arguments for examining univariate outcomes within the confines of a given experiment. For example, Huberty and Morris (1989) offer four circumstances in which conducting a series of ANOVAS within a single experiment might be a reasonable/logical approach:

> Multiple ANOVAS might be conducted to (a) study the effects of some treatment variable or variables on conceptually independent outcome variables; (b) explore new treatment-outcome variable bivariate relationships; (c) reexamine bivariate relationships within a multivariate context; and (d) select a comparison group in designing a study. (pp. 303-304)

It is important to note that Huberty and Morris are not advocating wholesale and unconstrained multiple univariate tests within experiments. Indeed, these authors propose a wide range of supportive statistical procedures when examining multiple dependent variables within a given study. These authors emphasize that concurrent univariate tests measured within a single experiment should be accompanied by a report of all the intercorrelations among the variables under consideration. Huberty and Morris state, "Typically, these correlations would be reported in the form of a matrix" (p. 307). Undoubtedly, providing sets of correlation matrices within studies that employ univariate multiple comparisons would allow readers to determine subjectively the degrees to which dependent variables may be orthogonal or correlated to one another; however, there are more parsimonious and efficient strategies for analyzing such data. Hochberg and Tamhane (1987) provide a classic guide to conducting multiple comparison techniques while guarding against the inflation of the Type I error rate. One of the most robust solutions for precluding inflation of Type I error rate is offered by Leary and Altmaier (1980):

> The solution to the problem of inflated error rates with multivariate studies (those having more than one dependent variable) is to analyze the data using multivariate techniques; the most common examples are Hotelling's T^2 and multivariate analysis of variance (MANOVA), multivariate analogues of the t test, and ANOVA, respectively. A multivariate analysis is, as its name implies, a statistical method that allows the simultaneous consideration of more than one dependent variable (Kleinbaum & Kupper, 1978), as distinguished from the more commonly used univariate analyses (t test, ANOVA, chi-square), that can handle only one dependent variable at a time. By considering several dependent variables at once, multivariate analyses allow the researcher to hold the probability of making one Type I error at alpha. (p. 613)

Well beyond the details described above, our appendix provides enrichment material in the form of a more mathematically rigorous account of multivariate confidence intervals, the Bonferroni Theorem, and related variables influencing changes in the experimentwise error rate within applied univariate and multivariate statistics.

Reviewer Comments and Suggestions

In the material below, we provide several excerpts from hypothetical/simulated reviews focusing on issues that have come into question fairly often in the course of conducting actual reviews for several journals, including, but not limited to, *Behavior and Social Issues*.

Failing to Find Significant Differences

In the excerpt below, the simulated submission included a strategy to establish "sameness of groups previous to treatment" by demonstrating that the null hypothesis could not be rejected when contrasting potentially contaminating variables. This strategy is not entirely unusual; on several occasions, researchers have attempted to demonstrate that variables such as age, political affiliation, academic performance level, ethnic origin, etc., are not experimental artifacts that might influence participants' behavior by conducting *t*-tests (or other univariate procedures) and then "failing to reject the null hypothesis." As the reviewer indicates below, failing to reject the null hypothesis does not provide evidence that the variables being analyzed are, in any way, equivalent to one another.

Reviewer Comments: In selecting participants for group assignment, it appears that *failure to find statistically significant difference*s was employed as a strategy for showing that participants are the same with regard to particular physiological characteristics. For example, on p. 19 within the Results section, the authors indicate:

> We found no statistically significant differences with regard to physiological attributes in terms of body mass index, $t(2, 12) = 0.914$, $p > .05$, resting blood pressure, $t(2, 12) = 0.1968$, $p > .05$, and serum cholesterol in mg/dl, $t(2, 12) = 0.06276$, $p > .05$.

I think it is important to be very clear about what it means to fail to reject the null hypothesis. Forgive the annoying use of jargon and double negatives, but this is the foundational logic of normal curve theory and null hypothesis testing, and we are obligated to employ it properly. Failing to reject the null hypothesis does not in any way suggest that groups demonstrate equivalent physical characteristics or performances. Failing to reject the null hypothesis only means that one cannot rule out the chance that the differences between groups could be a function of sampling error. Thus, I can see no methodological advantage regarding the inclusion of any of these details within a revised version of this manuscript. There is, however, an alternative statistical strategy that might be employed if the authors of this study are committed to demonstrating that groups are essentially the same with regard to potentially contaminating variables. "Equivalence testing" is a valuable and increasingly popular statistical technique that the researcher uses to determine an interval where the means can be considered as being the same or "equivalent." If the authors are inclined to employ some form of statistical analysis demonstrating sameness regarding a particular attribute, I believe they will find this *test of equivalence* approach especially helpful (Chow & Liu, 2000; Wellek, 2002).

Employing a Series of Univariate *t*-Tests

As indicated above, the probability of finding statistically significant differences (between or among groups) increases with the number of dependent variables selected for univariate testing conducted within a given experiment. In the reviewer's comments below, the imaginary

submission included multiple dependent variables analyzed separately in univariate fashion. In the simulated reviewer comments below, the reviewer informs the author/s that employing multiple dependent variables and testing the null hypothesis with a series of univariate *t*-tests are very likely to inflate the experimentwise Type I error rate.

Reviewer Comments: Several statistical anomalies occur within the Results section of the current submission, and it appears that the authors have employed multiple concurrent *t*-tests regarding the same participants within the same experiment:

> We conducted a *t*-test to identify changes in levels of interest of participants who were exposed to rules indicating that some facial expressions were more desirable than they were during post-test 1 and post-test 2. Findings of the initial *t*-test established that the change in participants' levels of interest seen during post-test 1 was significant only for those who had difficulty maintaining eye contact during conversations (M = 42.376, SD = 24.912) and for participants demonstrating no difficulties associated with maintaining eye contact during conversations (M = 85.464, SD = 16.31), $t(2, 14) = 0.022$, $p < .05$. Interestingly, the change in participant interest during post-test 2 did not reach significance, $t(2, 14) = 0.4611$, $p > .05$.

As additional univariate statistical tests are employed in a given experimental preparation, the likelihood that one or more of the calculated probabilities being identified as significant increases dramatically as a function of chance. With each univariate test conducted, the differences between groups are more likely to become a function of sampling error. Since the authors have run pretests and post-tests on two groups (while measuring changes on several dependent variables simultaneously), it seems reasonable to suggest an analysis of covariance (ANCOVA) test. If the authors are determined to employ more than one dependent measure, a multivariate procedure that analyzes change from pretest to post-test will be most appropriate. Moreover, since the authors have employed several dependent variables concurrently, a multivariate analysis of covariance (MANCOVA) is the statistical technique of choice.

Multiple χ^2 Tests within a Single Experiment

In the section below, the authors have conducted a series of χ^2 tests with reference to the same participants within the same experiment. In this particular study, the obtained χ^2 values are unusually large, and several techniques to address the inflation of α are suggested. **Reviewer Comments:** On p. 101, within the Results section, the following statistical outcomes are provided:

> A chi-square test was conducted in order to identify potential changes in the target behavior directly related to the type of treatment protocol employed. This nonparametric statistical test demonstrated that the form of treatment significantly changed the participants' ability to perform accurate sound discriminations, χ^2 (1, n = 2,133) = 17.88, $p < .001$. An overview of the proportions of correct versus incorrect phoneme segmentation and correct versus incorrect word comprehension as a function of the protocol type is illustrated in Table 3. The chi-square analysis of the phoneme segmentation (correct versus incorrect) changed significantly in accordance with the duration of protocol implementation, χ^2 (2, n = 2,133) = 31.61, $p < .001$. Likewise, word comprehension (correct versus incorrect) changed significantly in accordance with the type of protocol employed during treatment, χ^2 (2, n = 2,133) = 32.5, $p < .001$.

From the above description of procedures and outcomes, it appears that separate χ^2 analyses were conducted with reference to the same participants within the same study. Such a strategy results in what often is described as probability pyramiding. Under such experimental arrangements, the authors inadvertently have increased the likelihood of obtaining significant results where they may not exist (see Stevens, 2009, for a discussion). On the other hand, the authors have obtained χ^2 values that are extremely large (albeit obtained concurrently), and it appears very likely that a Bonferroni adjustment might reveal statistical significance among "all" the measures. This popularly employed adjustment procedure may well add statistical precision to the authors' findings while maintaining more than adequate levels of significance among all the obtained findings.

From my point of view, the Bonferroni adjustment is the most straightforward correction strategy. This procedure obtains new required *P*-values by allowing the researcher to calculate adjusted probabilities and to keep the familywise α value at .05 (or another specified value). I will, however, mention that Bonferroni has the propensity to sacrifice statistical power because the familywise error calculation is based on the supposition that the null hypothesis is true for all comparisons made within a given experiment.

One could arrive at the required 0.05 alpha level by dividing 0.05 by 5 and obtaining 0.01 as the critical value representing a 0.05 for the experimentwise error rate. It is important for the author to understand that the "traditional Bonferroni adjustment" assumes all tests are orthogonal when calculating the familywise error rate. As something of an overgeneralization, this only requires dividing the authors' current required error rate (0.05) by the number of χ^2 tests employed:

Adjusted α-level = required α for your study / number of tests for significance.

In the current submission, the authors have employed three significance tests. Accordingly, they can employ the Bonferroni correction by dividing their Type I error rate by the number of dependent variables analyzed (0.05/3 = .01666 as the required level of significance for each of the three obtained χ^2 values). I will mention that numerous substitutes for the Bonferroni have been developed (see Olejnik, Li, Supattathum, & Huberty, 1997, for a review). Also, Šidák (1967) suggests several simple modifications of the Bonferroni formula that will not contribute to the likelihood of a Type II error. Many of these procedures are more powerful than the basic Bonferroni correction, and they have the advantage of being applicable to most of the commonly employed parametric and nonparametric procedures (e.g., *t*-tests, *F*-tests, or χ^2).

It might be useful for the authors to take into consideration the fact that "modified Bonferroni procedures" were developed for more diversified types of investigations than the χ^2 test. As mentioned previously, Olejnik et al. (1997) reviewed the modified Bonferroni procedures and their computations, and the modified Bonferroni techniques have clear benefits beyond the original version of the Bonferroni adjustment. Given the authors' large obtained χ^2 values, it appears that the basic Bonferroni adjustment will reveal statistically significant findings among "all" the outcomes in this study. Again, this frequently employed strategy is very likely to add statistical precision and sustain more than adequate levels of significance among all of the authors' current findings.

Multiple *F*-Tests

In the excerpt below, the authors' paper included findings that were obtained by way of conducting a series of multiple *F*-tests within the same experiment. In some parts of the simulated reviewer comments below, particular figures are mentioned; however, since this is a hypothetical review, no actual figures are included.

Reviewer Comments: Running a sequence of post hoc comparisons by way of the Tukey method to access various *t*-values [subsequent to identifying significant *P*-values for ANOVA] is essential. However, when a researcher conducts several univariate tests [in this case, *F*-tests] within "the same experiment," significant results can be found due to chance at a level that exceeds that specified by the experimenter. On p. 8 of the authors' current submission, the following material contains a series of univariate *F*-tests. These univariate tests were conducted in the absence of techniques that would protect the experimentwise error rate:

> A subsequent series of ANOVA tests were employed to analyze marital status as it relates to military experience and ability to sleep. A borderline significant main effect was obtained regarding marital status on the mean level for REM sleep, $F(2,46) = 2.490$, $p = .071$. A post hoc Tukey HSD test suggested that the mean for female veterans (M = -0.87, SD = 4.21) approached levels of statistical significance as it differed from the mean for male veterans (M = -3.76, SD = 2.92, $p < .061$) while the mean for male and female non-veterans (M = -3.89, SD = 2.31) was not significantly different from female veterans ($p = .747$) or male veterans ($p = .409$). Female veterans demonstrated higher levels of REM sleep than male veterans. This series of tests indicated no statistically reliable differences pertaining to marital status on the mean level of REM sleep, $F(2,46) = 0.211$, $p = .834$.

I understand all too well that employing a series of univariate *F*-tests "within the same experiment" has a certain natural logic and, perhaps, even an intuitive appeal. Plainly, this type of analysis has been published very often in the psychological and general behavioral science literature; nevertheless, these strategies represent a series of inaccurately calculated statistical findings [i.e., inflated levels of significance] when employed in this manner. Indeed, several investigations have brought this issue to the attention of the behavioral science community. As described by Neher (1967) and forwarded by Schatz, Jay, McComb, and McLaughlin (2005), the issue of "probability pyramiding" is not uncommon where multiple ANOVAs are employed within the same experimental preparation/study. Schatz et al. state:

> In most cases, researchers treat each statistical test individually, instead of examining the results as a whole, demonstrating a lack of control for Type I error. Performing multiple ANOVAs or following a MANOVA with univariate ANOVAs without adjusting the alpha level accordingly commonly results in Type I error (Dar, Serlin & Omer, 1994). The Bonferroni correction, commonly referred to in methodology and statistical texts and articles, is a simple-to-use control for inflated Type I error (e.g., Cohen, 1990; Stevens, 2009). This simple procedure involves decreasing your alpha level to account for the number of statistical analyses conducted on that independent variable, and is often referred to as a means of reducing "familywise error." For example, the researcher analyzing the effects of an intervention on five separate dependent measures (but using a sample size too small to meet the assumptions of a MANOVA) would "correct" the P-value to .01 to maintain an acceptable likelihood of Type I error. Whereas, five separate analyses performed with a .05 alpha level would result in a 25% likelihood of Type I

> error (5×.05 = .25), "correcting" the alpha level to .01 maintains a "familywise" error rate of 5% likelihood of Type I error (5×.01 = .05). (p. 1054)

In the study currently under review, the use of fragmented univariate *F*-tests act to inflate the overall Type I error rate, and the researchers may have concluded that treatments were effective in conditions where chance might well be operating. It is possible that a conventional multivariate analysis and *post hoc* tests (or Bonferroni adjustments) could demonstrate statistical significance among several of the dependent measures. Alternatively, multiple comparison techniques as described by Hochberg and Tamhane (1987) might be conducted. Any of these strategies would go a long way toward precluding the authors' current complications associated with the inflation of alpha.

Generally speaking, the most commonly employed technique that controls the Type I error rate is multivariate analysis of variance. As discussed by Stevens (2009), "…with the use of multiple criterion measures, we can obtain a more complete and detailed description of the phenomena under investigation…" (p. 2). And, as described by Leary & Altmaier (1980):

> The solution to the problem of inflated error rates with multivariate studies (those having more than one dependent variable) is to analyze the data using multivariate techniques; the most common examples are Hotelling's T^2 and multivariate analysis of variance (MANOVA), multivariate analogues of the *t* test, and ANOVA, respectively. (p. 613)

I should mention that there are other relatively straightforward solutions for remediating the current statistical complications. First, the authors might simply employ their existing descriptive statistics and graph their results in accordance with the current collection of electroencephalograph (EEG) measures in Figures 5 and 6. In doing so, the authors could eliminate references to the probability values obtained by way of multiple univariate tests within the first and second training procedures. Given that the current version of statistical findings is inflated, I believe straightforward "descriptive statistics" and graphical representations of findings are the most appropriate remedial strategy. From my perspective, visual inspection of the authors' excellent graphs in conjunction with their descriptive statistics is sufficiently compelling. In this particular study, more than in most, the graphs alone address the findings in a manner that is more persuasive than any of the presently listed *P*-values—particularly, when one considers that the *P*-values were obtained by way of conducting multiple univariate procedures within the same experiment.

Neural Networking Alternatives

Although neural networking systems as applied to the classification and/or prediction of behavioral outcomes is beyond the scope of this paper, it is important to know that there are a rapidly evolving series of neural network systems aimed at classifying/differentiating groups and predicting behavioral findings. This is because in many studies within the behavioral and related sciences, the objective may not revolve around identifying the probability of finding differences between groups; rather, the researcher/s may be interested in forecasting future outcomes. For example, can we predict which graduate students will make the best teaching assistants (e.g., Rumph, Ninness, & Lawson, under review)? Are we able to recognize the phonological features of speech (Kohonen, Makisara, & Saramaki, 1984)? Are we able to predict the time and location of particular financial crises (Erdal & Ekinci, 2013)? Can we accurately forecast students who

are at risk for specific types of academic problems (Ninness et al., 2005)? Can we correctly predict the voting behaviors of particular legislatures (Ninness et al., 2012)? Many researchers in the behavioral and physiological sciences might believe that, to a large extent, we are already capable of making amazingly accurate predictions or classifications regarding such measures by way of our conventional regression methodologies, and if the data at hand approximate linearity and conform to normal curve theory, they would be correct. However, our conventional statistical techniques have limitations when a dataset is extremely nonlinear and is not consistent with the key assumptions within normal curve theory.

As discussed by Ninness et al. (2013), in the behavioral and related sciences, where the number of multivariate, nonindependent, and nonlinear variables are continually rising, the sheer volume of new types of academic measurements is almost overwhelming (James, 1985). The exasperating, yet unavoidable, fact is that an increasing part of the data we collect in an effort to answer our complex academic questions have become a substantial part of those very questions (Gigerenzer, 2004). Artificial neural networks offer a range of modern alternatives to traditional univariate and multivariate statistical tests that may be limited by the key assumptions in normal curve theory. To the extent that the available data includes some form of functional relationship/s, behavioral scientists are likely to find a large number of neural networks that are capable of recognizing and predicting patterns even when the data of interest is extremely nonlinear and non-normal (Gonzales & DesJardins, 2002; Rumph, Ninness & Lawson, under review). On the other hand, when the available data "do" conform to the assumptions of normal curve theory, classical statistical techniques are very likely to perform as well (and much faster) than neural network systems aimed at performing the same types of operations (see Navarro & Bennun, 2014, for a related discussion).

Discussion

As described above, when we have data that conform to normal curve theory and the data at hand are consistent with the underlying assumptions pertaining to the analyses, conventional statistics provide a wealth of supplemental information that does not exist in most of the currently available neural network technology. For example, predictions generated by neural networks are not accompanied by known margins of error. That is, when a network algorithm forecasts a score or a series of scores, behaviors, or classifications, these predictions are not supplemented by values indicating a bandwidth of accuracy within which subsequent results will fall with a known margin of error. This does not mean that neural network predictions cannot use *cross-validation* procedures or perform follow-up operations that verify the accuracy of a given set of predictions. It does mean that there are a number of clear advantages to making decisions and predictions by way of traditional statistical methodology when it is possible to do so. For example, when employing classical statistics and rejecting the null hypothesis, the researcher can access confidence intervals within which true mean differences or effect sizes fall within a known margin of error. When making predictions regarding a person's performance on a particular test, the researcher has access to the intervals within which future performances are likely to fall. When making predictions with classical statistics, the researcher can see the extent to which each independent variable is weighted and contributes to the prediction of future outcomes. Such details are simply not available when conducting an analysis with most of the currently available neural network systems (Sharma, Rai, & Dev, 2012).

There are, however, several caveats with regard to employing classical statistics and obtaining all of the valuable heuristic details that accompany these traditional procedures, even when the available dataset represents an excellent approximation of a normal distribution. None of the standardized error values, confidence intervals, or *P*-values serves any real purpose if the procedures used to test the null hypotheses have inflated the Type I error. As described above, when one or more dependent variables within the same study are analyzed univariately the external validity is compromised and the Type I error rate becomes inflated. And, despite any impressive appearing *P*-values that may accompany such forecasts and despite any apparently impressive confidence intervals, and/or coefficients of determination, etc., none of the obtained statistical calculations are what they appear if the Type I error has been inflated as a result of probability pyramiding.

Effect Size

In a general sense, *effect size* can be described as the magnitude of change that occurs in one or more dependent variables that is a function of the independent variable/s. Even if a familywise Type I error is appropriately accounted for, there is still the potential issue of whether or not a statistically significant effect is of any practical significance in the context of the problem of interest. Consequently, any significant test result, regardless of how small its associated *P*-value might be, should be critically assessed for its practical relevance. Simple evaluation of the estimated magnitude of the effect itself may suffice, or evaluation of the estimated confidence limits for the effect may be more relevant in some contexts.

Clearly, the risk of finding statistically significant results that are so small in magnitude as to not be considered of any practical value is highest in the presence of very large samples. So while *P*-values help determine the presence (or not) of an effect of interest, they fail to address the hanging question of: If an effect is present, how large is it? This is the role of estimation and is why virtually every finding of a statistically significant effect should lead to a subsequent estimation of the size of the effect, and preferably, a confidence interval estimate of the effect reflecting the remaining level of uncertainty involved in the associated experiment. And, of course, consistent with the argument throughout this article, the size of these intervals will depend on how many such intervals are to be constructed in order to preserve the experiment-wide confidence level for all the intervals presented (see enrichment material that follows for some further details).

References

American Psychological Association. (1996). *Task force on statistical inference report*. Washington, DC: Author. doi: 10.1037/e404402005-016

Anderson, T.W. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). New York, NY: John Wiley.

Armstrong, S. A. & Henson, R. K. (2005). Statistical practices of *IPJT* researchers: A review from 1993-2003. *International Journal of Play Therapy, 14*(1), 7-26. doi: 10.1037/h0088888

Austin, P. C. & Brunner, L. J. (2004). Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. *Statistics in Medicine, 23*(7), 1159-1178. doi: 10.1002/sim.1687

Baldwin, S. A., Murray, D. M., & Shadish, W. R. (2005). Empirically supported treatments or type I errors? Problems with the analysis of data from group-administered treatments. *Journal of Consulting and Clinical Psychology, 73*(5), 924-935. doi: 10.1037/0022-006x.73.5.924

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society,* Series B (Methodological), *57*(1), 289–300. Retrieved from http://www.jstor.org/stable/2346101

Cafri, G., Kromrey, J. D., & Brannick, M. T. (2010). A meta-meta-analysis: Empirical review of statistical power, type I error rates, effect sizes, and model selection of meta-analysis published in psychology. *Multivariate Behavioral Research, 45*(2), 239-270. doi: 10.1080/00273171003680187

Chow, S-C., & Liu, J-P. (2000). *Design and analysis of bioavailability and bioequivalence studies*, (2nd ed.). New York, NY: Marcel-Dekker. doi: 10.1002/sim.834

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304–1312. doi: 10.1037//0003-066x.45.12.1304

Dar, R., Serlin, R. C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology, 62*(1), 75–82. doi: 10.1037/0022-006x.62.1.75

Edgington, E. S. (1995). *Randomization tests,* New York, NY: Marcel Deckker. doi: 10.1002/0471667196.ess2169.pub2

Erdal, H. I., & Ekinci, A. (2013). A comparison of various artificial intelligence methods in the prediction of bank failures. *Computation Economics, 42*(2), 199-215. doi: 10.1007/s10614-012-9332-0

Fish, L. (1988). Why multivariate methods are usually vital. *Measurement and Evaluation in Counseling and Development, 21,* 130-137.

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics, 33*(5), 587–606. doi:10.1016/j.socec.2004.09.033

Gonzáles, J. M. B., & DesJardins, S. L. (2002). Artificial neural networks: A new approach to predicting application behavior. *Research in Higher Education, 43*(2), 235-272. doi: 10.1023/a:1014423925000

Good, P. (1994). *Permutation tests: A practical guide to resampling methods for testing hypotheses.* New York, NY: Springer-Verlag.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York, NY: Wiley. doi: 10.1002/9780470316672

Huberty, C. J., & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analyses. *Psychological Bulletin, 105*(2), 302-308. doi: 10.1037/0033-2909.105.2.302

James, M. (1985). *Classification algorithms*. Hoboken, NJ: Wiley.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Upper Saddle River, NJ: Prentice-Hall.

Kleinbaum, D. G., & Kupper, L. L. (1978). *Applied regression analysis and other multivariable methods*. Pacific Grove, CA: Duxbury.

Kohonen, T., Makisara, K., & Saramaki, T. (1984). Phonotopic maps – insightful representation of phonological features for speech recognition. *Proceedings of the 6th International Conference on Pattern Recognition*, 182-185. Silver Spring, MD: IEEE Computer Society Press.

Leary, M. R. & Altmaier, E. M. (1980). Type I error in counseling research: A plea for multivariate analyses. *Journal of Counseling Psychology, 27*(6), 611-615. doi: 10.1037/0022-0167.27.6.611

Navarro, H., & Bennun, L. (2014). Descriptive examples of the limitations of artificial neural networks applied to the analysis of independent stochastic data. *International Journal of Computer Engineering & Technology (IJCET), 5*(10), 40-42.

Neher, A. (1967). Probability pyramiding, research error, and the need for independent replication. *Psychological Record, 17*(2), 257-262. Retrieved from http://search.proquest.com/openview/2061cb08642bff457e7bc43c35652362/1?pq-origsite=gscholar

Ninness, C., Rumph, M., Clary, L., Lawson, D., Lacy, J.T., Halle, S., McAdams, R., Parker, S. & Forney, D. (2013). Neural network and multivariate analysis: Pattern recognition in academic and social research. *Behavior and Social Issues, 22,* 49-63. doi: 10.5210/bsi.v22i0.4450

Ninness, C., Lauter, J., Coffee, M., Clary, L., Kelly, E., Rumph, M., Rumph, R., Kyle, B., & Ninness, S. (2012). Behavioral and biological neural network analyses: A common pathway toward pattern recognition and prediction. *The Psychological Record, 62(*4), 579-598.

Ninness, C., Rumph, R., McCuller, G., Harrison, C., Vasquez, E., Ford, A., Ninness, S., & Bradfield, A. (2005). A relational frame and artificial neural network approach to computer-interactive mathematics. *The Psychological Record, 55,* 561-570.

Ninness, C., Rumph, R., Vasquez, E., & Bradfield, A. (2002). Multivariate randomization tests for small-n behavioral research. *Behavior and Social Issues, 12*(1), 64-74. doi: 10.5210/bsi.v12i1.80

Olejnik, S., Li, J., Supattathum, S., & Huberty, C. J. (1997). Multiple testing and statistical power with modified Bonferroni procedures. *Journal of Educational and Behavioral Statistics, 22(4),* 389-406. doi: 10.3102/10769986022004389

Roscoe, J. T. (1975). *Fundamental research statistics for the behavioral sciences* (2nd ed.). New York, NY: Holt, Rinehart, & Winston. doi: 10.2307/2284880

Rumph, M. L., Ninness, C., & Lawson, D. (2015). Prediction Analytics Revisited: Statistical and Artificial Neural Network Approaches. Manuscript submitted for publication.

Schatz, P., Jay, K. A., McComb, J., & McLaughlin, J. R. (2005). Misuse of statistical tests in *Archives of Clinical Neuropsychology* publications. *Archives of Clinical Neuropsychology, 20*(8), 1053-1059. doi: 10.1016/j.acn.2005.06.006

Sharma, V., Rai, S., & Dev, A. (2012). A comprehensive study of artificial neural networks. *International Journal of Advanced Research in Computer Science and Software Engineering, 2(10),* 278-284.

Sidman, M. (1960). *Tactics of scientific research; evaluating experimental data in psychology.* New York, NY: Basic Books.

Šidák, Z. K. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association, 62*, 626–633. doi: 10.2307/2283989

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). Hillsdale, NJ: Erlbaum. doi.org/10.1111/j.1751-5823.2009.00095_13.x

Tatsuoka, M. M. (1973). Multivariate analysis in educational research. In F. N. Kerlinger (Ed.), *Review of research in education* (pp. 273-319). Itasca, IL: Peacock. doi: 10.3102/0091732x001001273

Tukey, J. (1949). Comparing individual means in the analysis of variance. *Biometrics, 5*(2), 99-114. doi: 10.2307/3001913

Wellek, S. (2002). *Testing statistical hypotheses of equivalence.* Boca Raton, FL: Chapman & Hall/CRC Press. doi: 10.1201/9781420035964

Appendix: Enrichment Material

Confidence Intervals

Due to the inherent uncertainty in drawing conclusions about populations using data acquired from samples of those populations, it often is desirable to attempt to quantify the uncertainty involved by providing confidence interval estimates of the variable parameters or treatment effects of interest. For commonly considered location parameters, the basic form of a confidence interval is outlined in Figure 1. The most common form of a 100(1-α) percent confidence interval for a population mean μ is that obtained from inverting a common ***t***-test, and is given as:

$$\overline{X} - t_{(n-1,1-\alpha/2)}S/\sqrt{n} \text{ to } \overline{X} + t_{(n-1,1-\alpha/2)}S/\sqrt{n} \text{ ,}$$

where $t_{(n-1,1-\alpha/2)}$ is the $100(1-\frac{\alpha}{2})^{th}$ percentile of a student's ***t***-distribution with n-1 degrees of freedom, and $\overline{X}$ and S are the sample mean and sample standard deviation, respectively. The form of this interval is entirely consistent with the diagram in Figure 1. It simply is comprised of the best point estimate of the parameter of interest plus and minus a suitable multiplier times an estimate of the standard deviation of that best point estimate.

Multivariate Confidence Intervals

In most experiments and/or research, there is more than a single parameter or treatment effect to be evaluated. Certainly, each parameter or effect of interest can be evaluated separately using univariate statistical procedures similar or related to those discussed above. However, the experimenter/researcher should be aware of the potential for propagation of error when adopting a univariate approach.

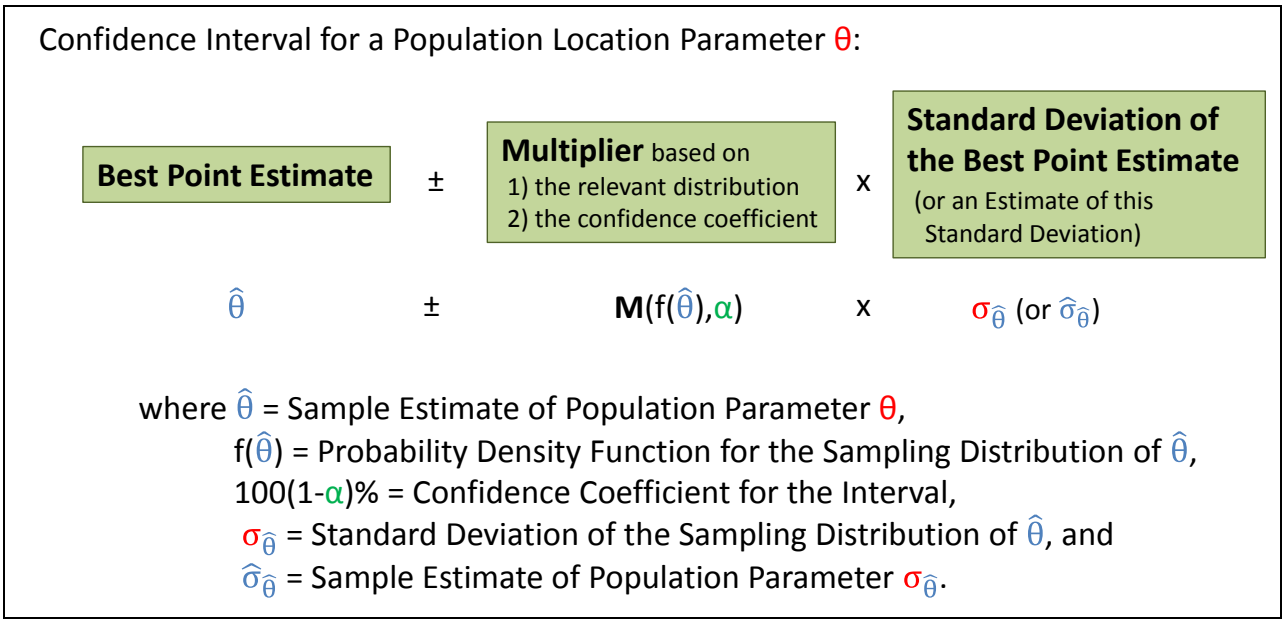

Figure 1. Basic structure of a confidence interval for a location parameter

As a basic example, consider a set of two test scores (perhaps, separate verbal and quantitative test scores) for a sample of n subjects representative of a population of interest (perhaps, for subjects from a geographic region or subjects receiving some kind of specialized advanced test preparation). Further, suppose the experimenter/researcher is interested in the mean test scores for the population from which this sample was obtained (i.e., μ_V & μ_Q). Separate univariate confidence intervals can be obtained for both of these population parameters using the approach described above. With an $\alpha = 0.10$, $n = 25$, $\overline{X}_V = 420$, $S_V = 90$, $\overline{X}_Q = 380$, and $S_Q = 125$, the resultant 90% confidence intervals are given as seen in Figure 2.

While each of the intervals above has only a 10% chance of not including its respective population parameters, it is important to understand that the chance that both intervals are in error is at least 10%, and that this chance could be as large as twice the potential rate for each individual interval, which in this case would be 20%.

Bonferroni Theorem

Exactly how large the experiment-wide error rate is will not be known, but it can be bounded. The Bonferroni Theorem ensures that for multiple intervals all having the same individual error rate α, the experiment-wide error rate will be no larger than the minimum of one and the sum of the individual error rates (i.e., mα, where m is the number of intervals being produced). Consequently, if it were desired that the overall, experiment-wide error rate for the intervals in this example was to be no larger than 10%, then two individual 95% confidence intervals could be constructed. Such intervals are relatively easy to construct as they only require a change to the multiplier used in their calculation.

Interval for μ_V:	Interval for μ_Q:
$\overline{X}_V \pm t_{(n-1,1-\alpha/2)} S_V/\sqrt{n}$	$\overline{X}_Q \pm t_{(n-1,1-\alpha/2)} S_Q/\sqrt{n}$
$420 \pm t_{(24,0.95)} 90/\sqrt{25}$	$380 \pm t_{(24,0.95)} 125/\sqrt{25}$
$420 \pm 1.711(18) \approx$	$380 \pm 1.711(25) \approx$
420 ± 30.8	380 ± 42.8
389.2 to 450.8	337.2 to 422.8

Figure 2. Univariate student's t percent confidence intervals

For the example considered here, the original multiplier used was $t_{(n-1,1-\alpha/2)} = t_{(24,0.95)} = 1.711$; however, this could produce an experiment-wide error rate as large as 20%. To preserve an experiment-wide error rate of no more than 10%, a multiplier of $t_{(n-1,1-\alpha/[2m])} = t_{(24,0.975)} = 2.064$ (note here, m = 2 intervals are being considered) could have been used for each individual interval. While these intervals will be over 20% (i.e., $\frac{2.064}{1.711} - 1 \approx 20.63\%$) wider than the original intervals, they ensure that the experiment-wide error rate is no more than 10%, while for the original intervals the experiment-wide error rate was almost certainly larger than 10%.

Just as it is necessary to pay a penalty for not knowing the true population standard deviation in construction of a univariate confidence interval (i.e., having to use a ***t***-distribution percentile as the multiplier that will always be larger than its corresponding normal distribution percentile), it also is necessary to pay a penalty to manage the propagation of error that occurs when constructing multiple intervals (or making multiple inferences). In both cases, the "penalty" comes in the form of wider confidence intervals that are produced through use of larger and more appropriate multipliers of the standard error, or more accurately, an estimate of the standard error.

The Bonferroni Theorem is based on a probability argument and simply provides an upper bound for the overall error rate. The actual error rate could be, and most likely is, less than the Bonferroni bound. While the Bonferroni approach recognizes that the considered variables might be correlated and impact the respective experiment-wide error rate, it makes no attempt to directly account for (or estimate) the relevant correlation(s).

A well-known multivariate statistic, Hotelling's T^2, does make an attempt to directly account for any correlation that might be present among the variables being considered by an experimenter/researcher. Hotelling's T^2 is generally defined as:

$$T^2 = n(\bar{\underline{x}} - \underline{\mu})^T \mathbf{S}^{-1} (\bar{\underline{x}} - \underline{\mu}),$$

where $\bar{\underline{x}} = \frac{1}{n}\sum_{i=1}^{n} \underline{x}_i$, with $\underline{x}_i$ an mx1 vector of results for the i^{th} sample unit/subject,

$\underline{\mu}$ = an mx1 vector of the m variable population mean values,

$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\underline{x}_i - \bar{\underline{x}})(\underline{x}_i - \bar{\underline{x}})^T$ an mxm sample variance-covariance matrix,

$\underline{v}^T$ = a 1xm transpose of the mx1vector $\underline{v}$, and

$\mathbf{A}^{-1}$ = an mxm inverse of the mxm matrix $\mathbf{A}$.

When the random vectors $\underline{x}_i$ originate from an m-dimensional multivariate normal distribution, $\frac{n-m}{m(n-1)}T^2$ follows an F-distribution with m and n-m degrees of freedom, allowing for construction of m-dimensional confidence ellipsoids. Such an m-dimensional ellipsoid is actually an m-dimensional region having the specified confidence of including the m-dimensional population mean vector $\underline{\mu}$. This is a region that has the specified confidence of containing all the μ_j values for the j = 1 to m parameters/effects of interest. When m = 1, this confidence region reduces to the confidence interval previously discussed.

For the example considered here, assume the sample correlation between an individual's Verbal and Quantitative test scores is 0.8, then the relevant covariance is 0.8 times the product of the two variable standard deviations (i.e., $S_v * S_Q = 90*125 = 11250$), or $S_{VQ} = 9000$. From the T^2 statistic noted above, the 2-dimensional 90% confidence ellipsoid for the population parameters vector $\underline{\mu}^T = [\mu_V, \mu_Q]$ is as appears in Figure 3.

While this confidence ellipsoid can be displayed when there are only two parameters/effects being evaluated, and perhaps, a three-dimensional ellipsoid could be displayed, if more than three parameters/effects are being considered, then display of the resultant ellipsoid is not possible. In addition, it is often desirable to be able to present confidence intervals that are not dependent on the unknown values of the other parameters/effects. These can be obtained from the approach described here and are of the same basic form as all the intervals described above and as displayed in Figure 1.

The best point estimates are still the relevant sample averages, and the estimates of their standard deviations are the same. The only difference is again found in the multiplier used to obtain the intervals for the individual parameters. While the univariate and Bonferroni interval multipliers were appropriate percentiles of student's ***t*-**distribution with n-1 degrees of freedom, the multipliers based on the T^2 statistic are given as:

$$M_{T^2} = \sqrt{\frac{m(n-1)}{n-m} F_{(m,n-m,1-\alpha)}}$$

where m = number of parameters/effects being considered,

n = number of multivariate observations (i.e., the sample size), and

$F_{(df_1,df_2,p)}$ = p^{th} percentile of an *F*-distribution with df_1 & df_2 degrees of freedom.

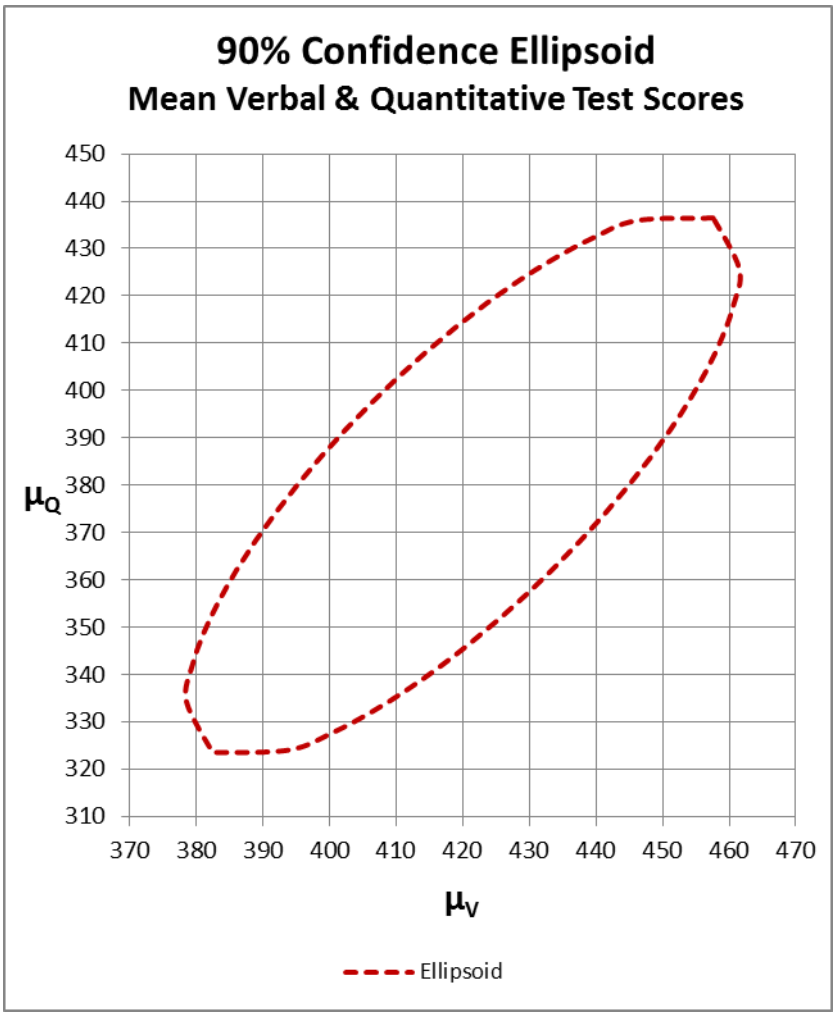

Figure 3. Ninety percent confidence ellipsoid for mean verbal and quantitative test scores

For the example considered here, $m = 2$, $n = 25$, $1-\alpha = 0.90$, and $F_{(2,23,0.9)} \approx 2.5493$, giving $M_{T^2} \approx 2.307$. Note that this multiplier produces an interval wider than both the Bonferroni and the univariate intervals, where the multipliers are 2.064 and 1.711, respectively. As with the Bonferroni intervals, these T^2 intervals also provide an error rate of no more than α that all the intervals include their respective parameters/effects. However, since they are wider than the Bonferroni intervals, the latter are often preferred by most analysts.

All the intervals appear in Figure 4, where the relationship between them can be readily observed. It should be noted that the T^2 intervals are the same width as projections, or shadows of the ellipsoid on the respective axes. In comparison to the exact ellipsoidal interval, the T^2 (multivariate) intervals are clearly conservative. The Bonferroni intervals are not as obviously conservative from the display but are known to be by probability theory. Probability theory also ensures the univariate intervals are liberal (i.e., have less than the stated 90% coverage for both parameters/effects, simultaneously).

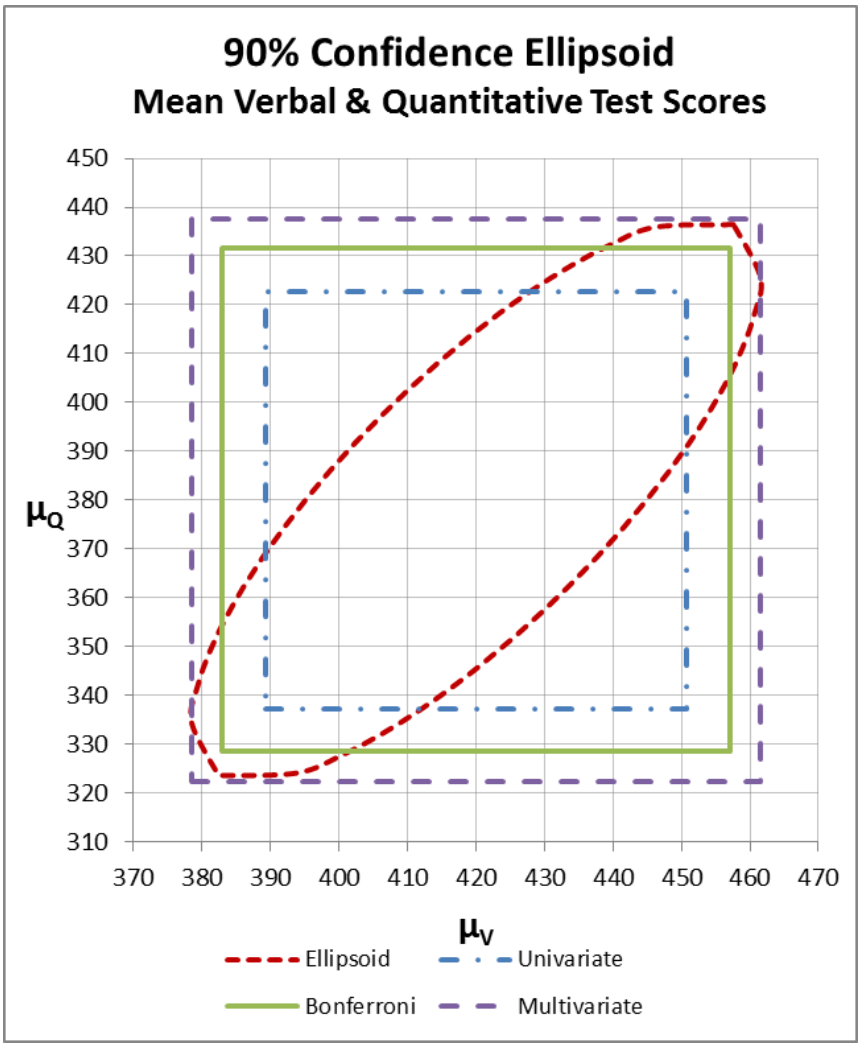

Figure 4. Comparison of univariate, Bonferroni, multivariate, and ellipsoidal intervals

A reasonable question at this point might be why the T^2 statistic is of any value at all if the confidence intervals it provides are always wider than the corresponding Bonferroni intervals. The value of the T^2 statistic is in conducting inference through hypothesis testing.

Note that if the experimenter/researcher involved in the example being considered here were interested in testing whether or not the mean test scores across the test types were equal or not (i.e., H_0: $\mu_V = \mu_Q$), then being willing to consider a Type I error rate of $\alpha = 0.10$, the researcher would reject this hypothesis using the T^2 statistic. However, considering only univariate approaches at the same error rate would fail to reject the hypothesis of equal mean test score values for the population of interest.

This can be observed in Figure 5, where the line $\mu_V = \mu_Q$ has been superimposed on Figure 4. The line lies entirely above the confidence ellipsoid; however, it passes through all the rectangles.

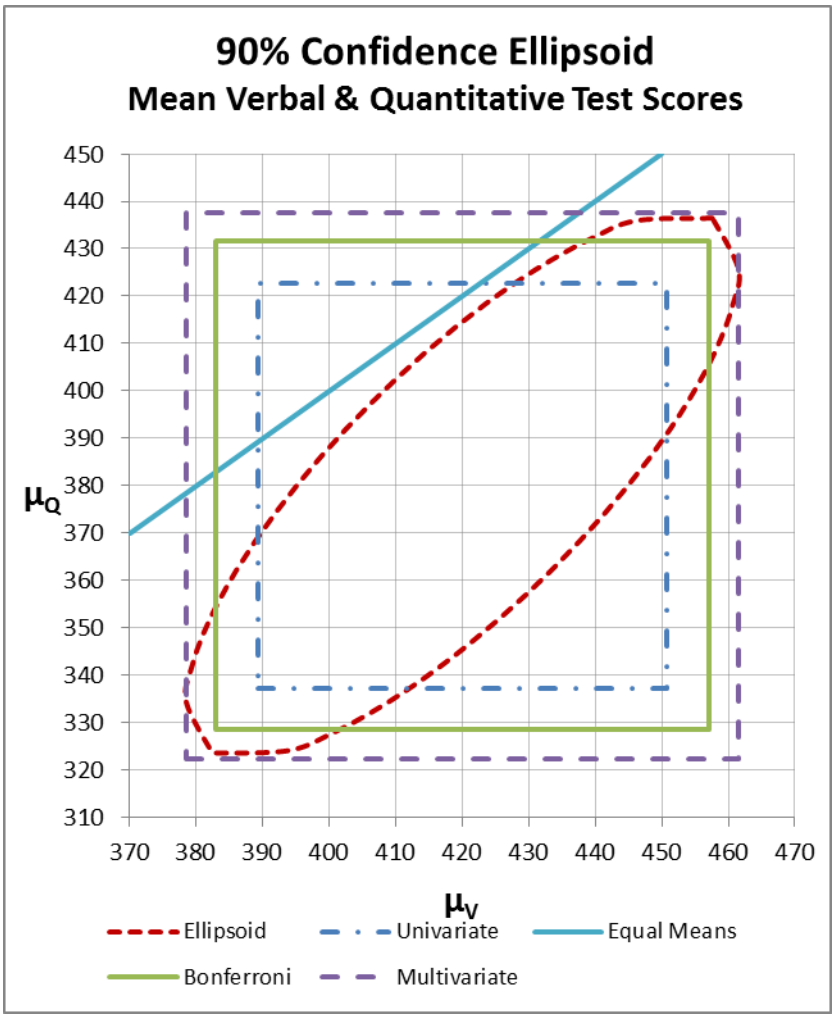

Figure 5. Evaluation of hypothesis of equal test score population means

This type of situation, where the multivariate test statistic indicates that the population parameters/effects are not equal, but all associated confidence intervals suggest that they may indeed be equal, can be encountered whenever more than a single parameter/effect is being evaluated. It is more likely to occur when the variables involved are highly correlated than when they are not. However, when variables are highly correlated, they essentially carry much of the same information, and perhaps, the measurement of both is unnecessary.

Reducing the dimensionality (e.g., reducing the number of variables under consideration or combining them in some manner) of an experiment is almost always desirable as it avoids some of the issues, complications, and complexities of effectively analyzing multivariate data. Certainly, removing essentially redundant measurements from consideration is one means to reduce dimensionality. This can be an attractive alternative in situations where one of the correlated variables is very costly to obtain compared to the other. However, in situations where removal is not a viable option, there are other approaches to reducing the dimensionality of the data to a more manageable and inherently understandable low (i.e., 2 to 3) dimensional space.

Discussion of such alternatives is really beyond the scope of this dialogue (see Johnson & Wichern, 2007; Anderson, 2003).

Conclusion

When an experimenter/researcher is considering more than one parameter/effect, using purely univariate approaches to separately evaluate each one is rarely the appropriate analysis approach. It frequently leads to misstatements relative to overall, experiment-wide error rates, as error is propagated across perhaps many univariate tests/intervals.

When considering confidence intervals for parameters/effects, simple adjustments to the commonly applied univariate intervals can be made to ensure low (i.e., stated) experiment-wide error rates. These adjustments take the form of slightly larger multipliers being used in the common construction of such intervals. However, all such intervals are still conservative and may not entirely align with their associated hypothesis test results. In such cases, the exact multivariate confidence ellipsoids become the only appropriate intervals. Producing displays of these is greatly facilitated if the dimensionality of the problem can be reduced to a more readily manageable level.