Title: Outcome Based Effect Size for the Marginals of a Multivariate Normal Distribution

Yan Ling, 5786 Field Brook Dr., East Amherst, NY, 14051, yli3333@gmail.com Corresponding Author

Paul I. Nelson, 101 Dickens Hall, Manhattan, KS, 66506, nels@ksu.edu

Running Title: Effect Size Multivariate Normal

Key Words: irreproducible results, randomized block designs, repeated measure designs, Hotellings T^2, bioequivalence, heteroscedasticity, split plot designs.

AMS Classifications: 62H99, 62F99

Abstract

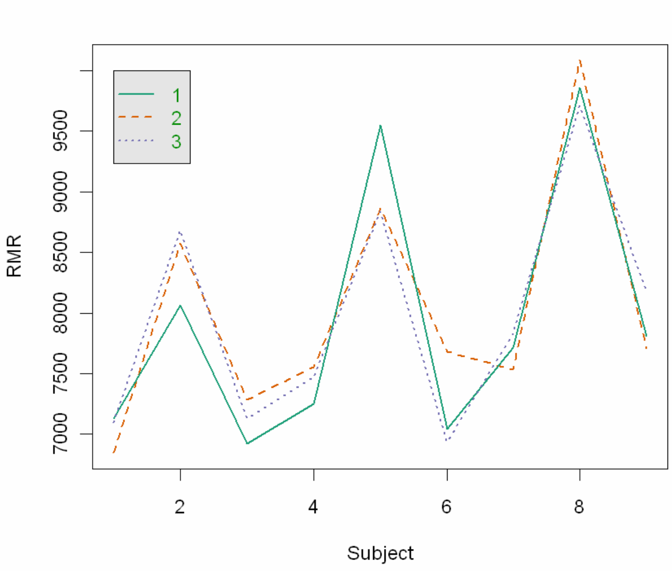
We develop and explore an effect size parameter based on a maximal contrast in observable outcomes that can be used to assess the degree of separation among two or more treatments modeled as the marginals of a multivariate normal distribution obtained from experiments replicated in blocks. Our effect size has observable consequences that can help experimenters calibrate its magnitude. Simulations indicate that since confidence intervals for our effect size can be quite wide, containing both small and large values when there are only a few blocks, a common occurrence, researchers should more often than is current practice, reserve judgment rather than conclude either that some new treatments are in a practical sense *significantly* better than existing ones or not *significantly* different in settings such as bioequivalence studies. We believe that such caution may be an appropriate response to the growing concern about experimental results which cannot be verified by replication.

1. Introduction

An effect size is a location-scale invariant index which can be used to quantify the separation among distributions of responses to treatments. The following example illustrates the application and utility of the new effect size developed here for correlated data.

Example 1.1. [2] reported on a study to find out if three protocols (methods) for measuring resting metabolic rate (*rmr*) in adult males were essentially equivalent. All three methods were used on each of a sample of n = 9 subjects in a randomized block design. A profile plot of protocol rmr's across subjects, presented in Figure 1.1, indicates relatively large differences among subjects and small differences among the protocols. Testing for equal protocol means results in p-values of 0.795 and 0.898 for the standard F-test and Hotelling's T^2. The authors concluded that their analysis implied that the protocols could be used interchangeably. However, the 0.95 two sided confidence interval for our effect size, described below, includes values that allow the possibility of systematic differences among the protocols. This added information that, although protocol means are close on the *rmr* scale and not *statistically significantly different*, some differences among the protocols could be of *clinical significance*.

Figure 1.1 Profiles of RMR Responses Across Subjects



Effect sizes for *k* treatments based on correlated data have largely been limited to paired comparison designs, *k* = 2, implemented by applying one sample methods to the differences between responses. [7] and [1] proposed and studied effect size statistics for a broad class of designs, including those with k ≥ 2 repeated measurements. Their effect sizes are ratios of aggregated sums of squares obtained from univariate ANOVA's. Although intuitively appealing as proportions of explained variance observed in a particular experiment, these ad hoc statistics applied to correlated data do not appear to correspond to intrinsic properties of the treatments, such as meaningful parameters, and provide no basis for estimating their values under repeated runs of the experiment. Specifically, even if the restrictive Huyn-Feldt conditions on covariances, given in [4], were to hold, inference for these effect sizes would be difficult to carry out and interpret. Here, we extend the effect size, denoted π_{MAX}, developed by [6] based on a maximal contrast in observable outcomes from independent random samples to $k \geq 2$ multivariate normal responses in designs carried out as independently replicated blocks. In the context of the correlated responses in model (1.1), we show that π_{MAX} has observable consequences that can help distinguish between statistical and practical significance, an important issue since there are almost always some treatment effects. Unlike the independent random samples setting of [6] where only approximate inference is possible, exact inference for π_{MAX} here can be obtained from confidence intervals and tests for the non-centrality parameter of a one sample Hotelling's T^2 statistic. Our simulations indicate that since confidence intervals for π_{MAX} can be quite wide, containing both small and large values when there are only a few blocks, researchers should, more often than is current practice, reserve judgment rather than conclude either that some new treatments are in a practical sense *significantly* better than existing ones or not *significantly* different in settings such as bioequivalence studies. We believe that such caution may be an appropriate response to the growing concern about experimental results which cannot be verified by replication, as described in the journal Nature's archive 'Challenges in Irreproducible Research.' At the end of Section 3 we indicate how our global, non-directional effect size π_{MAX} can be adapted to settings where a targeted, directional inference is of interest, as is the case in non-inferiority trials. We do not require the common assumptions of equal treatment variances or special covariance structures. Our approach can be applied to the following data structure.

2. Data Structure

Let $\{\mathbf{Y}_i; i = 1,2,...,n\}$ be iid (independent, identically distributed) copies of an N x 1 multivariate normally distributed vector $\mathbf{Y}$,which represent the responses to the same experiment independently carried out in *n* replicates, called blocks. Assume the mixed model

$$\{\mathbf{Y}_i = \mathbf{W\Gamma} + \mathbf{Zu}_i + g_i\mathbf{1} + \mathbf{v}_i, i = 1,2,...,n\} \tag{2.1}$$

where $\mathbf{W}$ and $\mathbf{Z}$ are known matrices of constants, $\mathbf{\Gamma}$ is an unknown vector of regression parameters, $\{\mathbf{u}_i\}$ are vectors of random effects, $\{g_i\}$ are iid block effects and $\{\mathbf{v}_i\}$, $\{g_i\}$ and $\{\mathbf{u}_i\}$ are independent, normally distributed mean zero vectors. Suppose that it desired to estimate an effect size for k- treatments whose vector of mean responses is the k -dimensional vector of linear combinations given by $\boldsymbol{\mu} = E(\mathbf{HY}) = \mathbf{HW\Gamma}$, where $\mathbf{H}$ is an $k\ x\ N$ matrix of constants of rank $k < n$ and $\mathbf{H1} = \mathbf{1}$, i.e. row sums are 1 and. Note that $\boldsymbol{\mu}$ is estimable and letting $\mathbf{X} = \mathbf{HY}$,

$$\mathbf{X}_i = \boldsymbol{\mu} + g_i\mathbf{1} + \boldsymbol{\varepsilon}_i \ , i = 1,2,\dots,\mathrm{n}, \tag{2.2}$$

$$\sim \mathrm{MVN}(\boldsymbol{\mu}, \mathbf{\Sigma} = \sigma_g^2\mathbf{11}' + \mathbf{\Sigma}_\varepsilon)$$

are independent vectors of k correlated measurements recorded on each block with error terms $\{\boldsymbol{\varepsilon}_i = \mathbf{H}(\mathbf{Zu}_i + \mathbf{v}_i) \sim \mathrm{MVN}(\mathbf{0}, \mathbf{\Sigma}_\varepsilon)\}$. Then, $\mathbf{X}^\# \equiv \{\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{ik})' \ , i = 1,2,\dots,n\}$ represents the data available to estimate the desired effect size. This framework encompasses single effects, main effects and interactions in such useful designs as randomized block, repeated measures and split plot designs which are completely replicated in randomly selected blocks. For the special case of a randomized complete block design, $\mathbf{\Sigma} = \sigma_g^2\mathbf{J} + \mathbf{D}(\boldsymbol{\sigma}^2)$, where $\mathbf{J}^{kxk} = (1)$, $\mathbf{1}$ is a $k \times k$ vector of ones and $\mathbf{D}(\boldsymbol{\sigma}^2)$ is a diagonal matrix with diagonal entries $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2)$ so that covariance $(X_{ij}, X_{im}) = \sigma_g^2$ for $m \neq j$ and $\sigma_g^2 + \sigma_j^2$ for $m = j$.

Our effect size, denoted by π_{MAX} , which depends on both the vector of means and the matrix of covariances in (2.2) for treatment combinations such as main effects that are obtained by averaging over other sources of variation in the particular experiment at hand, as is illustrated in the split plot example given below, is hence experiment specific and not an intrinsic property of the treatment distributions which might be investigated in different settings. This fact raises questions about the utility of effect sizes estimated from a meta analysis conducted across possibly very different experiments and environments.

Split Plot Designs in Field Trials: Suppose an experiment is conducted to compare crop yields that would be obtained by using *a* fixed levels of factor A and *b* fixed levels of factor B . Farms $\{G_i, i = 1,2,...,n\}$ are selected at random and act as blocks. Independently, each farm is divided into *a* whole plots and each of these is further divided into *b* subplots. Independently within each farm, at random, one level of factor A is applied to each whole plot and each level of factor *B* to the subplots. Letting y_{ijk} denote the response in farm (block) *i*, level *j* of factor *A* and level *k* of factor *B*, the standard model for this setup is given by:

$$y_{ijk} = \mu + g_i + \alpha_j + \delta_{j(i)} + \beta_k + (\alpha\beta)_{jk} + v_{k(ij)} \quad , \tag{2.3}$$

where the block effects $\{g_i; i = 1,2,...,n\}$ are iid $N(0,\sigma_g^2)$, the whole plot error terms $\{\delta_{i(j)}\}$ are iid $N(0,\sigma_\delta^2)$ and the split plot errors $\{v_{k(ij)}\}$ are iid $N(0,\sigma_v^2)$. All three random vectors are taken to be jointly independent. Further, take the fixed effects to sum to zero: $\sum \alpha_j = \sum \beta_k = \sum_j (\alpha\beta)_{ij} = \sum_i (\alpha\beta)_{ij}$ = 0. Note that $E(y_{ijk}) = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} = \mu_{jk}$. See Example 3.3 for an illustration of this model given by [11]. Effect sizes for *A*, *B* and the interaction between *A* and *B* interaction may be obtained from (2.3) as follows:

A Effect: Set $x_{ij} = \sum_{k=1}^{b} y_{ijk} / b = \mu + \alpha_j + g_i + \delta_{j(i)} + \bar{v}_{\square(ij)} = \mu + \alpha_i + g_i + \delta_{j(i)} + \varepsilon_{ij}$.

B Effect: Set $x_{ik} = \sum_{j=1}^{a} y_{ijk} / a = \mu + \beta_k + g_i + \bar{\delta}_{\square(i)} + \bar{v}_{k(i\square)} = \mu + \beta_k + g_i + \varepsilon_{ik}$.

A x B Effect: The standard null hypothesis of testing for an *A* × *B* interaction is that all *a* × *b* $\{(\alpha\beta)_{jk}\}$ parameters are *zero*. However, setting $x_{ijk} = (\hat{\alpha\beta})_{ijk} = y_{ijk} - \bar{y}_{i\cdot k} - \bar{y}_{ij\cdot} + \bar{y}_{i\cdot\cdot}$, $1 \le j \le a-1$, $1 \le k \le b-1$, $1 \le i \le n$, the linearly independent residuals constructed from subtracting estimates of *A* and B main effects from the observations in each cell in each block, although possible, would result in an effect size that assesses the extent to which the distributions of these *(a-1)(b-1)* observations differ, something difficult to interpret and of doubtful practical utility. Instead, we suggest basing interaction effect size on the usually interesting differences among the levels of A across the levels of B and conversely, as follows.

(I) Compare A_l to A_m at B_k, k = 1,2,...,b, 1 ≤ l < m ≤ a,

Set $x_{ilmk} = y_{ilk} - y_{imk}$.

(ii) Compare B_l to B_m at A_j, j = 1,2,...,a, 1≤l < m ≤ b,

Set $x_{ijlm} = y_{ilk} - y_{imk}$.

3. Effect Size π_{MAX}

Without loss of generality, assume that the greater the magnitude of a response, the more favorable the outcome. For the model in (2.2), for any k x 1 contrast vector of constants $\mathbf{l} = (l_1, l_2, ..., l_K)'$, $\mathbf{l'1} = \sum l_i = 0,$ we define a location-scale invariant contrast superiority ordering by the event $\{ \mathbf{l'X} > 0 \}$ and set

$$\pi(\mathbf{l}) \equiv P(\mathbf{l'X} > 0) = P(\mathbf{l'}(\boldsymbol{\mu} + b\mathbf{1} + \boldsymbol{\varepsilon}) > 0)$$
$$= \Phi(\mathbf{l'}\boldsymbol{\mu} / \sqrt{\mathbf{l'\Sigma l}}) = \Phi(\mathbf{l'}\boldsymbol{\mu} / \sqrt{\mathbf{l'}\boldsymbol{\Sigma}_\varepsilon \mathbf{l}}),$$

where Φ denotes the distribution function of a standard normal. Following [6], we define a *global*, non-directional effect size π_{MAX} for this problem by

$$\pi_{MAX} \equiv \sup\{\pi(\mathbf{l}); \mathbf{l} \in \mathbf{L}^{ALL}\} = \sup\{Max\{\pi(\mathbf{l}), 1 - \pi(\mathbf{l})\}; \mathbf{l} \in \mathbf{L}^{ALL}\}$$
$$= \Phi([(\boldsymbol{\mu} - \bar{\mu}_\varepsilon \mathbf{1})' \boldsymbol{\Sigma}_\varepsilon^{-1} (\boldsymbol{\mu} - \bar{\mu}_\varepsilon \mathbf{1})]^{.5}) \qquad (3.1)$$
$$\equiv \Phi(\gamma)$$

$$= \pi(c\mathbf{l}_{MAX})$$

$$= \pi[c\boldsymbol{\Sigma}_{\varepsilon}(\boldsymbol{\mu} - \bar{\mu}_{\varepsilon}\mathbf{1})],$$

where, $\mathbf{L}^{ALL} = \{\mathbf{l}\}$ denotes the collection of all nonzero K-dimensional contrast vectors of constants, $\mathbf{1}$ denotes a vector of ones, $\bar{\mu}_{\varepsilon} = \mathbf{1}'\boldsymbol{\Sigma}_{\varepsilon}^{-1}\boldsymbol{\mu} / \mathbf{1}'\boldsymbol{\Sigma}_{\varepsilon}^{-1}\mathbf{1}$,

$$\gamma^2 = (\boldsymbol{\mu} - \bar{\mu}_{\varepsilon}\mathbf{1})'\boldsymbol{\Sigma}_{\varepsilon}^{-1}(\boldsymbol{\mu} - \bar{\mu}_{\varepsilon}\mathbf{1}) \qquad (3.2)$$

and c is an arbitrary, non-zero constant. A proof of (3.1) is given in the appendix. Note that π_{MAX} = $\Phi(\gamma)$. Inference for π_{MAX} is based on $n\gamma^2$ being the non-centrality parameter of Hotelling's T^2 when used to test $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_k)'$. The following properties hold: (i) as should be the case, γ , and hence π_{MAX} , does not depend on the variance component due to blocks, which plays no role in calibrating differences among the treatment/conditions; (ii) π_{MAX} is a monotone increasing function of the power function of Hotelling's T^2 one sample test; (iii) if $\boldsymbol{\Sigma}_{\varepsilon}$ = $\sigma^2\mathbf{I}$, π_{MAX} is a monotone function of $(\boldsymbol{\mu} - \bar{\mu}\mathbf{1})'(\boldsymbol{\mu} - \bar{\mu}\mathbf{1}) / \sigma^2$, a commonly used effect size for a paired comparison design (RCBD with k = 2); (iv) π_{MAX} = 0.5, its minimum value, only if all the response means are equal, but π_{MAX} = 0.5 does not imply that the response distributions are identical; (v) π_{MAX} is invariant with respect to location-scale changes in $\mathbf{X}$; (vi) for k = 2, π_{MAX} is a two sided Mann-Whitney ordering, Max{ $P(X_1 > X_2)$, $P(X_1 < X_2)$ } =$\Phi[| \mu_1 - \mu_2 | / \sqrt{(1'\Sigma_{\varepsilon}1)}]$. See Chapter 5 of [3] for a discussion of effect size for comparing k = 2 distributions based on this ordering. Although, as with all effect sizes, calibration as to what is big and what is small are subjective judgments, the magnitude of π_{MAX} has observable consequences that can aid researchers in making meaningful interpretations of its values, as follows.

A contrast in means, $\mathbf{l}'\boldsymbol{\mu}$, or outcomes, $\mathbf{l}'\mathbf{X}$, makes a comparison between two combinations of treatments/conditions, those with positive coefficients vs. those with negative coefficients. Unlike π_{MAX} , neither of these types of comparisons is scale invariant, a particularly important issue in our setting where treatments/conditions may differ in spread as well as location. First, in comparing treatments, a small value of π_{MAX} indicates that in many independent applications to blocks, no combination of treatments/conditions would frequently be observed to be better than any other combination, pointing towards what in the context of comparing drugs is called *bioequivqlence*, even though un-standardized differences among some of the means may be large. On the other hand, a large value of π_{MAX} indicates that some specific treatment/condition combinations that *look good* in one application to a block would actually be good across many repeated applications. Since π_{MAX} depends on covariances as well as means, in complex designs such as (2.1) its value for a treatment whose levels have means $\boldsymbol{\mu} = E(\mathbf{HY}) = E(\mathbf{X}) = \mathbf{HW\Gamma}$ potentially depends on all sources on variation within a block and is hence, as noted above, experiment specific. Simply put, π_{MAX} calibrates the likelihood of separation orderings between groups of treatments that would actually be observed upon many independent replications of a *particular* experiment. If for example, a field trial is conducted under artificially homogeneous conditions, a treatment which appears to have a large effect in the laboratory could have small effect in real world settings. Finally, π_{MAX} is a statement about the ordering of the distributions and, being scale free, contains no information as to how far apart the means are in a particular unit of measurement. However, in situations where interest lies in the separation among the means in an intrinsically meaningful unit of measurement recorded under conditions where the distributions have an approximately equal standard deviation whose magnitude is close to what it would be in practice, π_{MAX} could be used to augment a traditional analysis based on inference for means. See the examples below.

4. Inference for π_{MAX}

For a user input proportion $\pi_0 \in [.5,1)$, we first propose, when appropriate, testing two sets of hypotheses using the data $\{\mathbf{X}_i = \mathbf{x}_i; i = 1,2,...,n\}$,

$$H_0 : \pi_{MAX} \le \pi_0 \quad \text{vs.} \quad H_1 : \pi_{MAX} > \pi_0 \ , \tag{4.1}$$

$$H_0 : \pi_{MAX} \ge \pi_0 \quad \text{vs} \quad H_1 : \pi_{MAX} < \pi_0 \ . \tag{4.2}$$

Rejection of H_0 could be used to support bioequivalence in (4.2) and the statement that there is*,* a difference among the treatments in (4.1) that is of practical importance. For π_0 = 0.5, H_0 in (4.1) is the traditional null hypothesis of equal means. A decision to reject this null hypothesis is often, mistakenly in our view, taken as support of the conclusion that there are practically significant and not just statistically significant differences among the treatments. As shown below, confidence intervals for π_{MAX} can be constructed by inverting these tests.

Let $\bar{\mathbf{X}} = (\bar{X}_1, \bar{X}_2, ..., \bar{X}_k)'$ denote the vector of sample means, $\mathbf{S} = \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})'(\mathbf{X}_i - \bar{\mathbf{X}}) / (n-1)$ the sample covariance matrix, $n > k$, and for any $k-1$ x k matrix of constants $\mathbf{C}$ whose rows are linearly independent vectors of contrasts . Then, from (2.2) we have that

$$T^2 = n(\mathbf{C}\mathbf{X})'(\mathbf{C}\mathbf{S}\mathbf{C}')^{-1}\mathbf{C}\mathbf{X} \sim T^2_{k-1,n-1,n\gamma^2} \ , \tag{4.3}$$

where $T^2_{r,\nu,\omega}$ denotes an *r*-variate non-central Hotelling's T^2 with ν degrees of freedom and non-centrality parameter ω . The fact that T^2 is invariant with respect to the choice of $\mathbf{C}$ and has non-centrality parameter $\omega = n\gamma^2$ follows from an identity given in [10]. Since $F = (n-k+1)T^2 / ((n-1)(k-1)) \sim F_{k-1,n-k+1,n\gamma^2}$, an *F*-distribution with *k*-1 and *n*-*k*+1 degrees of freedom and non-centrality parameter $n\gamma^2$, inference for $n\gamma^2$, and hence π_{MAX}, can easily be carried out using widely available software. Recall that for fixed numerator and denominator degrees of freedom, *F*-distributions are stochastically ordered in their non-centrality parameters. Hence, having observed $T^2 = t_{obs}^2$ and setting $f_{obs} = (n-k+1)t_{obs}^2 / ((n-1)(k-1))$, exact size α, unbiased tests for (4.1) and (4.2) are given respectively by rejection regions $f_{obs} \geq F_{1-\alpha,k-1,n-k+1,n\gamma_0^2}$ and $f_{obs} \leq F_{\alpha,k-1,n-k+1,n\gamma_0^2}$, where $F_{\lambda,\nu_1,\nu_2,\omega}$ denotes the λ quantile of a F-distribution with degrees of freedom ν_1, ν_2 and non-centrality parameter ω. Let $'\mathbf{l}'$ denote the lower alternative in (4.2) and $'\mathbf{u}'$ denote the upper alternative in (4.1). Rejecting the null hypothesis in (4.1) if the $p-value_{\mathbf{u}}(\pi_0) \equiv P(V \geq f_{obs}) \leq \alpha$ leads to an exact size α test, where $V \sim F_{k-1,n-k+1,n\gamma_0^2}$. Similarly, an exact size α test for (4.2) can be carried out by rejecting the null hypothesis if $p-value_{\mathbf{l}}(\pi_0) \equiv P(V \leq f_{obs}) \leq \alpha$. The power functions at $\pi = \pi^*$ for (4.1) and (4.2) are given respectively by $\kappa_{\mathbf{u}}(\pi^*) = P(V \geq F_{1-\alpha,k-1,n-k+1,n\gamma_0^2})$ and $\kappa_{\mathbf{l}}(\pi^*) = P(V \leq F_{\alpha,k-1,n-k+1,n\gamma_0^2})$, where $V \sim F_{k-1,n-k+1,n\omega^*}$, $\omega^* = (\Phi^{-1}(\pi^*))^2$. Inverting these tests, a two sided, $1-\alpha$ confidence interval for π_{MAX} is given by

$$\{\pi_{MAX} = \Phi(\gamma); F_{\alpha/2,k-1,n-k+1,n\gamma^2} < f_{obs} < F_{1-\alpha/2,k-1,n-k+1,n\gamma^2}\}, \tag{4.4}$$

which is easily constructed via a grid search or by using a bisection algorithm. One sided confidences can be constructed similarly. Specifically, (4.4) consists of those values of π_0 for which neither (4.1) or (4.2) is rejected using the data at hand at type 1 error rate $\alpha/2$. If $p-value_{\mathbf{u}}(0.50) > \alpha/2$, set the lower endpoint of the interval at π_{MAX} = 0.50. If in addition $p-value_{\mathbf{u}}(0.50)$ > 1- $\alpha/2$, the data do not restrict the parameter space and we then set (4.4) equal to the parameter space, [.50, 1.0)

The *global*, non-directional effect size $\pi_{MAX} = Max\{\pi(\mathbf{l}); l \in \mathbf{L}^{ALL}\}$ can be modified to accommodate targeted comparisons based on contrasts spanned by a pre-selected family of linearly independent contrast vectors $\mathbf{L} = \{\mathbf{l}_j; j = 1,2,...,m; m < k-1\} \subset \mathbf{L}^{ALL}$ by taking the rows of the matrix $\mathbf{C}$ in (4.3) to be the vectors $\{\mathbf{l}'_j; j = 1,2,...,m\}$ and setting $\pi_{MAX(SP(\mathbf{L}))} = Max\{\pi(\mathbf{l}); \mathbf{l} \in SP(\mathbf{L})\}$, where $SP(\mathbf{L})$ denotes the space spanned by $\{\mathbf{l}'_j; j = 1,2,...,m\}$. Of greater practical interest in cases like this, would probably be $\pi_{MAX(\mathbf{L})} = Max\{\pi(\mathbf{l}); \mathbf{l}_j \in \mathbf{L}, j = 1,2,...,m\}$ If, for example, the main interest of a study lies in *m* specific pairwise comparisons, each of the vectors in $\mathbf{L}$ would have two nonzero entries, one is 1 and the other is negative 1 (-1). Further, if the goal were to compare *k*-1 treatments to control and $\pi_{MAX(\mathbf{L})} = .51$, it would be difficult to argue that in a practical sense any of the treatments differed from the control even though the distributions of all *k* treatments were not identical. Note that this formulation allows for one sided stochastic orderings used in non-inferiority trials, since both $\mathbf{l}$ and $-\mathbf{l}$ need not be in $\mathbf{L}$. Approximate inference for $\pi_{MAX(\mathbf{L})}$ may be based on combinations of individual p-values, as follows. Let $t_{obs}(\mathbf{l}) = \sqrt{n}\mathbf{l}'\bar{\mathbf{x}} / \sqrt{\mathbf{lSl}}$ and $p\text{-}value(\mathbf{l}) = P(V \geq t_{obs}(\mathbf{l}))$, where $V \sim t_{n-1, n(\Phi^{-1}(\pi_0))^2}$ and $t_{v,\lambda}$ denotes a non-central t-distribution with v degrees of freedom and non-centrality parameter λ. We have that that $p\text{-}value(\mathbf{l}) \geq p\text{-}value(\mathbf{l}_{MAX}) = p - value_{\mathbf{u}}$. Let $\{p - value_{(j)}; j = 1,2,...,m\}$ denote the ordered p-values from testing each of the contrasts in $\mathbf{L}$, all using either the hypotheses in (4.1) or (4.2). [9] provides ways of combining these correlated p-values to arrive at a composite test that has overall type 1 error rates approximately bounded above by a desired α. For example, such a test could be given by ' reject H_0 if $m*(p - value(\mathbf{l}_j)) \leq j\alpha;$ for any $\mathbf{l}_j \in \mathbf{L}\}$ '. Further study of these tests is needed.

5. Examples

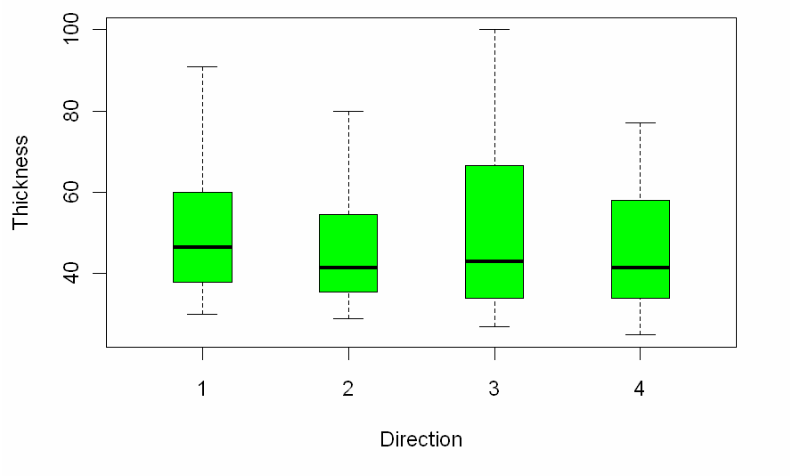
We present three examples where confidence intervals for the effect size π_{MAX} modify the conclusions that would be reached just based on statistical significance.

Example 1.1 (continued) The 0.95 two sided confidence interval for the protocol main effect π_{MAX} , [.500, .750], includes a very wide range of values that indicates the large uncertainty in the information that these data contain about practical differences among protocols. As noted in the introduction, this confidence interval does not rule out the possibility, at the high end, of systematic differences among the protocols. Specifically, although protocol means are close on the *rmr* scale and not statistically significantly different, some specific combination of protocols might result in responses systematically different from others as much as three quarters of the time they were used.

Example 5.1. [8] measured the thicknesses of cork borings on each of $n = 28$ trees in the $k = 4$ four directions of the compass. We treat these trees as having resulted from a random sample taken from some population of interest Although side-by side boxplots of the data presented in Figure 5.1 exhibit considerable overlap of the responses in the four directions, the p-values reported by the standard *F*-test, which assume equal direction variances, and Hotelling's T^2 for equal direction means are respectively 0.0039 and 0.0021. The 0.95 two sided confidence interval for π_{MAX} obtained by using (4.4) is (0.62, 0.89), a wide range of values that like the interval in Example 1.1 fails to strongly support either concluding that there is a large or small separation among the distributions. Specifically, the null hypothesis in (4.1) would only be rejected for $\pi_0 \le 0.62$ and the null hypothesis in (4.2) only rejected for $\pi_0 \ge 0.89$, both at α = 0.025, where both values of π_{MAX} are closer to indicating statistical significance rather than practical significance. Again, the large width of this confidence interval for π_{MAX} is a cautionary warning against interpreting the small p-values obtained from testing for equal means as evidence of practical significance among the direction distributions.

Example 5.2. [11] describes an agricultural experiment carried out as a split plot design as given in (2.2), where the whole plot treatments are specific varieties, $\mathrm{v}_1, \mathrm{v}_2$ and v_3, and the split plot treatments are four specific levels of applied nitrogen, n_0, n_1, n_2, n_3 , independently replicated in $n = 6$ blocks, assumed

Figure 5.1 Boxplots of Thickness



here to be a random sample of all such blocks. Side by side boxplots of the responses across blocks to amounts of nitrogen and the varieties are presented in Figure 5.2 and Figure 5.3. The p-values for standard tests of no variety, no nitrogen and no nitrogen x variety interaction effects are, respectively, 0.2724, < 0.001 and 0.9322. A .95 confidence interval for π_{MAX} for varieties is given by [.50, .88), a wide range which is consistent at the low end with the considerable overlap seen in the boxes in Figure 5.3 and the lack of statistical significance for the test of no variety effect. Most important, the wide range of this

Figure 5.2 Boxplot of Yields by Nitrogen Levels

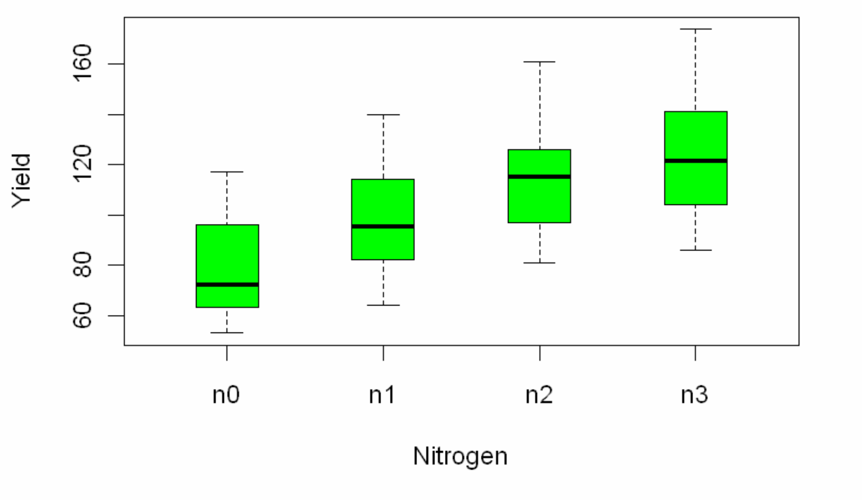
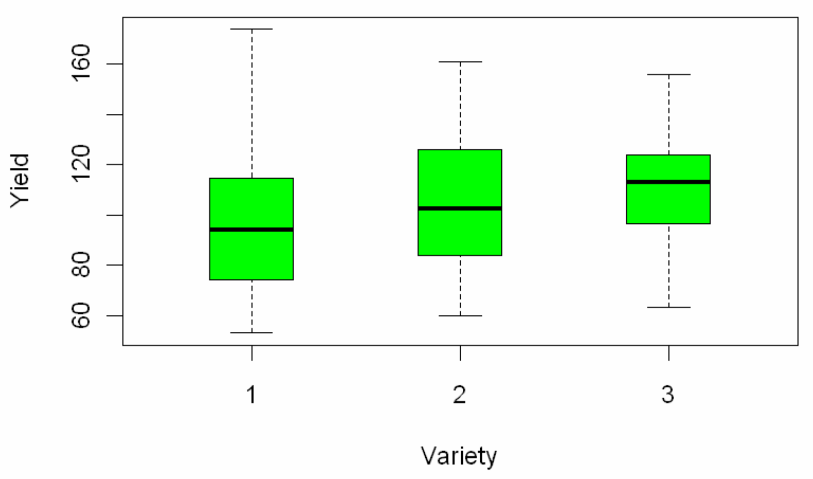


Figure 5.3 Boxplot of Yields by Varieties



interval indicates that this experiment tells us very little about the importance of the impact, which could be large or small, of variety on yield in this particular setting. On the other hand, the very narrow .95 confidence interval for π_{MAX} for the nitrogen main effect has the form (c, 1.0), where due to the limits in the accuracy of our algorithm for obtaining cumulative probabilities of the non-central F- distribution, we only know that c > 0.99998. Thus, in addition to the evidence provided by the very small p-value and distinct separation among the boxplots in Figure 5.2, the very large lower endpoint of the confidence interval for π_{MAX} provides evidence that some combinations of nitrogen levels are very reliably better than others. To investigate the interaction between nitrogen and variety, the 0.95 confidence intervals for the interaction π_{MAX} obtained from pairwise comparisons of the differences between nitrogen levels across the varieties are: [0.50, 0.70] for n_0 vs n_1; [0.50, 0.80] for n_0 vs n_2; [0.50, 0.86] for n_0 vs n_3; [0.50, 0.85] for n_1 vs n_2; [0.50, 0.83] for n_1 vs n_3 and [0.50, 0.85] for n_2 vs n_3. Except for a somewhat smaller range of values in comparing n_0 vs n_1, the wide range of these intervals indicate that these data have little to say about the lack of uniformity among pairwise differences in nitrogen levels across varieties. Similarly, investigating the interaction among varieties across levels of nitrogen, we find that the 0.95 confidence intervals for π_{MAX} are: [0.50,1) for v_1 vs v_2 and [0.50, 0.94] for both v_1 vs v_3 and

v_2 vs v_3, all containing small and large values. In sum, this effect based interaction analysis is unable to rule out the possibilities that interactions between varieties and nitrogen could be of little or great practical importance, a potentially useful augmentation to a traditional analysis which would simply report that there is insufficient evidence to support concluding that there is a variety by nitrogen interaction, p = 0.9322.

6. Simulation

For a fixed number of blocks *n* and number of treatments *k*, without loss of generality, we set σ_g = 1 and independently generated $\{\sigma_i\}$ from a uniform distribution on the interval (.2,1) and, independent of these, generated independent means $\{\mu_i\}$ from a uniform distribution on (0,1). The lower bound of .2 was used to avoid very large γ^2 values which would have required computing cumulative *F*-distribution probabilities for arguments so big that numerical algorithms become unstable. This random choice, a process we carried out 100 times, of parameter settings covered a wide range of cases, including heavy doses of homoscedasticity. We took *k* = 2, 3 and 5 and for each *k* took *n* = k + 1, *k*+ 6, *k*+11, *k*+21, *k* + 31 and *k* + 41. For each of these settings we independently generated 10000 data sets. Since power functions and coverage rates of confidence intervals are exact and given by easily computed explicit formulas, we only briefly summarize some of our representative findings on mean widths of two sided 0.95 confidence intervals for π_{MAX} . From Figure 6.1, consisting of side by side box plots of simulated relative mean interval widths = mean width/ π_{MAX} ,denoted RMW , for different numbers of blocks *n* with *k* = 5, we see that mean relative widths are large for small *n* and decrease steadily as *n* increases.

Specifically, the median of the mean relative widths decreases by about 25% from the smallest sample size to the largest. The pattern and values are very similar for *k* = 2 and 3. Aggregated over n, median $RMW's$ are about 0 .40 for all values of *k*, as can be seen in the side by side box plots in Figure 6.2.

TheFigure 6.1 Boxplots: Mean Relative Width, k = 5

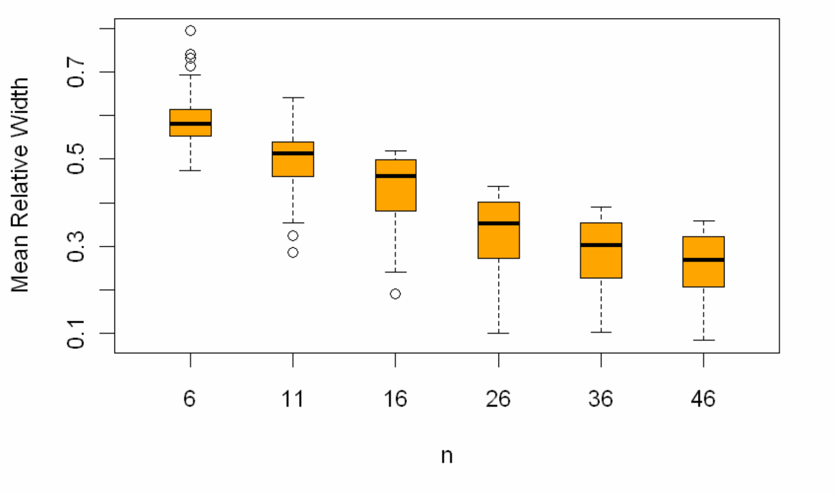
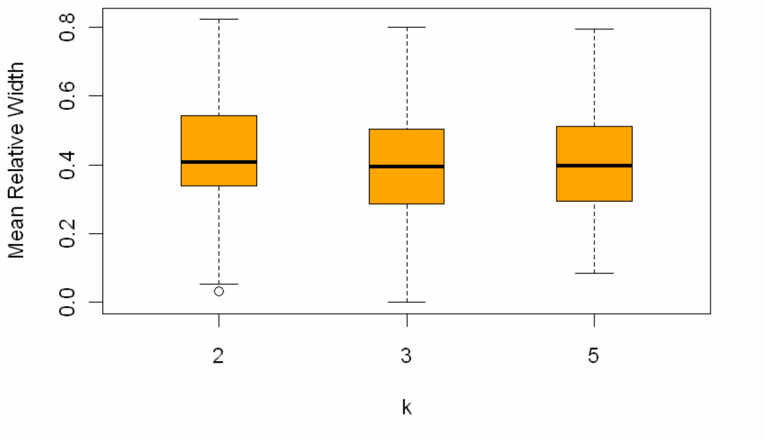


Figure 6.2 Boxplots: Mean Relative Width, k = 2,3,5



plot in Figure 6.3 for $k = 3$ illustrates that RMW is quite stable over values of the standard deviation ratio $RSIG = \max\{\sigma_i\} / \min\{\sigma_i\}$ for all n. The representative plot for $k = 2$ in Figure 6.4 of mean width vs π_{MAX} is approximately quadratic except for very small n, with a maximum approximately at 0.75, the center of the possible values of π_{MAX}. An empirically determined least squares surface fitted to simulated mean widths, denoted MW, results in the surface

$\hat{MW} = -.64 - .00584 * n + .01129 * k + 2.95 * \pi_{MAX} - 2.07 * \pi_{MAX}^2$, $R^2 = .88$. All of the sources of variation are statistically significant and their signs and magnitudes are consistent with the discussion above. In particular, using this surface, everything else held fixed, we estimate that increasing sample size by ten corresponds to a decrease in mean width of about 0.06, that mean width is about 0.03 greater when $k = 5$ than when $k = 2$ and the quadratic relation evident in the plot is supported by the negative coefficient of π_{MAX}^2 .

Figure 6.3 Plot of RMW vs RSIG, k = 3

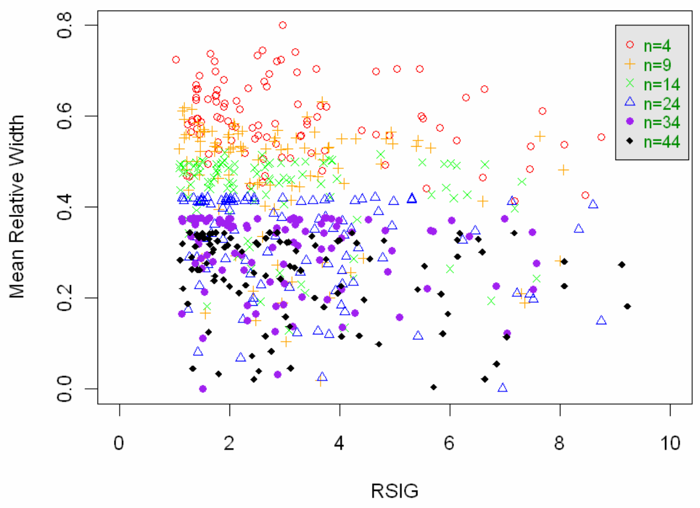
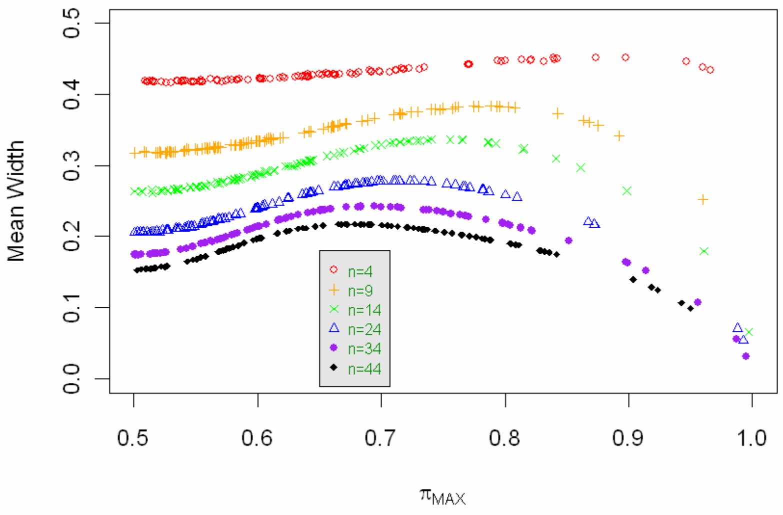


Figure 6.4 Plot of MW vs π_{MAX} , k = 2



7. Conclusion

The effect size π_{MAX}, developed here in the context of correlated responses recorded on a block, is a parameter that can aid in distinguishing between practical and statistical significance. As noted above, although, like all effect sizes, the values of π_{MAX} that determine practical significance are subjective, π_{MAX} can be calibrated in terms of observable consequences and not just as a function of means and covariances, parameters which are never observed. Since inference for π_{MAX} requires estimating the covariance matrix of the responses, the number of blocks, *n*, must be greater than the number of levels of the treatment, *k*. It is not surprising that confidence intervals for π_{MAX} are relatively wide when $n - k$ is small, a fact researchers should consider when designing their experiments. [5] argues that the failure of many published statistically significant results to be practically significant is due to the small effect size of the treatments. We hope that the inference presented here will help researchers better appreciate the often large uncertainty in their estimates of effects and show caution in announcing practical significance.

Appendix

Let $\mathbf{C} = \{\underline{c}_v\}$ be the class of all nonzero, k - dimensional, column vectors of constants and note that $\mathbf{L}^{all} \subset \mathbf{C}$.

Lemma 1: For a fixed vector $\mathbf{x}$ and positive definite matrix **B,**

$$\text{Max}\{\mathbf{c}'\mathbf{x}\mathbf{x}'\mathbf{c} / \mathbf{c}'\mathbf{B}\mathbf{c}\,;\ \mathbf{c} \in \mathbf{C}\} = \mathbf{x}'\mathrm{B}^{-1}\mathbf{x} \equiv Q(\mathbf{c}^{\#}, \mathbf{x}, \mathbf{B}),$$

where $Q(\mathbf{c}, \mathbf{x}, \mathbf{B}) = \mathbf{c}'\mathbf{x}\mathbf{x}'\mathbf{c} / \mathbf{c}'\mathbf{B}\mathbf{c}$ and $\mathbf{c}^{\#} = \mathbf{B}^{-1}\mathbf{x}$.

Proof: See (1f.1.1) on page 60 of Rao (1973).

Lemma 2: Proof of (3.1)

From Lemma 1, for any positive definite matrix $\mathbf{\Sigma}$, since $\Phi(\square)$ is non-decreasing, setting $\mathbf{c}^{\#} \equiv \mathbf{\Sigma}^{-1}(\boldsymbol{\mu} - \bar{\mu}_w\mathbf{1}) \in \mathbf{L}^{All}$ and $\bar{\mu} = \mathbf{1}'\mathbf{\Sigma}^{-1}\boldsymbol{\mu} / \mathbf{1}'\mathbf{\Sigma}^{-1}\mathbf{1}$, we have that

$$\pi_{MAX} \equiv Sup\{\pi(\mathbf{l}); \mathbf{l} \in \mathbf{L}^{all}\}$$

$$= \Phi\left(Sup\{\sqrt{Q(\mathbf{l}, \boldsymbol{\mu}, \mathbf{\Sigma})}; \mathbf{l} \in \mathbf{L}^{all}\}\right)$$

$$= \Phi\left(Sup\{\sqrt{Q(\mathbf{l}, \boldsymbol{\mu} - \bar{\mu}\mathbf{1}, \mathbf{\Sigma}); \mathbf{l} \in \mathbf{L}^{all}}\}\right)$$

$$\leq \Phi\left(\sqrt{Sup\{Q(\mathbf{c}, \boldsymbol{\mu} - \bar{\mu}\mathbf{1}, \mathbf{\Sigma}); \mathbf{c} \in \mathbf{C}\}}\right)$$

$$= \Phi\left(\sqrt{(\boldsymbol{\mu} - \bar{\mu}\mathbf{1})'\,\mathbf{\Sigma}^{-1}(\boldsymbol{\mu} - \bar{\mu}\mathbf{1})}\right)$$

$$= \Phi\left(\sqrt{Q(\mathbf{c}^{\#}, (\boldsymbol{\mu} - \bar{\mu}\mathbf{1}), \mathbf{\Sigma})}\right)$$

$$= \Phi\left(Sup\{\sqrt{Q(\mathbf{l}, \boldsymbol{\mu} - \bar{\mu}\mathbf{1}, \mathbf{\Sigma}); \mathbf{l} \in \mathbf{L}^{all}}\}\right)$$

$$= \pi_{MAX}.$$

References

[1] R. Bakeman, Recommended effect size for repeated measure designs, *Behav. Res. Meth*. 37(3), (2005), pp. 379-84.

[2] R.C. Bullough and C.L. Melby, Effect of inpatient vs outpatient measurement protocol on resting metabolic rate and respiratory exchange ratio. *Annals of Nutrition and Metabolism*, 37, 1993, pp. 24-32.

[3] R.J. Grissom and J.J. Kim, Effect Size for Research: Univariate and Multivariate Applications, Second Edition, Routlege, NY, 2012.

[4] H. Huyn and L.S. Feldt, Conditions under which mean square ratios in repeated measure designs have *F*-distributions. *JASA*, 65, (1970), pp. 1582- 1589.

[5] J.P.A. Ioannidis, Why most published research findings are false. PloS Med 2(8):e124. doi:10:13, (2005).

[6] Y. Ling and P.I. Nelson, P.I. Effect size for comparing two or more normal distributions based on maximal contrasts in outcomes, *Stat. Methods Appl. 23*, (2014), pp. 381-399.

[7] S. Olejnik and J. Algina, Generalized Eta and Omega squared statistics: Measurements of effect size for some common research designs, *Psychological Methods*, 8, (2003), pp. 434-47.

[8] C. R. Rao, Tests of significance in Multivariate Analysis. *Biometrika.35*, (1948), pp. 58-79.

[9] R.J. Simes, An improved Bonferroni Procedue for multiple tests of significance. *Biometrika*, 73, (1986), pp. 751-754.

[10] E.J. Williams, Comparing means of correlated variates, *Biometrika*, *57*, (1970), pp. 459-461.

[11] F. Yates, Complex Experiments, Reprinted in Experimental Design, Selected Papers, Griffen, London, (1970), pp. 71-117.

Captions for Figures