Published in final edited form as:

Commun Stat Theory Methods. 1996 ; 25(7): . doi:10.1080/03610929608831787.

BIAS IN LINEAR MODEL POWER AND SAMPLE SIZE CALCULATION DUE TO ESTIMATING NONCENTRALITY

Douglas J. Taylor and

Dept. of Biostatistics, CB#7400 University of North Carolina Chapel Hill, North Carolina, 27599

Keith E. Muller

Dept. of Biostatistics, CB#7400 University of North Carolina Chapel Hill, North Carolina, 27599

Abstract

Data analysts frequently calculate power and sample size for a planned study using mean and variance estimates from an initial trial. Hence power, or the sample size needed to achieve a fixed power, varies randomly. Such calculations can be very inaccurate in the General Linear Univariate Model (GLUM). Biased noncentrality estimators and censored power calculations create inaccuracy. Censoring occurs if only certain outcomes of an initial trial lead to a power calculation. For example, a confirmatory study may be planned (and a sample size estimated) only following a significant result in the initial trial.

Computing accurate point estimates or confidence bounds of GLUM noncentrality, power, or sample size in the presence of censoring involves truncated noncentral F distributions. We recommend confidence bounds, whether or not censoring occurs. A power analysis of data from humans exposed to carbon monoxide demonstrates the substantial impact on sample size that may occur. The results highlight potential biases and should aid study planning and interpretation.

Keywords

effect size; noncentrality; confidence bounds

1. INTRODUCTION

1.1 Motivation

Study planning usually includes power calculations. Computing a power or sample size (N) value typically involves estimating one or more distribution parameters. Hence computed power, for a fixed N , varies randomly. Likewise the sample size necessary to obtain a fixed power varies randomly. Data analysts often fix mean values at “clinically significant” differences (δ) and estimate power based upon an existing variance estimate. Here we focus on the many instances with power estimated, sometimes naively, assuming random variance and mean values. For example, the Food and Drug Administration generally requires evidence of efficacy from more than a single study before approving a new drug for commercial use. In planning a confirmatory trial the observed (and hence random) mean values from an initial, significant result may provide the parameter values for a power calculation. Likewise, following a non-significant result one may wonder whether running a subsequent trial with a few more subjects might allow winning the scientific bet. A researcher may estimate the N necessary to obtain a desired power in the second trial based on a mean difference observed in the initial study. Validity of traditional power computations requires that the chosen δ has been “fixed,” when in fact it was observed randomly. As demonstrated in §3 and §4, such calculations can be very inaccurate in the General Linear Univariate Model (GLUM).

In the GLUM power varies as a one-to-one function of the noncentrality parameter of an F distribution. While noncentrality lies in the interval $[0, \infty)$, power lies in the bounded interval $[\alpha, 1)$, with α the desired significance level of the test.

At least two sources of inaccuracy may arise in GLUM power calculations. First consider the use of biased estimators of noncentrality (see §3.1), with the obvious (and typical) estimate constructed as a function of separate variance and mean values from an initial study. Alternately, maximum likelihood estimates depend only on a single random variable, the test statistic, and do not require separate estimation of means and variances. As seen in §4, however, these estimators typically have negative bias for large values of true power and positive bias for small values. Although an unbiased estimator of noncentrality exists, it can yield improper (negative) estimates. The restriction of power to the interval $[\alpha, 1)$ exacerbates the problem for power estimation.

In contrast to the difficulties with point estimates, accurate confidence bounds always allow useful interpretations with respect to specified confidence levels. As a special case, an exact 50% confidence limit for noncentrality corresponds to a median unbiased point estimate of non centrality.

A second type of bias occurs when power calculations for a planned *target* study are only performed if an initial *screening* study's test statistic achieved, or alternately failed to achieve, a desired significance level. This behavior acts as a censoring mechanism which, when neglected, can lead to inaccurate estimates of power and N .

Either left or right censoring can occur in practice. Assume larger values of the test statistic correspond to smaller p -values. Then left censoring occurs with power calculations based on estimates from earlier "significant" results. For example, a scientist may test many compounds for efficacy or toxicity, but only plan to do a confirmatory study, and therefore desire a power calculation, following a significant result. Hence a power calculation only results if the original test statistic achieved a desired *screening* significance level, α_s . Note that α_s may correspond to an overall type I error rate of, say, .20, and not to the significance level required for the *target* study, α_t (often .05). Right censoring occurs with power analyses based on earlier, non-significant, results. Consider a hypothesis test resulting in a claim of no effect at the α_s significance level. A researcher might subsequently choose to estimate the N necessary to obtain a significant result in a second trial, based on mean differences observed in the initial study. The analyst only performs the sample size calculation because the screening test statistic failed to achieve a desired significance level, α_s .

Failing to consider censoring mechanisms in power calculations typically results in optimistic bias under left censoring and pessimistic bias under right censoring (see §3 and §4). Either invalidates standard computations of power and N . The new results shed light on practical aspects of power calculations, and should serve to instill caution in scientists hoping to find significant results in a planned study when little evidence for such findings exists.

1.2 Literature Review

Many authors have described approximate and exact methods for computing power and N for variations of the GLUM assuming fixed expected value and variance parameters. See Muller, LaVange, Ramey, and Ramey (1992) for a review in the more general multivariate context. Such traditional presentations fail to address bias in power calculations due to estimation of model parameters.

Confidence bounds for the noncentrality parameter of an F distribution have been examined in the past. Wright and O'Brien (1988) extended results of Dwass (1955) to provide bounds for noncentrality and used these to compute bounds for power. Their methods were implemented in JMP® Version 2 (1989). The approximate techniques can provide liberal results. Venables (1975) suggested a truncated Cornish-Fisher expansion to provide bounds for noncentrality. The approximation works best for large power. Other authors, including Hardison, Langston, and Quade (1980), Lam (1987), and Guirguis (1990) have presented algorithms which allow calculating confidence bounds for noncentrality. The various algorithms, together with proper treatment of certain boundary conditions not discussed by any of the authors, allow computing exact bounds for noncentrality in the absence of censoring. These methods can be extended to provide confidence bounds for power and N , but may be very inaccurate when censoring mechanisms affect study planning.

1.3 Statement of the Model and Hypothesis

For N sampling units and q predictors state the GLUM as (Ch. 5, Searle, 1971):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \quad (1.1)$$

$N \times 1 \quad N \times q \times 1 \quad N \times 1$

Here $\mathbf{X}$, of rank $r \leq q < N$, contains fixed, known constants and $\boldsymbol{\beta}$ contains fixed, unknown parameters. Assume $\mathbf{e} \stackrel{d}{=} N_N(0, \sigma^2 \mathbf{I})$, with $0 < \sigma^2 < \infty$. The general linear hypothesis involves the secondary parameter $\boldsymbol{\theta} = C\boldsymbol{\beta}$. Interest lies in testing

$$\mathbf{H}_0: \boldsymbol{\theta}_{\nu_1 \times 1} = \boldsymbol{\theta}_0 \quad (1.2)$$

versus

$$\mathbf{H}_A: \boldsymbol{\theta} \neq \boldsymbol{\theta}_0. \quad (1.3)$$

Both $\boldsymbol{\theta}_0$ and C contain fixed constants chosen by the data analyst. Assume full rank for C . The likelihood-ratio test statistic equals

$$f_{obs} = \frac{SSH(\hat{\boldsymbol{\theta}}, N) / \nu_1}{SSE / \nu_2} = \frac{SSH(\hat{\boldsymbol{\theta}}, N) / \nu_1}{\hat{\sigma}^2}, \quad (1.4)$$

in which

$$SSH(\hat{\boldsymbol{\theta}}, N) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \left[C(\mathbf{X}'\mathbf{X})^- C' \right]^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0). \quad (1.5)$$

$\hat{\boldsymbol{\theta}}$ represents the least squares and maximum likelihood estimate of $\boldsymbol{\theta}$ and $(\cdot)^-$ indicates a generalized inverse. The test statistic follows an F distribution with ν_1 (rank of C) numerator degrees of freedom (df), $\nu_2 = N - r$ denominator df , and noncentrality parameter

$$\omega = \frac{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \left[C(\mathbf{X}'\mathbf{X})^- C' \right]^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)}{\sigma^2} = \frac{SSH(\boldsymbol{\theta}, \mathbf{N})}{\sigma^2}. \quad (1.6)$$

The notation $SSH(\cdot, N)$ emphasizes that SSH depends on sample size via $\mathbf{X}$.

Represent the cumulative distribution function (CDF) of a noncentral F with ν_1 df, ν_2 df, and noncentrality ω by $F_F[\cdot|\nu_1, \nu_2, \omega]$, with corresponding density function $f_F[\cdot|\nu_1, \nu_2, \omega]$. Using this notation, write the power of the test as

$$P = 1 - F_F[f_{crit}(1 - \alpha_t) | \nu_1, \nu_2, \omega], \quad (1.7)$$

with $f_{crit}(1 - \alpha_t)$ the $100 \cdot (1 - \alpha_t)$ percentile of a central F with ν_1 and ν_2 df. The CDF may be expressed as a Poisson-weighted infinite sum (Johnson and Kotz, p192, 1970). In practice, CDF and density values may be accurately calculated with computer algorithms (Abramowitz and Stegun, p947, 1964; Lentz, 1987; Posten, 1989; Odeh and Fox, Chapter 5, 1991).

Some authors, notably Cohen (1987, p8), have discussed effect size as a generic concept encompassing disparate analyses. For the special cases of the GLUM considered by Cohen, effect size and noncentrality can be expressed as one-to-one functions of each other. Although distinct cases of the GLUM yield distinct formulae for effect size, they yield only a single formula for noncentrality. Hence we restrict our attention to noncentrality in order to maximize generality.

2. THE TRUNCATED DISTRIBUTIONS

Traditional GLUM power calculations depend on knowing the values of all relevant distribution parameters. In practice, a power analysis often requires an estimate of ω based on the components of a noncentral F statistic from an initial screening study, f_s . Maximum likelihood estimates and confidence bounds for ω follow from maximizing or integrating the likelihood of f_s , respectively (see §3). Under censoring, however, a power analysis only results given that f_s achieved a particular screening criterion. The appropriate likelihoods on which to base such conditional power calculations follow from *truncated* noncentral F distributions.

Consider planning a power analysis for a confirmatory trial only if an initial screening study achieves a desired level of significance, α_s , specified in advance (*left censoring*). The p -value from the initial study must be less than α_s or no confirmatory trial will be conducted. The appropriate likelihood for a power analysis based on f_s , given that $0 \leq f_s \leq f_{crit}(1 - \alpha_s)$, follows from a left-truncated noncentral F distribution. The likelihood equals zero for f_s in the interval $[0, f_{crit}(1 - \alpha_s))$, while otherwise

$$f_L[f_s; \nu_1, \nu_2, \omega] = \frac{f_F[f_s | \nu_1, \nu_2, \omega]}{1 - F_F[f_{crit}(1 - \alpha_s) | \nu_1, \nu_2, \omega]}. \quad (2.1)$$

Setting $\alpha_s = 1$ corresponds to the uncensored case, and reduces equation (2.1) to the usual noncentral F likelihood.

Next consider planning a power analysis for a second trial only if an initial study fails to achieve a desired significance level, α_s (*right censoring*). The p -value from the screening study must be greater than α_s or a second trial will not be conducted. The appropriate likelihood for a power analysis based on f_s , given that $f_s \geq f_{crit}(1 - \alpha_s)$, follows from a right-truncated noncentral F distribution. The likelihood equals zero for f_s in the interval $[f_{crit}(1 - \alpha_s), \infty)$, while otherwise

$$f_R[f_s; \nu_1, \nu_2, \omega] = \frac{f_F[f_s | \nu_1, \nu_2, \omega]}{F_F[f_{crit}(1 - \alpha_s) | \nu_1, \nu_2, \omega]}. \quad (2.2)$$

Setting $\alpha_s = 0$ corresponds to the uncensored case, and reduces equation (2.2) to the usual noncentral F likelihood.

It is important to recognize that neither the distribution of f_s , nor the distribution of the planned target study's test statistic, f_t , are truncated. Equations (2.1) and (2.2) are only specified to obtain mle's and confidence bounds for ω based on the observed value of f_s , given that *censoring occurred*.

The above formulations and related results extend in an obvious way to two-sided censoring mechanisms. For example, a second trial may only be planned if the original study p -value falls within the range $[a, b]$, in which $0 < a < b < 1$. Such extensions will not be considered further in this paper.

3. ESTIMATING NONCENTRALITY, POWER, AND SAMPLE SIZE

3.1 Candidate Estimators of Noncentrality

Various point estimators of ω have some appeal. For example, the obvious estimate of noncentrality,

$$\hat{\omega}_o = \frac{SSH(\hat{\theta}, N)}{\hat{\sigma}^2} = \nu_1 \cdot f_s, \quad (3.1)$$

equals ω when the observed statistics $\hat{\theta}$ and $\hat{\sigma}^2$ equal the true population values. In practice this estimator is biased. Based on the expected value of f_s , a linear function of (3.1) exists which results in an unbiased estimator. However, it can produce improper (negative) estimates. The mle of ω ($\hat{\omega}_{mle}$) follows by maximization of (2.1) or (2.2). Proper treatment of censoring fails to remove the potentially substantial bias of the mle, however.

The less attractive features of point estimators mostly arise from the strong asymmetry typically present in the distributions of interest. In contrast, confidence bounds have very appealing and useful interpretations. As a special case, exact 50% confidence bounds correspond to median unbiased point estimates of ω (overestimating ω half the time and underestimating ω half the time). The new results presented here allow computing exact bounds for ω , power and N , regardless of censoring.

3.2 Confidence Bounds for Noncentrality

All calculations require carefully maintaining the distinction between the initial (*screening*) study, which provides a parameter estimate, and the planned (*target*) study, for which the data analyst desires a power value. For example, the value of $f_{crit}(1 - \alpha_s)$ depends on the screening study error d_f , while the value of $f_{crit}(1 - \alpha_t)$ requires the target study error d_f . To help facilitate the necessary distinction use the notation ν_{2s} and ν_{2t} to represent the error d_f associated with the screening and target studies, respectively.

Calculate exact confidence bounds for ω as follows. Assign a lower tail probability $\alpha_q \in (0,1)$ corresponding to the desired confidence level. Then solve

$$F_C[f_s; \nu_1, \nu_{2s}, \hat{\omega}_q] = 1 - \alpha_q \quad (3.2)$$

for $\hat{\omega}_q$, in which the index $C \in \{L, R, N\}$ denotes Left, Right, or Not censored, respectively, and $F_C[f_s; \cdot]$ represents the corresponding likelihood integrated from zero to f_s . An upper one-sided $100 \cdot (1 - \alpha_q)$ percent confidence interval follows as $[\hat{\omega}_q, \infty)$, while solving equation (3.2) for two distinct α_q 's yields a two-sided interval. Letting $\alpha_q = .50$ provides a

median unbiased estimate of ω . Notice that in the event $f_s < F_C^{-1}[1 - \alpha_q | \nu_1, \nu_{2s}, 0]$, no positive $\hat{\omega}_q$ exists to satisfy equation (3.2). When this occurs set $\hat{\omega}_q = 0$ to guarantee a proper estimate and exact coverage for an upper one-sided interval. The proof follows from close consideration of the probability of this event, given the true value of ω .

3.3 Confidence Bounds for Power

Point estimates and confidence bounds for power follow from the equation

$$\hat{P} = 1 - F_F[f_{crit}(1 - \alpha_t) | \nu_1, \nu_{2s}, \hat{\omega}], \quad (3.3)$$

in which $\hat{\omega}$ equals the corresponding point estimate or confidence bound for ω . The smooth, strictly increasing dependence of power on ω insures that

$$\Pr\{\hat{\omega} \leq \omega\} = \Pr\{\hat{P} \leq P\}. \quad (3.4)$$

Consequently exact confidence bounds for ω lead to exact bounds for power. Note that equation (3.3) provides confidence bounds for the power of a target study based on a strict replication of the screening trial.

3.4 Confidence Bounds for Sample Size

A confidence bound for sample size ($\hat{N}_t$) provides the N which insures with confidence $(1 - \alpha_q)$ a desired power, P , for the *target* study. For the ratio $R_N = \hat{N}_t / N_s$, with N_s the screening study sample size, recognize that

$$\Pr\{\hat{\omega}_q \leq \omega\} = \Pr\{R_N \cdot \hat{\omega}_q \leq R_N \cdot \omega\}, \quad (3.5)$$

with $\hat{\omega}_q$ the bound for noncentrality in a strict replication of the initial study. The appropriate value of $\hat{N}_t$ satisfies the equation

$$\hat{\omega}_{N_t} = R_N \cdot \hat{\omega}_q, \quad (3.6)$$

under the constraint on $\hat{N}_t$ that

$$P = 1 - F_F[f_{crit}(1 - \alpha_t); \nu_1, \nu_{N_t}, \hat{\omega}_{N_t}], \quad (3.7)$$

with both ν_{N_t} and $\hat{\omega}_{N_t}$ taken as functions of $\hat{N}_t$. Iterative evaluation of (3.6) and (3.7) allows straightforward estimation of $\hat{N}_t$.

No $\hat{N}_t$ satisfies equation (3.7) when $\hat{\omega}_q = 0$, since $\hat{\omega}_{N_t}$ then equals zero for all choices of R_N . This should warn that the initial trial provides little evidence for the existence of an effect, and a second study may simply waste resources.

4. SIMULATION STUDIES

4.1 Overview

Simulations were conducted with SAS IML® to demonstrate the impact of censoring and to assess the accuracy of the new techniques. In the first simulation 50% confidence bounds and mle's for ω and power of a one-way ANOVA were calculated adjusting for left censoring. In addition, 50% confidence bounds and the "obvious" estimates of ω and power

(see §3.1) were calculated ignoring the censoring mechanism. The second simulation paralleled the first, but with right censored data and without calculation of mle's. In the third simulation adjusted upper one-sided confidence intervals were estimated under both left and right censoring, for $\alpha_q \in \{.05, .50\}$. All three simulations used $\alpha_t = .05$ with all combinations of target power ($P \in \{.1, .5, .9\}$), number of treatment groups ($G \in \{2, 4\}$), cell size ($N_G \in \{4, 8, 16\}$), and screening level ($\alpha_s \in \{.05, .20\}$).

4.2 Method

Without loss of generality, ($\sigma^2 = 1$ for all simulations. Cell means were chosen as $\beta = \gamma \cdot [0 \ 1 \ \dots \ G - 1]'$, with γ calculated by iterative use of a power program (see the Appendix) to ensure the target power for each combination of G , N_G , and α_q . A total of 1500 pseudo-random samples of Gaussian data were generated for each combination, independently for each simulation.

The “obvious” estimates of noncentrality and power ($\hat{P}_o$) were calculated using the noncentrality estimator in equation (3.1). Confidence bounds for noncentrality and power ($\hat{P}_q$) *unadjusted* for censoring were obtained with the SAS© FNONCT function. Without access to SAS, the algorithms described by Lam (1987) and Guirguis (1990) may be used.

Adjusted mle's for ω and power under left censoring ($\hat{P}_{L,mle}$) were obtained by applying an interval halving algorithm to equation (2.1). Adjusted q -percent confidence bounds for ω and power under left censoring ($\hat{P}_{L,q}$) were obtained by applying an interval halving algorithm to solve for $\hat{\omega}_q$ in the equation

$$\frac{F_F[f_s; \nu_1, \nu_{2s}, \hat{\omega}_q] - F_F[f_{crit}(1 - \alpha_s); \nu_1, \nu_{2s}, \hat{\omega}_q]}{1 - F_F[f_{crit}(1 - \alpha_s); \nu_1, \nu_{2s}, \hat{\omega}_q]} = 1 - \alpha_q. \quad (4.1)$$

Similarly, bounds for ω and power adjusted for right censoring ($\hat{P}_{R,q}$) were obtained from

$$\frac{F_F[f_s; \nu_1, \nu_{2s}, \hat{\omega}_q]}{F_F[f_{crit}(1 - \alpha_s); \nu_1, \nu_{2s}, \hat{\omega}_q]} = 1 - \alpha_q. \quad (4.2)$$

Boundary conditions discussed in §3.2 were treated appropriately. A maximum tolerance of 10^{-5} was used for all mle and confidence bound calculations.

Values of $F_F[\cdot | \nu_1, \nu_2, \omega]$ were obtained with the SAS© PROBF function. Without access to SAS, the algorithms described by Lenth (1987) and Posten (1989) may be used. The PROBF function failed to compute some extreme values (for SAS 6.08 for OS/2). Severo and Zelen's approximation was used (Johnson and Kotz, p 194, 1970) to overcome the limitation. If the absolute value of the approximate unit normal was greater than 6.0, $F_F[\cdot | \nu_1, \nu_2, \omega]$ was rounded to zero or one, as appropriate.

4.3 Results for Left Censoring

As discussed in §3.2 and §3.4, a confidence bound for noncentrality may need to be truncated at zero. When this occurs α_t provides an exact bound for the power of a strict replication of the initial study, while the sample size necessary to achieve a desired target power (greater than α_t) in a subsequent trial can not be estimated. The importance of this event ($\hat{P} = \alpha_t$) led us to tabulate the percentage of time it occurred in our simulations.

Table I contains a summary of simulation results under left censoring. The “obvious” and $\hat{P}_{0.5}$ estimates of power ignoring censoring were extremely optimistic for low true power, with bias decreasing as power increased. Even adjusted for censoring, $\hat{P}_{L,mle}$ estimates were liberal for low power and conservative for high power. On average the $\hat{P}_{L,.5}$ estimates were likewise inaccurate. As stated previously, however, this estimator is *median* unbiased, overestimating P half the time and underestimating P half the time. For example, mean coverage for 50% confidence bounds was $.501 (\pm .0012)$. Hence unlike the other point estimators, there is reasonable justification for using $\hat{P}_{L,.5}$. Accurate coverage was also obtained using 95% confidence intervals, with mean coverage of $.948 (\pm .0028)$ across all conditions in our simulation.

The percent truncation $(\hat{P}=\alpha_t)$ was between 3% ($P = .90$ and $\alpha_s = .20$) and 42% ($P = .10$ and $\alpha_s = .05$) for $\hat{P}_{L,.5}$ estimates, and between 30% and 92% for the 5% confidence bound. Thus even after a significant result in an initial trial, a point estimate or confidence bound for the power of a target study may equal α_t (and $\hat{\omega}_q=0$) with non-trivial probability.

4.4 Results for Right Censoring

Table II contains a summary of simulation results under right censoring. The “obvious” and $\hat{P}_{0.5}$ estimates of power ignoring censoring were extremely pessimistic for high power, with bias decreasing as true power decreased. Even adjusted for censoring, average $\hat{P}_{R,.5}$ estimates were liberal for low power and conservative for high power. As stated previously, however, this point estimator is median unbiased, overestimating P half the time and underestimating P half the time. Averaged across all simulation conditions, mean coverage for 50% confidence intervals was $.501 (\pm .0034)$. Accurate coverage was also obtained using 95% confidence intervals, with mean coverage of $.948 (\pm .0015)$ across all simulated conditions.

The percent truncation $(\hat{P}=\alpha_t)$ was between 2% ($P = .90$ and $\alpha_s = .05$) and 44% ($P = .10$ and $\alpha_s = .20$) for $\hat{P}_{R,.5}$ estimates and between 48% and 93% for the 5% confidence bound. Thus even after failing to obtain a significant result in an initial trial, a useful point estimate or confidence bound for the N necessary to achieve a significant result in a subsequent study can sometimes be obtained.

5. AN EXAMPLE

Equivocal results as to whether exposure to a low level of carbon monoxide (CO) reduces human performance (Laties and Merigan, 1979) led the U. S. EPA and Army to fund Benignus, Muller, Barton, and Prah (1987) to replicate results by Putz, Johnson, and Setzer (1976). The 1976 data were used for power analyses. As effects of any size were of potential consequence to the U. S. population, mean values were not considered fixed for these analyses. Design factors included CO (0, 35, and 70 ppm; between subject), Frequency (high or low; within subject), and Hour (1, 2, 3, 4; within subject). Effects significant at the .05 level in the original analysis were of primary interest for replication. Hence in this example power estimates were left censored, with $\alpha_s = .05$.

Table III contains power estimates based on the “univariate” approach to repeated measures ANOVA. Computed values reflect the impact of considering only effects significant at the .05 level in the initial study. Table III includes both $\hat{P}_{L,.5}$ estimates adjusting for censoring,

$\hat{P}_{.5}$ estimates ignoring censoring, and one-sided 90% confidence intervals. Results for effects of $\text{CO} \times \text{Hr}$ and $\text{CO} \times \text{Fr} \times \text{Hr}$ differed little due to censoring, with both having p -values less than .001 in the initial study. In contrast, results for the $\text{CO} \times \text{Fr}$ interaction term ($p = .035$) differed substantially. Ignoring censoring led to a power estimate of $\hat{P}_{.5} = .608$, while adjusting for censoring led to $\hat{P}_{L,.5} = .05$. Hence estimated power for a strict replication of the hypothesis test was actually small. The fact that the p -value for the initial test barely achieved the screening criterion agrees with this result. Note that the investigators did not have the censored quantile information presented here. This likely would have influenced the choice of study plan.

6. BIAS WITH FIXED MEANS

The methods in §3 apply to power calculations based on random variances and random means, in contrast to calculations based on random variances and *fixed* means. Taylor and Muller (1995) provided exact confidence intervals in this setting under the restriction of no censoring. However, censoring can occur in practice. For example, a study's power for a clinically significant difference following a negative result may be important. Regulatory decisions often hinge on such evidence. With human or animal subjects at risk, or with great cost associated with each observation, such bias may be important. Accurate methods for that setting requires substantial additional research, in progress at this time.

7. CONCLUSIONS

Data analysts often perform power calculations based on random means. For example, a standard function in at least one popular statistical package calculates the sample size needed to obtain a desired power, based on a randomly observed value of the hypothesis sums-of-squares $\left(SSH(\hat{\theta})\right)$. A naive practitioner may ignore possible censoring mechanisms and inadvertently assume this to be the sample size required for a subsequent trial to have a desired power. At other times one simply seeks to replicate an initial, significant result. Here an observed value from the first trial may be the most informative choice of mean difference for a sample size calculation. In either instance the censoring mechanism may importantly affect study planning. Consider the example in §5. Ignoring left censoring generated optimistic bias, with inaccuracy greatest when f_s fell near $f_{crit}(1 - \alpha_s)$. Values of f_s in this neighborhood leave the researcher least certain of results, and therefore most in need of accurate power estimation.

Not surprisingly the obvious estimates and 50% confidence bounds for power, ignoring censoring, were very inaccurate in our simulations, emphasizing the potential for bias. Inaccuracy due to censoring can be eliminated by considering truncated noncentral F distributions. The mle's derived from such distributions may still have substantial bias. In contrast, unbiased bounds for power and N may be stated with desired levels of confidence. We prefer such values, whether or not censoring occurs.

Confidence bounds for ω sometimes must be truncated to zero, which precludes defining an estimate of N necessary to achieve a target power in a subsequent trial. This can happen even if the hypothesis test in an initial study was significant (see Table I). Attempting to estimate large confidence bounds reduces the risk of truncation. In fact, a range of bounds exists which guarantees $\hat{\omega}_q > 0$. However, zero estimates of the 50% or larger bound should warn of little support for a confirmatory trial. In contrast, 50% confidence bounds for ω obtained from a study which failed to achieve significance need not always equal zero (see Table II). Hence useful information for planning a second trial may still be obtained from the initial study.

The new methods do not apply to all GLUM power calculations. In some instances a scientist can not objectively select an appropriate α_s . At other times interest centers on a fixed mean difference, or a range of differences. Even when the methods do not apply directly, the results illustrate the potential for serious bias in study planning, and suggest solutions.

Acknowledgments

Taylor's work was supported in part by NIEHS Training Grant 5-T32-ES07018 and NCI program project grant PO1 CA47 982-04. Muller's work was supported in part by NICHD Mental Retardation Research Center Grant 2-P30-HD03110-21, NCI program project grant PO1 CA47 982-04, and NIH Clinical Research Center Grant RR00046. The authors gratefully acknowledge helpful comments of anonymous reviewers on earlier versions of this paper.

APPENDIX

The power software described in Muller, LaVange, Ramey, and Ramey (Appendix A, 1992) resides at <ftp://ftp.uga.edu/pub/sas/contrib/cntb0014>. To get the files via FTP, connect to HOST <ftp.uga.edu>, as USER *anonymous*, with PASSWORD your_email_address. Change the remote directory to /pub/sas/contrib/cntb0014 and get all seven files (example1-3.sas, manual2.doc, powerlib.sas, README, and tables.sas). Although FTP is preferred, one can obtain the software by sending E-mail with SUBJECT *cntb0014: download* to USERID sascontrib@Sasserv.uga.edu. The message body is ignored. The user will receive several E-mails in return.

BIBLIOGRAPHY

- Abramowitz, M.; Stegun, IA. Handbook of Mathematical Functions. Dover Publications; New York: 1964.
- Benignus VA, Muller KE, Barton CN, Prah JD. Effect of low level carbon monoxide on compensatory tracking and event monitoring. *Neurotoxicology and Teratology*. 1987; 9:227–234. [PubMed: 3627086]
- Cohen, J. Statistical Power Analysis for the Behavioral Sciences. revised edition. Academic Press; New York: 1977.
- Dwass M. A note on simultaneous confidence intervals. *Annals of Mathematical Statistics*. 1955; 26:146–147.
- Guirguis GH. A note on computing the noncentrality parameter of the noncentral F-distribution. *Communications in Statistics (B)*. 1990; 19:1497–1511.
- Hardison CD, Langston RD, Quade D. SAS functions to compute chi-squared, F, and T distributions and their inverses allowing for noncentrality and fractional degrees of freedom. *SUGI Supplemental Library User's Guide - Version*. 1980; 5:413–416.
- Johnson, NL.; Kotz, S. Continuous Univariate Distributions. Vol. Vol. 2. Houghton Mifflin; Boston: 1970.
- Lam, Y. Confidence limits for noncentrality parameters of non central chi-square and F distributions. *Proceedings of the Statistical Computing Section, American Statistical Association Annual Meetings*; 1987. p. 441–443.
- Lattes V, Merigan W. Behavioral effects of carbon monoxide on animals and man. *Annual Review of Pharmacology and Toxicology*. 1979; 19:357–392.
- Lenth RV. Computing noncentral beta probabilities. *Applied Statistics*. 1987; 36:241–244.
- Muller KE, LaVange LM, Ramey SL, Ramey CT. Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association*. 1992; 87:1209–1226.
- Odeh, RE.; Fox, M. Sample Size Choice. Marcel Dekker; New York: 1991.

- Posten HO. An effective algorithm for the noncentral chi-squared distribution function. *The American Statistician*. 1989; 43:261–263.
- Putz, VR.; Johnson; Setzer, JV. Effects of CO on vigilance performance. U. S. Dept. of Health, Education and Welfare, National Institute for Occupational Safety and Health; Cincinnati, Ohio: 1976.
- SAS Institute. JMP© User's Guide. Version 2. SAS Institute; Cary, N. C.: 1989.
- Searle, S. *Linear Models*. Wiley; New York: 1971.
- Taylor DJ, Muller KE. Computing confidence bounds for power and sample size of the general linear univariate model. *The American Statistician*. 1995; 49:43–47. [PubMed: 24039272]
- Venables W. Calculation of confidence intervals for noncentrality parameters. *Journal of the Royal Statistical Society, B*. 1975; 37:406–412.
- Wright, S.; O'Brien, RG. Power analysis in an enhanced GLM procedure: what it might look like. *SAS Users Group International Conference Proceedings*; 1988. p. 1097-1100.

TABLE I

Power and Confidence Bound Mean Estimates Across Conditions ($\times 100$), Ignoring and Adjusting for Left Censoring (all standard errors = 1)

P	α_s	<u>Ignoring</u>		<u>Adjusting</u>				
		$\hat{P}_0$	$\hat{P}_{.5}$	$\hat{P}_{L, mle}$	$\hat{P}_{L, .5}$	$(\hat{P} = \alpha_t)$	$\hat{P}_{L, .05}$	$(\hat{P} = \alpha_t)$
10	.05	74	69	24	31	(42%)	6	(92%)
	.20	53	45	21	26	(41%)	6	(91%)
50	.05	81	76	43	49	(25%)	11	(77%)
	.20	67	62	46	49	(18%)	12	(70%)
90	.05	90	87	71	74	(9%)	26	(48%)
	.20	86	83	78	79	(3%)	31	(30%)

TABLE II

Power and Confidence Bound Mean Estimates Across Conditions ($\times 100$), Ignoring and Adjusting for Right Censoring (standard error = 1, but 2 for *)

P	α_s	<u>Ignoring Censoring</u>		<u>Adjusting for Censoring</u>			
		$\hat{P}_0$	$\hat{P}_{.5}$	$\hat{P}_{R,.5}$	$(\hat{P} = \alpha_t)$	$\hat{P}_{R,.05}$	$(\hat{P} = \alpha_t)$
10	.05	21	14	25	(42%)	6	(92%)
	.20	15	8	31	(44%)	7	(93%)
50	.05	34	25	53	(12%)	11	(73%)
	.20	19	11	52	(22%)	10	(84%)
90	.05	43	35	75	(2%)	19	(48%)
	.20	23	14	69	(6%)	17*	(73%)

TABLE III

Power Estimates and 90% Confidence Lower Bounds ($\times 100$) for Compensatory Tracking Data (Putz, Johnson and Setzer; 1976), Ignoring and Adjusting for Left Censoring ($\alpha_s = .05$, $\alpha_t = .05$, $v_2 = 81$)

Source	v_1	f_s	p -value	Estimates		Lower Bounds	
				$\hat{P}_{.5}$	$\hat{P}_{L, .5}$ Not Adjusted	Adjusted	Adjusted
CO $\times$ Fr	2	3.82	.035	60.8	5.0	15.8	5.0
CO $\times$ Hr	6	10.16	<.001	>99.9	>99.9	99.6	99.6
CO $\times$ Fr $\times$ Hr	6	5.78	<.001	98.8	98.7	82.1	78.3