Published in final edited form as:
Commun Stat Theory Methods. 2000 January ; 29(12): . doi:10.1080/03610920008832631.

Some Distributions and Their Implications for an Internal Pilot Study With a Univariate Linear Model

Christopher S. Coffey and
A-1124 Medical Center North, Dept. of Preventive Medicine, Vanderbilt Univ. School of Med., Nashville, Tennessee 37232-2637

Keith E. Muller
3105C McGavran-Greenberg, Dept. of Biostatistics, CB#7400, University of North Carolina, Chapel Hill, North Carolina 27599

Abstract

In planning a study, the choice of sample size may depend on a variance value based on speculation or obtained from an earlier study. Scientists may wish to use an internal pilot design to protect themselves against an incorrect choice of variance. Such a design involves collecting a portion of the originally planned sample and using it to produce a new variance estimate. This leads to a new power analysis and increasing or decreasing sample size. For any general linear univariate model, with fixed predictors and Gaussian errors, we prove that the uncorrected fixed sample F-statistic is the likelihood ratio test statistic. However, the statistic does not follow an F distribution. Ignoring the discrepancy may inflate test size. We derive and evaluate properties of the components of the likelihood ratio test statistic in order to characterize and quantify the bias. Most notably, the fixed sample size variance estimate becomes biased downward. The bias may inflate test size for any hypothesis test, even if the parameter being tested was not involved in the sample size re-estimation. Furthermore, using fixed sample size methods may create biased confidence intervals for secondary parameters and the variance estimate.

Key Words and Phrases

Interim power analysis; sample size re-estimation

1. Introduction

1.1 Motivation and Literature Review

In designing a study, researchers want to collect a sample large enough to detect a specified effect for a given test size (α_t) and target power (P_t). Scientists often rely on an educated guess or variance estimate of uncertain validity to conduct a power analysis and choose a sample size. Wittes and Brittain (1990) introduced the concept of an internal pilot study for the two sample *t*-test, in which some fraction of the planned observations are used to re-estimate error variance but not the effect of interest. Using the new variance estimate in a fixed sample power calculation then modifies the final sample size. Wittes and Brittain suggested ignoring the randomness of the final sample size for testing.

Coffey and Muller (1999) extended the idea to any General Linear Univariate Model (GLUM) with fixed predictors and Gaussian errors. They derived an exact algorithm for computing test size and power of the primary hypothesis. They also illustrated the strong dependence of test size inflation on interactions among a number of study features.

Many important questions remain unanswered. Carefully evaluating the analytic properties of the approach will greatly help in determining the impact of using an internal pilot design. In particular, 1) detailed knowledge of analytic properties of the random variables in the test statistic would allow characterizing the inflation. 2) Additional results are needed for the general GLUM setting to allow testing secondary hypotheses other than the one upon which sample size re-estimation was based. 3) The ability to provide a defensible confidence interval for the variance observed in a study would aid researchers planning similar studies in the future.

1.2 Notation

Indicate the cumulative distribution function (CDF) of a random variable, U, with parameters α_1 to α_k, as $F_U(u; \alpha_1 \ldots \alpha_k)$, with pth quantile $F_U^{-1}(p;\alpha_1 \ldots \alpha_k)$, and density $f_U(u; \alpha_1 \ldots \alpha_k)$. Let $\chi^2(\nu, \omega)$ indicate a noncentral χ^2, $\chi^2(\nu)$ a central χ^2, and $F(\nu_1, \nu_2, \omega)$ a noncentral F variable (Johnson, Kotz, and Balakrishnan, 1995, Chapters 29 and 30). Also, let $\chi_T^2(\nu;l,u)$ indicate a doubly truncated central χ^2 with lower truncation point l and upper truncation point u (Coffey and Muller, 2000).

We consider the same model as in Coffey and Muller (1999), which includes the two sample t-test as a special case. For a specified design, write a GLUM with fixed predictors and Gaussian errors as

$$\boldsymbol{y}_+ = \boldsymbol{X}_+\beta + \boldsymbol{e}_+$$

$$\begin{bmatrix} \boldsymbol{y}_1 \\ n_1 \times 1 \\ \boldsymbol{y}_2 \\ N_2 \times 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 \\ n_1 \times q \\ \boldsymbol{X}_2 \\ N_2 \times q \end{bmatrix} \beta + \begin{bmatrix} \boldsymbol{e}_1 \\ n_1 \times 1 \\ \boldsymbol{e}_2 \\ N_2 \times 1 \end{bmatrix}, \quad (1)$$

with partitioning corresponding to the internal pilot and second samples. Table 1 contains four categories of notation: 1) design parameters, which are properties required for any sample size calculation, 2) sample size allocation rules, which determine the size of the internal pilot sample and limit the final sample size, 3) unknown fixed parameters, and 4) random variables to be observed. Let $\mathrm{Es}(\boldsymbol{X}_j)$ represent the matrix created by deleting any duplicate rows from $\boldsymbol{X}_j$ (Helms, 1988). We require that $\mathrm{Es}(\boldsymbol{X}_1) = \mathrm{Es}(\boldsymbol{X}_2)$, with the possible exception that a "block" effect may be added, which indicates whether the observation was collected in the internal pilot or second sample. We also impose the restriction that all possible observed samples differ by a multiple of a fixed number of observations, m. It follows that n_1, N_2, and N_+ will be multiples of m. For example, consider increasing sample size by always taking two control subjects for each experimental one ($m = 3$).

In testing H_0: $\theta = \theta_0$, with $\theta = \boldsymbol{C}\beta$, we assume $\boldsymbol{C}$ to be an $a \times q$ matrix with $a = \mathrm{rank}(\boldsymbol{C})$. Without loss of generality assume $\theta_0 = 0$ (Coffey and Muller, 1999). The unadjusted testing method computes $F(n_+)$, the fixed sample size F statistic and rejects H_0 if $F(n_+) > f\mathrm{F} = F_F^{-1}(1 - \alpha_t, a, \nu_+)$.

1.3 Known Results

Wittes and Brittain (1990) considered using an internal pilot design with no adjustment to testing. A fixed sample power calculation determines the random N_+ as a function of $\hat{\sigma}_1^2$. They used simulations to evaluate test size, power, and expected sample size for a t-test involving roughly 100 total observations.

Wittes, Schabenberger, Zucker, Brittain, and Proschan (1999) derived exact test size in this setting. They also showed that $\hat{\sigma}^2(N_+)$ is biased downward, but did not provide an expression for the bias.

Coffey and Muller (1999) provided a number of exact results for the more general GLUM setting. For a specified value of n_+, define $\omega_t(n_+)$ to be the solution to $P_t = 1 - F_F[f_F; a, \nu_+, \omega_t(n_+)]$. Hence

$$\sigma^2(n_+)=\frac{\theta_*'\left[\boldsymbol{C}(\boldsymbol{X}_+'\boldsymbol{X}_+)^-\boldsymbol{C}'\right]^{-1}\theta_*}{\omega_t(n_+)} \quad (2)$$

equals the largest value of $\hat{\sigma}_1^2$ which leads to a final sample size of n_+ or smaller. Also define

$$\begin{aligned} q(n_+,\gamma)&=\frac{\nu_1\sigma^2(n_+)}{\sigma^2}\\ &=\frac{\nu_1\sigma^2(n_+)}{\gamma\sigma_0^2}, \end{aligned} \quad (3)$$

which equals the largest value of SSE_1/σ^2 leading to a sample size of n_+ or smaller. Hence the probability of a particular random final sample size is

$$\begin{aligned} \Pr\{N_+=n_+\}&=\Pr\{\sigma^2(n_+-m)<\hat{\sigma}_1^2<\sigma^2(n_+)\}\\ &=\Pr\{q(n_+-m,\gamma)<SSE_1/\sigma^2<q(n_+,\gamma)\}\\ &=F_{\chi^2}[q(n_+,\gamma);\nu_1]-F_{\chi^2}[q(n_+-m,\gamma);\nu_1]. \end{aligned} \quad (4)$$

Note that $F_{\chi^2}[q(n_{+,\min}-m,\gamma);\nu_1]=0$ and $F_{\chi^2}[q(n_{+,\max},\gamma);\nu_1]=1$. In theory, $n_{+,\max}$ may be infinite. However, budgetary and time constraints often restrict $n_{+,\max}$ to some small multiple of n_0. Coffey and Muller (1999) used a double conditioning argument to describe an algorithm for computing the power of the unadjusted test, for any θ.

$$P(\gamma,\theta)=1-\sum_{n_+=n_{+,\min}}^{n_{+,\max}}\int_{q(n_+-m,\gamma)}^{q(n_+,\gamma)}\Pr\left\{\left(\frac{\nu_+}{f}\,\mathrm{F}a\chi^2[a,\omega(n_+,\gamma,\theta)]-\chi^2(n_2)\le t\right)f_{\chi^2}(t;\nu_1)\,dt, \quad (5)$$

with

$$\omega(n_+,\gamma,\theta)=\frac{\theta'[\boldsymbol{C}(\boldsymbol{X}_+'\boldsymbol{X}_+)^-\boldsymbol{C}']^{-1}\theta}{\sigma^2}. \quad (6)$$

In practice, the results of Coffey and Muller (1999) and the new results in this paper do not require determining N_+ with a fixed sample calculation. The rule for choosing sample size need only determine $\{\sigma^2(n_+)\}$ in a way that maps regions of $\hat{\sigma}_1^2$ into values of N_+. Changing the rule merely changes the corresponding truncation regions.

2. New Analytic Results

2.1 Error Bound for Power and Test Size Algorithm

Even without a finite upper limit on N_+ (allowing $n_{+,\max}=\infty$), practical computations require truncating the distribution of N_+ at a value beyond which the probability of a sample

size more extreme is negligible. Indicate the error due to this truncation by $P_E(\gamma, \theta)$. Let N_L be the lower truncation point, *i. e.*, the largest value of N_+ such that $F_{\chi^2}[q(n_+,\gamma);\nu_1]<\epsilon$. Let N_U be the upper truncation point, *i. e.*, the smallest value of N_+ such that $1 - F_{\chi^2}[q(n_+,\gamma);\nu_1]<\epsilon$. Truncating at N_L and N_U leads to an error of

$$P_E(\gamma,\theta)=\sum_{n_+=n_{+,\min}}^{N_L}\int_{q(n_+-m,\gamma)}^{q(n_+,\gamma)}F_Q(t;n_+,\gamma,\theta)f_{\chi^2}(t;\nu_1)dt+\sum_{n_+=N_U}^{n_{+,\max}}\int_{q(n_+-m,\gamma)}^{q(n_+,\gamma)}F_Q(t;n_+,\gamma,\theta)f_{\chi^2}(t;\nu_1)dt, \quad (7)$$

with

$$F_Q(t;n_+,\gamma,\theta)=\Pr\left\{\left(\frac{\nu_+}{f}\right.\right. \mathrm{Fa}\chi^2[a,\omega(n_+,\gamma,\theta)]-\chi^2(n_2)\le t. \quad (8)$$

Replacing $F_Q(\cdot)$ with 1 gives an upper bound on the error in the last equation:

$$P_E(\gamma,\theta)\le\sum_{n_+=n_{+,\min}}^{N_L}\int_{q(n_+-m,\gamma)}^{q(n_+,\gamma)}f_{\chi^2}(t;\nu_1)dt+\sum_{n_+=N_U}^{n_{+,\max}}\int_{q(n_+-m,\gamma)}^{q(n_+,\gamma)}f_{\chi^2}(t;\nu_1)dt\le F_{\chi^2}(N_L;\nu_1)+[1-F_{\chi^2}(N_U;\nu_1)]\le 2\epsilon. \quad (9)$$

If only one end requires truncation the bound reduces to ϵ.

2.2 The Likelihood and Related Properties

Using an internal pilot study causes N_+ to be random. In a Bayesian framework, the stopping rule does not affect inference and the likelihood principle is not affected by this randomness (Jennison and Turnbull, 2000, p.338). Thus it seems intuitive that the maximum likelihood estimates and likelihood ratio test statistic for internal pilot and fixed sample designs coincide. Nevertheless, our interest in a wide range of scenarios compelled us to provide a formal proof.

Use conditioning arguments to write the likelihood for the GLUM as

$$\mathscr{L}(\beta,\sigma^2;\boldsymbol{y}_1,\boldsymbol{y}_2,N_+)=\mathscr{L}(\beta,\sigma^2;\boldsymbol{y}_1)\cdot\mathscr{L}(\beta,\sigma^2;\boldsymbol{y}_2|N_+,\boldsymbol{y}_1)\cdot\mathscr{L}(\beta,\sigma^2;N_+|\boldsymbol{y}_1) \quad (10)$$

The marginal likelihood of the first sample, $\mathscr{L}(\boldsymbol{\beta}, \sigma^2; \mathbf{y}_1)$ equals that of a random sample of n_1 observations from a Gaussian population. The likelihood of the second sample conditional upon N_+, $\mathscr{L}(\boldsymbol{\beta}, \sigma^2; \mathbf{y}_2|N_+, \mathbf{y}_1)$, equals that of a random sample of n_2 observations from a Gaussian population. Hence the total likelihood differs from that for a fixed sample with $n_+ = n_1 + n_2$ observations only through $\mathscr{L}(\boldsymbol{\beta}, \sigma^2; N_+ \mid \mathbf{y}_1)$. Observe that N_+ is discrete and write

$$\begin{aligned}\mathscr{L}(\beta,\sigma^2;N_+|\boldsymbol{y}_1)&=\Pr\{N_+=n_+|\boldsymbol{y}_1\}\\&=\begin{cases}1, & \text{if } q(n_+-m,\gamma)<SSE_1/\sigma^2\le q(n_+,\gamma)\\0, & \text{otherwise}\end{cases}\\&=I\{q(n_+-m;\gamma)<SSE_1/\sigma^2\le q(n_+;\gamma)\},\end{aligned} \quad (11)$$

in which $I(\cdot)$ represents an indicator function with a value of 1 if the expression is true. In turn, write the joint likelihood under an internal pilot design as

$$\mathscr{L}(\beta,\sigma^2;\boldsymbol{y}_1,\boldsymbol{y}_2,N_+)=I\{q(n_+-m,\gamma)<SSE_1/\sigma^2\le q(n_+,\gamma)\}\times(2\pi\sigma^2)^{-n_+/2}\exp[-(\boldsymbol{y}_+-\boldsymbol{X}_+\beta)'(\boldsymbol{y}_+-\boldsymbol{X}_+\beta)/2\sigma^2]. \quad (12)$$

Since the value of the indicator function does not depend on any unknown parameter, we may ignore it in estimation. Therefore the sufficient statistics and maximum likelihood estimates for σ^2, β, and θ, coincide with those from a fixed sample size analysis. Let $\hat{\sigma}^2(n_+)$ represent the maximum likelihood estimate of σ^2. In a fixed sample design we often prefer the unbiased estimate,

$$\tilde{\sigma}^2(n_+)=\left(\frac{n_+}{\nu_+}\right)\hat{\sigma}^2(n_+). \quad (13)$$

However, both $\hat{\sigma}^2(n_+)$ and $\hat{\sigma}^2(N_+)$ have bias, as detailed in §2.5.

Coincidence of the maximum likelihood estimates implies the coincidence of the likelihood test statistics. However, $F(n_+)$ will not follow an F distribution under an internal pilot design. In order to characterize the distribution of $F(n_+)$, we examine each component separately in the sections which follow.

2.3 The Distribution of $SSH(n_+)$

Conditional on $N_+ = n_+$, the numerator of the likelihood ratio test statistic has the same distribution as for a fixed sample design. To see this, observe that $\hat{\beta}(n_+)$ equals a linear combination of $\hat{\beta}_1$ and $\hat{\beta}_2(n_+)$, the independent estimates from the internal pilot and additional samples. The randomness of N_+ depends only on $\hat{\sigma}_1^2$ which is independent of $\hat{\beta}_1$, $\hat{\beta}_2(n_+)$, and, in turn, $\hat{\beta}(n_+)$. Hence the conditional distributions of $\hat{\beta}(n_+)$, $\hat{\theta}(n_+)$, and $SSH(n_+)$ coincide with the corresponding distributions for a fixed sample size design with n_+ observations. Incidentally, if Es($\boldsymbol{X}_+$) has full rank then $\hat{\beta}(n_+)$ will be unbiased.

Unconditionally, the numerator of the likelihood ratio test statistic for an internal pilot design has the same distribution as for a fixed sample design under the null, but not under the alternative. To see this, observe that if $\theta\neq\mathbf{0}$ then the CDF of $SSH(N_+)/\sigma^2$ equals a weighted sum of the conditional CDF's, with the weights corresponding to the probability of observing a specific n_+:

$$\Pr\{SSH(N_+)/\sigma^2<s\}=\sum_{n_+=n_{+,\min}}^{n_{+,\max}}F_{\chi^2}[s;a,\omega(n_+,\gamma,\theta)]\Pr\{N_+=n_+\}. \quad (14)$$

Under the null, $\omega(n_+, \gamma, \mathbf{0}) = 0$, the conditional distribution does not depend on n_+, and $SSH(N_+)/\sigma^2 \sim \chi^2(a)$. Hence any effect on *test size* due to using an internal pilot does not depend on the numerator of the test statistic.

2.4 The Distribution of $SSE(n_+)$

With a fixed n_+, $SSE(n_+)/\sigma^2 \sim \chi^2(\nu_+)$. However, with an internal pilot, the dependence of n_+ on $\hat{\sigma}_1^2$ complicates the distribution. Following Coffey and Muller (1999), write $SSE(n_+)$ as the sum of two independent quadratic forms:

$$SSE(n_+)=\boldsymbol{y}_+'\boldsymbol{A}_e\boldsymbol{y}_+=\boldsymbol{y}_+'(\boldsymbol{A}_e-\boldsymbol{A}_1)\boldsymbol{y}_+ +\boldsymbol{y}_+'\boldsymbol{A}_1\boldsymbol{y}_+=SSE_*(n_+)+SSE_1(n_+), \quad (15)$$

with $SSE_1(n_+)$ the error sum of squares from the internal pilot sample. Coffey and Muller (1999) proved that $SSE_*(n_+)/\sigma^2 \sim \chi^2(n_2)$. However, conditional upon observing a specific n_+, SSE_1 is restricted to the possible range of values which would have led to that final sample size. Therefore

$$SSE_1(n_+)/\sigma^2 \sim \chi^2_T[\nu_1, q(n_+ - m;\gamma), q(n_+;\gamma)]. \quad (16)$$

The characteristic function of $SSE(n_+)/\sigma^2$ has a simple form. Define

$$D(t) = F_{\chi^2}[q(n_+;\gamma)(1-2t);\nu_1] - F_{\chi^2}[q(n_+ - m;\gamma)(1-2t);\nu_1], \quad (17)$$

and assume $t \in [0, 1/2)$. A result in Coffey and Muller (2000) allows writing the characteristic function of $SSE_1(n_+)/\sigma^2$ as

$$\phi_{SSE_1(n_+)/\sigma^2}(t) = \left[\frac{D(it)}{D(0)}\right] \phi_{\chi^2}(t;\nu_1). \quad (18)$$

In turn, the independence of $SSE_1(n_+)$ and $SSE_*(n_+)$ implies

$$\begin{aligned} \phi_{SSE(n_+)/\sigma^2}(t) &= \left\{\left[\frac{D(it)}{D(0)}\right] \phi_{\chi^2}(t;\nu_1)\right\} \phi_{\chi^2}(t;n_2) \\ &= \left[\frac{D(it)}{D(0)}\right] \phi_{\chi^2}(t;\nu_+). \end{aligned} \quad (19)$$

Thus the characteristic function of $SSE(n_+)/\sigma^2$ under an internal pilot design equals the product of the characteristic function under a fixed design and a factor which accounts for truncation. Ignoring the randomness of N_+ and approximating the distribution with the fixed sample result ignores the factor.

The CDF of $SSE(n_+)/\sigma^2$ may be computed by inverting the characteristic function. Alternatively, a method similar to the one used by Coffey and Muller (1999) for the CDF of the test statistic may be used. Condition on $SSE_1(n_+)/\sigma^2$ and integrate numerically over its range of values:

$$\begin{aligned} &F_{SSE(n_+)/\sigma^2}(s;\gamma) \\ &= \int_{q(n_+-m,\gamma)}^{q(n_+,\gamma)} \Pr\{SSE_*(n_+)/\sigma^2 + t \le s\} \frac{f_{\chi^2}(t;\nu_1)}{\Pr\{N_+ = n_+\}} dt \\ &= \frac{1}{\Pr\{N_+ = n_+\}} \int_{q(n_+-m,\gamma)}^{q(n_+,\gamma)} F_{\chi^2}(s-t;n_2) f_{\chi^2}(t;\nu_1) dt. \end{aligned} \quad (20)$$

Compute the unconditional CDF via the law of total probability:

$$F_{SSE(N_+)/\sigma^2}(s;\gamma) = \sum_{n_+ = n_{+,\min}}^{n_{+,\max}} \Pr\{N_+ = n_+\} F_{SSE(n_+)/\sigma^2}(s;\gamma) = \sum_{n_+ = n_{+,\min}}^{n_{+,\max}} \int_{q(n_+-m,\gamma)}^{q(n_+,\gamma)} F_{\chi^2}(s-t;n_2) f_{\chi^2}(t;\nu_1) dt. \quad (21)$$

Truncating the sum will lead to the same size error as in §2.1 for computing the CDF of the test statistic (error < 2ϵ, with ϵ chosen as the truncation value).

2.5 The Bias of $\hat{\sigma}^2(n_+)$

Wittes, *et al*. (1999) showed that $\varepsilon[\hat{\sigma}^2(N_+)] \leq \sigma^2$ (for the *t*-test), but did not provide an expression for the bias. The importance of the bias arises from the fact that test size inflation varies directly with it. Using the standard result for the moment generating function of a linear transformation of a random variable,

$$M_{\hat{\sigma}^2(n_+)/\sigma^2}(t) = \left[\frac{D(t/\nu_+)}{D(0)}\right] M_{\chi^2(\nu_+)}(t/\nu_+). \quad (22)$$

Taking the first derivative and setting $t = 0$ gives the conditional bias:

$$\varepsilon\left[\frac{\hat{\sigma}^2(n_+)}{\sigma^2}\right] = 1 - \frac{2\{q(n_+,\gamma) f_{\chi^2}[q(n_+,\gamma);\nu_1] - q(n_+ - m,\gamma) f_{\chi^2}[q(n_+ - m,\gamma);\nu_1]\}}{\nu_+ \Pr\{N_+ = n_+\}}. \quad (23)$$

Applying Lemma 1 from Coffey and Muller (2000),

$$\varepsilon\left[\frac{\hat{\sigma}^2(n_+)}{\sigma^2}\right] = 1 - \frac{2\nu_1\{f_{\chi^2}[q(n_+,\gamma);\nu_1+2] - f_{\chi^2}[q(n_+ - m,\gamma);\nu_1+2]\}}{\nu_+ \Pr\{N_+ = n_+\}}. \quad (24)$$

Recall that $\chi^2(\nu_1 + 2)$ has a single mode at ν_1 (for $\nu_1 \geq 2$, as always true here). Hence $\hat{\sigma}^2(n_+)$ is (conditionally) biased in a direction depending on whether $\hat{\sigma}_1^2$ is greater than or less than σ^2. The law of total probability allows obtaining the *unconditional* expectation:

$$\varepsilon\left[\frac{\hat{\sigma}^2(N_+)}{\sigma^2}\right] = \sum_{n_+ = n_{+,\min}}^{n_{+,\max}} \varepsilon[\hat{\sigma}^2(n_+)/\sigma^2] \Pr\{N_+ = n_+\} = 1 - 2\nu_1 \sum_{n_+ = n_{+,\min}}^{n_{+,\max}} \frac{f_{\chi^2}[q(n_+,\gamma);\nu_1+2] - f_{\chi^2}[q(n_+ - m,\gamma);\nu_1+2]}{\nu_+}. \quad (25)$$

Factoring like terms and finding common denominators leads to

$$\varepsilon\left[\frac{\hat{\sigma}^2(N_+)}{\sigma^2}\right] = 1 - \sum_{n_+ = n_{+,\min}}^{n_{+,\max}} \left(\frac{\nu_1}{\nu_+}\right)\left(\frac{2m}{\nu_+ + m}\right) f_{\chi^2}[q(n_+,\gamma);\nu_1+2]. \quad (26)$$

Hence $\varepsilon[\hat{\sigma}^2(N_+)] \leq \sigma^2$. Furthermore, from (26) it is clear that large values of either n_1 or n_+ insure negligible bias. However, the dependence of N_+ on many of the parameters in Table 1 complicate any discussion of large sample properties. In general, any combination of parameters which leads to large values of n_1 or N_+ will reduce bias in $\hat{\sigma}^2(N_+)$. Finally, observing that $f_{\chi^2}(t;\ \nu_1 + 2)$ has a single mode at $t = \nu_1$, allows bounding the unconditional bias:

$$0 \leq 1 - \varepsilon\left[\frac{\hat{\sigma}^2(N_+)}{\sigma^2}\right] \leq f_{\chi^2}(\nu_1;\nu_1+2) \sum_{n_+ = n_{+,\min}}^{n_{+,\max}} \left(\frac{\nu_1}{\nu_+}\right)\left(\frac{2m}{\nu_+ + m}\right). \quad (27)$$

The origin of the bias may be characterized further. Obviously $\varepsilon(\hat{\sigma}_1^2) = \sigma^2$. Define $\hat{\sigma}_*^2(n_+) = SSE_*(n_+)/n_2$ and note that

$$\varepsilon[\hat{\sigma}_*^2(N_+)]=\sum_{n_+}\Pr\{N_+=n_+\}\varepsilon[SSE_*(n_+)/n_2]=\sigma^2. \quad (28)$$

Therefore $\sigma^2(N_+)$ equals a linear function of two unbiased estimators:

$$\hat{\sigma}^2(N_+)=\left(\frac{\nu_1}{\nu_+}\right)\hat{\sigma}_1^2+\left(\frac{n_2}{\nu_+}\right)\hat{\sigma}_*^2(N_+). \quad (29)$$

This form has three important implications. First, the randomness of the weights creates bias. Second, as n_2/n_1 increases, the second term in (29) dominates. Third, although neither of the two unbiased estimates uses all of the data, any combination of the estimates using fixed, positive weights that sum to one would create an unbiased estimate of σ^2 that uses all of the data.

2.6 Characteristic Function of a 1-1 Function of the Test Statistic

Let $F_{F(n+)}(f;\gamma,\theta)$ represent the conditional CDF of the internal pilot test statistic computed at an arbitrary point f. For example, letting $f = f_F$ allows computing the conditional power under an internal pilot design using the unadjusted approach for testing. The results in §2.3 and §2.4 imply that $F(n_+)$ differs from $F(f; a, \nu_+)$ only through the denominator. With $c(n_+, f) = \nu_+/fa$, express $F_{F(n+)}(f;\gamma,\theta)$ as

$$\begin{aligned}F_{F(n_+)}(f;\gamma;\theta)&=\Pr\left\{\frac{SSH(n_+)/a}{SSE(n_+)/\nu_+}\le f\right\}\\&=\Pr\{c(n_+,f)SSH(n_+)/\sigma^2-SSE_*(n_+)/\sigma^2-SSE_1(n_+)/\sigma^2\le 0\}\\&=\Pr\{c(n_+,f)\chi^2[a,\omega(n_+,\gamma,\theta)]-\chi^2(n_2)-\chi_T^2[\nu_1,q(n_+-m;\gamma),q(n_+;\gamma)]\le 0\}\\&=\Pr\{S(n_+,f)\le 0\}=F_{S(n_+,f)}(0;f,\gamma,\theta).\end{aligned} \quad (30)$$

Hence we may compute the CDF of $F(n_+)$ at the point f via the CDF of $S(n_+, f)$ evaluated at zero. Davies' (1980) algorithm inverts the characteristic function to compute the CDF of a weighted sum of χ^2's. Since $S(n_+, f)$ contains a doubly-truncated χ^2, Coffey and Muller (1999) computed the CDF by first conditioning on the doubly-truncated random variable, then applying Davies' algorithm to compute the CDF of the remaining sum of two χ^2's. The approach leads to a double numerical integral. Alternatively, we could directly invert the characteristic function of $S(n_+, f)$, which requires only one integration. Using standard results about characteristic functions gives:

$$\begin{aligned}&\phi_{S(n_+,f)}(t)\\&\quad=\phi_{c(n_+,f)SSH(n_+)/\sigma^2}(t)\phi_{-SSE_*(n_+)/\sigma^2}(t)\phi_{-SSE_1(n_+)/\sigma^2}(t)\\&\quad=\frac{D(-it)}{D(0)}\\&\quad\times\left\{(1+2it)^{-\nu_+/2}[1-2ic(n_+,f)t]^{-a/2}\exp\left[\frac{ic(n_+,f)t\omega(n_+,\gamma,\theta)}{1-2ic(n_+,f)t}\right]\right\}.\end{aligned} \quad (31)$$

The term in braces equals the characteristic function for the weighted sum which would arise in the fixed sample case, with n_+ observations. In turn, the factor $D(-it)/D(0)$ accounts for the truncation of $\hat{\sigma}_1^2$.

3. Numerical Examples

We illustrate the relationship between the bias in estimating the variance and inflation of test size with two examples. Although our results apply far more generally, for simplicity of exposition the examples correspond to the two sample setting. In each case, we assume α_t = 0.05, P_t = 0.90, and no finite bound on the size of the final sample ($n_{+,\max} = \infty$). Example A, described by Wittes and Brittain (1990), centers on detecting a mean difference of $\theta_* = 1$ with $\sigma_0^2 = 2$, n_0 = 86, n_1 = 44, and $n_{+,\min} = n_0$ = 86. Example **B**, described by Coffey and Muller (1999), centers on detecting a mean difference of $\theta_* = 1.6$ with $\sigma_0^2 = 1$, n_0 = 20, n_1 = 10 and $n_{+,\min} = n_1$ = 10. This allows the sample size to be reduced if the variance was originally overstated ($\gamma < 1$).

For $\gamma \in \{0.5, 0.75, 1, 1.5, 2\}$, Table 2 displays $\varepsilon[\hat{\sigma}^2(N_+)/\sigma^2]$ and the true test size for the examples. Bias was computed in SAS IML® with equation 24, while test size was calculated using the algorithm in Coffey and Muller (1999). Note how the amount of test size inflation closely tracks the amount of bias in $\hat{\sigma}^2(N_+)/\sigma^2$. Wittes and Brittain (1990) showed that, in Example A, an internal pilot can provide much better power than a fixed design while providing a negligible increase in test size. Table 2 illustrates that we have little bias in estimating $\hat{\sigma}^2(N_+)/\sigma^2$ as well. However, Example B can lead to test size as large as 0.065 and $\varepsilon[\hat{\sigma}^2(N_+)/\sigma^2]$ as small as 0.89. At least for the highly constrained design in Example A, with moderate to large sample sizes, an internal pilot design causes little worry about test size inflation or bias in estimating σ^2. However, the test size and bias are of much greater concern in small samples.

Coffey and Muller (1999) showed that the degree of test size inflation was highly dependent upon the choice of design parameters and sample size allocation rules. The examples illustrate the correlation in the bias of $\hat{\sigma}^2(N_+)/\sigma^2$ with test size inflation. This implies that there are combinations of design parameters and sample size allocation rules which cause nonignorable bias in $\hat{\sigma}^2(N_+)/\sigma^2$. Hence the possibility of bias should at least be examined before using the variance estimate from an internal pilot study to make inference or plan a future study.

4. Implications of the New Results

4.1 Testing

Any inflation of test size arises solely from a change in the distribution of the variance estimate, rather than a change in the distribution of the parameter estimate itself. In the GLUM setting, a test of any secondary parameter involves the variance estimate. For example, consider testing for a “block” effect in order to insure that there were no differences between the internal pilot and second samples with regards to the outcome. The biased estimate of σ^2 may inflate test size. Hence researchers must be wary of inflation even for hypothesis tests about secondary parameters that were not involved in the sample size re-estimation.

4.2 Confidence Regions for θ

The same complication that biases test size also biases confidence interval coverage. Inverting a test statistic and naively computing a $100(1 - p)\%$ confidence region for θ with standard fixed sample size linear models theory gives coverage less than or equal to the desired level. As with test size, the true coverage depends upon the unknown parameter γ. Furthermore such bias occurs with any secondary parameter.

4.3 Confidence Intervals for σ^2

An appropriate confidence interval for the variance observed in a particular study can be invaluable in planning future studies. Furthermore, in the fixed sample size case, computing confidence intervals for σ^2 allows computing confidence intervals for power and noncentrality (Taylor and Muller, 1996; Muller and Pasour, 1997). However, since $SSE(n_+)/\sigma^2$ does not follow a χ^2 distribution under an internal pilot design, forming a confidence interval for σ^2 using fixed sample size methods may not provide the desired coverage. As with confidence regions for θ, the true coverage depends on γ. However, confidence intervals for σ^2 may have more or less coverage than desired.

5. Conclusions

The following conclusions apply to any univariate linear model with fixed predictors and Gaussian errors.

1. Coffey and Muller's (1999) algorithm for power involves the distribution of N_+. With $n_{+,\max} = \infty$, the calculations require truncation of the distribution. The truncation region can be defined to insure a specified upper bound on error.
2. The likelihood ratio test statistic under an internal pilot design coincides with the statistic for a fixed sample design. However, the statistic does not follow an F-distribution because the variance estimate is not a scaled χ^2.
3. The distributions of $\hat{\theta}(n_+)$ and $SSH(n_+)/\sigma^2$ coincide with those from a fixed sample analysis with n_+ observations.
4. $SSE(n_+)/\sigma^2$ equals the sum of a $\chi^2(n_2)$ and a doubly-truncated $\chi^2(\nu_1)$, in contrast to the fixed sample result of $\chi^2(n_2 + \nu_1)$. This leads to $\mathcal{E}[\hat{\sigma}^2(N_+)] \leq \sigma^2$. Bias in test size and coverage of confidence intervals varies directly with the bias of the final variance estimate.
5. Random predictors (Sampson, 1974), greatly complicate the problem. Our results apply only conditional upon the observed value of (random) $\boldsymbol{X}_+$, at the conclusion of the study. Even with a fixed sample size, only limited results are available for power calculations with random predictors. The introduction of an internal pilot complicates the problem because $\boldsymbol{X}_2$ is random at the re-estimation stage. Clearly, further research is needed for such studies.
6. Fast, accurate approximations for computing power and test size would ease the burden of planning a study with an internal pilot design.
7. Methods for controlling test size merit future research.

Acknowledgments

Coffey's work was supported in part by NIEHS grant 5-T32-ES07018. A portion of the work was submitted by Coffey in partial fulfillment of the requirements for the Ph. D. in Biostatistics. Muller's work was supported in part by NCI program project grant P01 CA47 982-04. The authors wish to thank two anonymous reviewers and an associate editor for a number of helpful suggestions.

Bibliography

Coffey CS, Muller KE. Exact Test Size and Power of a Gaussian Error Linear Model for an Internal Pilot Study. Statistics in Medicine. 1999; 18:1199–1214. [PubMed: 10363340]

Coffey CS, Muller KE. Properties of Doubly-Truncated Gamma Variables. Communications in Statistics - Theory and Methods. 2000; 294 in press.

Davies RB. The Distribution of a Linear Combination of χ^2 Random Variables. Applied Statistics. 1980; 29:323–333.

Helms RW. Comparisons of Parameter and Hypothesis Definitions in a General Linear Model. Communications in Statistics - Theory and Methods. 1988; 17:2725–2753.

Jennison, C.; Turnbull, BW. Group Sequential Methods with Applications to Clinical Trials. Boca Raton: Chapman & Hall/CRC; 2000.

Johnson, NL.; Kotz, S.; Balakrishnan, N. Continuous Univariate Distributions-2. New York: Wiley; 1995.

Muller KE, Pasour VB. Bias in Linear Model Power and Sample Size Due to Estimating Variance. Communications in Statistics - Theory and Methods. 1997; 26:839–851.

Sampson AR. A Tale of Two Regressions. Journal of the American Statistical Association. 1974; 69:682–689.

Taylor DJ, Muller KE. Bias in Linear Model Power and Sample Size Calculations Due to Estimating Noncentrality. Communications in Statistics - Theory and Methods. 1996; 25:1595–1610.

Wittes J, Brittain E. The Role of Internal Pilot Studies in Increasing the Efficiency of Clinical Trials. Statistics in Medicine. 1990; 9:65–72. [PubMed: 2345839]

Wittes J, Schabenberger O, Zucker D, Brittain E, Proschan M. Internal Pilot Studies I: Type I error rate of the naive t-test. Statistics in Medicine. 1999; 18:3481–3491. [PubMed: 10611620]

Table 1

Internal Pilot Study Notation

Symbol	Definition
Design Parameters	
α_t	Target test size
P_t	Target Power
$\boldsymbol{\theta}_*$	"Scientifically Important" value of $\boldsymbol{\theta}$
σ_0^2	Variance value used for planning
n_0	Pre-planned sample size based on α_t, P_t, $\boldsymbol{\theta}_*$, σ_0^2
Sample Size Allocation	
π	Proportion of n_0 used in internal pilot
n_1	Internal pilot sample size; size of first sample, πn_0
ν_1	Internal pilot error degrees of freedom, $n_1 - \text{rank}(\boldsymbol{X}_1)$
$n_{+,\min}$	Minimum size of *final* sample
$n_{+,\max}$	Maximum size of *final* sample
Fixed, Unknown Parameters	
σ^2	True error variance
γ	Ratio of true variance to variance used for planning, σ^2/σ_0^2
$\boldsymbol{\theta}$	True value of secondary parameter, $\boldsymbol{C\beta}$, $a \times 1$ vector
Random Variables	
$\widehat{\sigma}_1^2$	Internal pilot variance estimate, $\boldsymbol{y}_1'[\boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-}\boldsymbol{X}_1']\boldsymbol{y}_1/\nu_1$
N_2	Size of second sample, with particular value n_2
N_+	Final sample size, $n_1 + N_2$, with particular value n_+
ν_+	Final sample error df, $N_+ - \text{rank}(\boldsymbol{X}_+)$
$\widehat{\boldsymbol{\beta}}(n_+)$	$(\boldsymbol{X}_+'\boldsymbol{X}_+)^{-}\boldsymbol{X}_+'\boldsymbol{y}_+$
$\widehat{\boldsymbol{\theta}}(n_+)$	Final estimate of secondary parameter, $\boldsymbol{C}\widehat{\boldsymbol{\beta}}(n_+)$
$SSH(n_+)$	Final hypothesis SS, $\widehat{\theta}(n_+)'[\boldsymbol{C}(\boldsymbol{X}_+'\boldsymbol{X}_+)^{-}C']^{-1}\widehat{\theta}(n_+)$
$\widehat{\sigma}^2(n_+)$	Final variance estimate, $\boldsymbol{y}_+'[\boldsymbol{I} - \boldsymbol{X}_+(\boldsymbol{X}_+'\boldsymbol{X}_+)^{-}\boldsymbol{X}_+']\boldsymbol{y}_+/\nu_+$
$F(n_+)$	Test statistic, $[SSH(n_+)/a]/\widehat{\sigma}^2(n_+)$

Table 2
Bias in Estimating γ and the Relationship with Test Size Inflation

	Example A		Example B	
Gamma	$E[\sigma^2(N_+)/\sigma^2]$	α	$E[\sigma^2(N_+)/\sigma^2]$	α
0.5	1.000	0.050	0.909	0.055
0.75	0.998	0.050	0.891	0.062
1.0	0.990	0.051	0.896	0.065
1.5	0.985	0.052	0.916	0.065
2	0.988	0.052	0.931	0.062