TECHNICAL REPORT # 58
APRIL 2012

Design resampling for interim sample size recalculation

Sergey Tarima, Peng He, Tao Wang, Aniko Szabo
Division of Biostatistics, Medical College of Wisconsin

Abstract

Internal pilot designs allow re-estimation of the sample size at the interim analysis using available information on nuisance parameters. In general, this affects the Type I and II error rates. We propose a method based on resampling the whole design at the interim analysis, starting with sample size recalculation at the observed interim analysis values of nuisance parameters, and finishing with the decision to accept or reject the null hypothesis. This internal resampling is performed under both the null and under the alternative hypotheses allowing the estimation of the bias of the type I error and power. Finally, the bias corrected error rates are used in the original sample size calculation procedure to obtain an updated sample size. We explore the proposed resampling approach under a set of simulation scenarios and compare it with several others previously published internal pilot designs.

KEYWORDS: Internal Pilot; Sample Size; Power Calculation; Hypothesis Testing; Study Design.

1 Introduction

Ethical, financial, and recruitment constraints prevent researchers from enrolling arbitrarily many patients for a study to achieve statistically significant results. Pilot studies are used to provide information on parameters needed to determine an appropriate sample size for a larger confirmatory

for the one sample t-test,

$$D_{1t;IPN}\ (; ; \ _0; \ _1; n_1; n_{max}) (2\ D_2) ;$$

is an alternative to D_{1t}, which does not use $^{(0)}$ but depends on n_1 and n_{max}. Its power function is

$$P(\ jD_{1t;IPN}) = Pr \quad T_{v(; ; \ _0; \ _1;\hat{\ })} > k(v)j ; D_{1t;IPN} \quad ; \tag{2}$$

where ^ depends on , n_1, n_{max} and possibly . In this manuscript we assume that ^ is independent of , that is ^ = ^ .

A naive internal pilot-based sample size recalculation for a two sample t-test will be denoted by $D_{2t;IPN}$. This design was first analyzed by Wittes and Brittain [9]. We also consider the internal pilot design $D_{2t;IPS}$ suggested by Stein [8], which slightly modifies the functional form of the two-sample t-statistic, whereas $D_{2t;IPN}$ uses the classical two sample t-statistic for T_v.

Internal sample size recalculation makes the final sample size a random variable, which makes the distribution of the test statistic T_v and therefore the critical value of the test difficult to calculate. Exact control of the type I error is achieved by $D_{2t;IPS}$, but this is rather an exception than a rule for internal pilot designs. In general, the true type I error rate is rarely controlled,

$$E\ \ (D_{2t;IPN}(; ; \ _0; \ _1; n_1; n_{max}) jH_0) = a(; \ jD_{2t;IPN}) \neq \ :$$

The desired power is not controlled in either Stein's or the naive internal pilot designs,

$$E\ \ (D(; ; \ _0; \ _1; n_1; n_{max}) jH_1) = 1 \quad b(; \ jD) \neq 1 \quad :$$

Sample size recalculation via resampling

We propose a new approach to sample size re-estimation after the internal pilot that maintains both the type I and type II error rates. This approach is applicable to any internal pilot design.

Key idea: For a design $D\ 2\ D_2$ we find $_{new}$ and $_{new}$ to control the desired type I error and power,

$$E\ \ (D(\ _{new}; \ _{new}; \ _0; \ _1; n_1; n_{max}) jH_0) =$$

and

$$E_{\theta}\left(\phi(D(\alpha_{new};\beta_{new};\theta_0;\theta_1;n_1;n_{max}))\,|\,H_1\right) = 1-\beta:$$

This definition leads to a fully defined internal pilot procedure $D^a(\alpha;\beta;\theta_0;\theta_1;n_1;n_{max})$, since all the details about sample size re-estimation, final hypothesis testing, etc are already defined in D.

Implementation: At the interim analysis we estimate $\hat{\sigma}$ and perform the following resampling procedure with M iterations. For each $i = 1;\dots;M$, we generate $Y_1^{(i)};\dots;Y_{n_1}^{(i)}$ from $f_Y(y|\theta_0;\hat{\sigma})$, estimate $v_i \in [n_1;n_{max}]$ based on these n_1 observations, generate additional $(v_i - n_1)$ observations $Y_{n_1+1}^{(i)};\dots;Y_{v_i}^{(i)}$ from $f_Y(y|\theta_0;\hat{\sigma})$, and calculate $T_{v_i}^{(i)}$ on this i^{th} sample. We add the subscript i to highlight dependence on iteration. The estimated type I error rate is

$$\hat{\alpha}(\alpha;\beta|D) = \frac{1}{M}\sum_{i=1}^{M} I\left(T_{v_i}^{(i)} > k_i\right) \ne \alpha;$$

where k_i is the critical value for an originally assumed distribution of $T_{v_i}^{(i)}$. On the logit scale ($\text{logit}(x) = \ln(x/(1-x))$) the bias-corrected α_{new} can be expressed as

$$\text{logit}(\alpha_{new}) = \text{logit}(\alpha) - [\text{logit}(\hat{\alpha}) - \text{logit}(\alpha)]$$

or

$$\alpha_{new} = \frac{\alpha^2(1-\hat{\alpha})}{(1-\alpha)^2\hat{\alpha} + \alpha^2(1-\hat{\alpha})}: \qquad (3)$$

Then, we perform a similar resampling procedure to find β_{new}. For $i = 1;\dots;M$, we generate $Y_1^{(i)};\dots;Y_{n_1}^{(i)}$ from $f_Y(y|\theta_1;\hat{\sigma})$, estimate $v_i \in [n_1;n_{max}]$ on these n_1 observations using α_{new} and β in the sample size formula, generate additional $(v_i - n_1)$ observations $Y_{n_1+1}^{(i)};\dots;Y_{v_i}^{(i)}$ from $f_Y(y|\theta_1;\hat{\sigma})$, and calculate $T_{v_i}^{(i)}$ on this i^{th} sample. The estimated power

$$1-\hat{\beta}(\alpha_{new};\beta|D) = \frac{1}{M}\sum_{i=1}^{M} I\left(T_{v_i}^{(i)} > k_i\right) \ne 1-\beta$$

leads to the bias-corrected value

$$_{\text{new}} = \frac{{}^{2}\ 1 \quad \hat{b}}{(1 \quad)^{2}\,\hat{b} + {}^{2}\ 1 \quad \hat{b} +}$$

Design $D_{1t;IPN}$ (; ; $_0$; $_1$; n_1; n_{max}) does not formally depend on and uses the internally estimated

$$\hat{} =$$

Table 1: Monte-Carlo Type I error, Power, and Sample Sizes; 10,000 simulations; one sample t-test designs, $n_1 = 10$, $n_{max} = 300$.

	D_{1t}	$D_{1t;IPN}$	$D^a_{1t;IPN}$
	Type I error		
1.6	0.0492	0.0643	0.0573
2	0.0500	0.0612	0.0513
3	0.0495	0.0553	0.0473
3.5	0.0494	0.0526	0.0473
	Power		
1.6	0.8177	0.8091	0.8367
2	0.8086	0.7841	0.8216
3	0.8040	0.7601	0.8001
3.5	0.8043	0.7517	0.7943
	EN (SD)		
1.6	23	22.73(9.33)	26.86(12.22)
2	34	33.89(14.80)	40.93(18.01)
3	73	73.17(33.29)	86.68(38.06)
3.5	99	98.53(45.05)	115.89(51.21)

Table 2: Monte-Carlo Type I error, Power, and Sample Sizes; 10,000 simulations; one sample t-test designs, $n_1 = 5$, $n_{max} = 300$.

	D_{1t}	$D_{1t;IPN}$	$D^{a}_{1t;IPN}$
	Type I error		
0.6	0.0501	0.0523	0.0515
1	0.0515	0.0727	0.0682
2	0.0487	0.0685	0.0519
3	0.0503	0.0589	0.0448
3.5	0.0504	0.0574	0.0458
	Power		
0.6	0.8985	0.9387	0.9335
1	0.8030	0.8327	0.8596
2	0.8076	0.7319	0.7897
3	0.8033	0.7057	0.7663
3.5	0.8034	0.6953	0.7560
	EN (SD)		
0.6	6	6.00(1.59)	6.18(2.29)
1	10	10.56(5.39)	13.22(8.51)
2	34	33.88(22.24)	46.87(30.06)
3	73	73.30(49.34)	97.85(61.65)
3.5	99	97.79(64.78)	127.84(76.90)

and

$$Y_{11}; \dots; Y_{n_{11}1}; \dots; Y_{v_1 1}; \dots \quad N(\;_2 + \quad ; \quad ^2_1);$$

where n_{10}, n_{11}, v_0 and v_1 satisfy

$$\frac{n_{10}}{n_{10} + n_{11}} = \frac{n_{10}}{}$$

Table 4: Monte-Carlo Type I error, Power, and Sample Sizes; 10; 000 simulations; two sample t-test designs; $n_1 = 10$ (5 per group); xed allocation, $r = 0:5$

1	D_{2t}	$D_{2t;IPS}$	$D_{2t;IPN}$	$D_{2t;IPNR}$	$D^{a}_{2t;IPN}$
	Type I error				
1	0.0507	0.0508	0.0636	0.0579	0.0526
1.5	0.0496	0.0503	0.0546	0.0537	0.0467
2	0.0499	0.0499	0.0510	0.0509	0.0469
2.5	0.0504	0.0496	0.0515	0.0515	0.0491
	Power				
1	0.8081	0.8140	0.8401	0.8446	0.8213
1.5	0.8093	0.8077	0.8261	0.8259	0.7995
2	0.8010	0.8030	0.8184	0.8183	0.7883
2.5					

t-distribution. However random allocation of subjects to groups leads to a di erent distribution. Since only the noncentrality parameter dependens on v_1 and v_2, the distribution under H_0 does not change, but under H_1 it becomes a mixture with

$$P(|T_v| > k \mid v \quad 2; \ _1; \ _0; \ _3) = \sum_{v_1=0}^{v} \frac{v!}{v_1! v_2!} \ _3^{v_1} (1 \quad _3)^{v_2} P(|T_v| > k \mid v \quad 2; !_2(v_1; v_2)) : \tag{9}$$

Moreover, the test statistic is not de ned if $\min(v_1; v_2) \quad 1$ and has to be extended to these possible situations. For example, at $v_1 = 1$ or $v_2 = 1$ one can estimate the pooled standard deviation on one sample only, for the case $v_1 = v_2 = 0$ one can set $T_v = 0$. Thus, even a xed sample size calculation faces substantial complications in deriving the distribution of the two sample t-test statistic under H_1.

In practice, the random aspect of the allocation is usually ignored in the sample size estimation formulas and the formula for a xed allocation is used instead. Fixed allocation sample size calculation leads to two number312(d)-339.35b31(.97

Table 5: Monte-Carlo Type I error, Power, and Sample Sizes; 10; 000 simulations; two samplet-test designs; $n_1 = 20$; random allocation

1	3	D_{2tr}	$D_{2tr;IPN}$	$D_{2tr;IPNR}$	$D^{a}_{2tr;IPN}$
		Type I error			
0.5	1	0.0480	0.0560	0.0502	0.0562
0.5	1.5	0.0500	0.0540	0.0535	0.0506
0.5	2	0.0499	0.0520	0.0520	0.0509
0.25	1	0.0508	0.0555	0.0529	0.0553
0.25	1.5	0.0497	0.0517	0.0516	0.0497
0.25	2	0.0502	0.0519	0.0519	0.0508
		Power			
0.5	1	0.8455	0.8543	0.9028	0.8070
0.5	1.5	0.8369	0.8444	0.8454	0.8181
0.5	2	0.8247	0.8384	0.8385	0.8116
0.25	1	0.8419	0.8669	0.8834	0.8264
0.25	1.5	0.8431	0.8515	0.8516	0.8235
0.25	2	0.8296	0.8429	0.8429	0.8145
		EN (SD)			
0.5	1	38.64(7.50)	41.12(16.36)	46.69(12.96)	36.78(16.54)
0.5	1.5	80.85(10.73)	90.51(36.48)	90.68(36.23)	84.99(33.60)
0.5	2	136.99(13.84)	158.68(62.09)	158.69(62.08)	147.03(56.55)
0.25	1	50.08(9.89)	58.99(28.36)	61.11(26.50)	54.06(29.26)
0.25	1.5	109.18(14.39)	128.22(59.05)	128.26(58.97)	120.27(58.77)
0.25	2	184.04(18.60)	222.00(95.56)	222.00(95.56)	206.66(95.59)

Measurements of prostate-specific antigen (PSA) levels are widely used for screening and diagnosing prostate cancer. PSA levels are known to be associated with measures of disease aggressiveness such as tumor stage as well as demographic characteristics predictive of screening behavior such as race/ethnicity, marital status, etc. A (hypothetical) investigator in Atlanta, GA wishes to conduct a study to evaluate whether the effect of Black versus White race on PSA levels is the same for localized versus regionally or distantly extended tumors. In practice he or she would turn to the SEER cancer registry, as we will for the source of data, but for the sake of the example let's assume that the information of interest is not available in the registry. In fact, PSA levels were not available in SEER until recently.

The specific goal of the study is to test the interaction effect of race (White vs Black) and tumor stage (localized vs others) on $\ln(PSA)$ values controlling for the effect of marital status (married vs others) and ethnicity (Hispanic vs others).

We use the linear regression model

$$\ln(PSA_i) = \beta_0 + \beta_1 W_i + \beta_2 L_i + \beta_3 W_i L_i + \beta_4 M_i + \beta_5 H_i + \epsilon_i; \quad (10)$$

where W_i, L_i, M_i, and H_i are, respectively, indicators of White race, localized tumor, married status, and Hispanic ethnicity of the i^{th} subject. The random noise ϵ_i is assumed to follow a normal model with the zero mean and a finite unknown variance σ^2. We formulate the research question about the interaction via $H_0: \beta_3 = 0$ and wish to design a study that would have 80% power to detect a 1.5-fold difference in the race effect among the localized versus non-localized tumors, corresponding to $\beta_3 = \ln(1.5)$.

To calculate the study sample size we use the formula proposed by Hsieh et al [6]. If X represents the predictor of interest and Z stands the other predictors, then the sample size required to detect an effect with a partial regression coefficient of β with power $100(1-\beta)\%$ at a two-sided significance

Table 6: Linear regression on internal pilot data, $n_1 = 100$.

	Estimate	Std.Error	t value	p value
Intercept ($\hat{}_0$)	5.4036	0.7208	7.497	<0.0001
White ($\hat{}_1$)	-1.2507	0.6519	-1.918	0.0581
Localized ($\hat{}_2$)	-1.4274	0.4916	-2.904	0.0046
Hispanic ($\hat{}_4$)	0.1849	0.5394	0.343	0.7326
Married ($\hat{}_5$)	-0.0928	0.2122	-0.437	0.6629
White Localized ($\hat{}_3$)	1.4046	0.6822	2.059	0.0423

To simulate the conduct of the study we extracted a sample of 142 proT432(c)3.56312(t)174(h52 Tf 1 0 0 1 128.4 539.64 3.)-2.26432]TJ -371.04 -14.4 Td a.26463

Table 7: Regression model for the total sample, $N = 1837$.

	Estimate	Std.Error	t value	p value
Intercept ($\hat{\ }_0$)	4.8725	0.1923	25.333	< 0.0001
White ($\hat{\ }_1$)	-0.1577	0.1267	-1.245	0.2134
Localized ($\hat{\ }_2$)	-0.4670	0.0989	-4.721	< 0.0001
Hispanic ($\hat{\ }_4$)	-0.0148	0.1706	-0.086	0.9311
Married ($\hat{\ }_5$)	-0.1396	0.0445	-3.136	0.0017