Journal of Data Science **9**(2011), 155-170

# Correction for Two-Group Sample Size Calculation with Uncertain Group Membership

Hung-Mo Lin[1], Shannon K. McClintock[2] and John M. Williamson[3]
[1]*Mount Sinai School of Medicine,* [2]*Emory University School of Public Health and* [3]*National Center for Zoonotic, Vector-Borne and Enteric Diseases, Centers for Disease Control and Prevention*

*Abstract*: Sample size and power calculations are often based on a two-group comparison. However, in some instances the group membership cannot be ascertained until after the sample has been collected. In this situation, the respective sizes of each group may not be the same as those prespecified due to binomial variability, which results in a difference in power from that expected. Here we suggest that investigators calculate an "expected power" taking into account the binomial variability of the group membership, and adjust the sample size accordingly when planning such studies. We explore different scenarios where such an adjustment may or may not be necessary for both continuous and binary responses. In general, the number of additional subjects required depends only slightly on the values of the (standardized) difference in the two group means or proportions, but more importantly on the respective sizes of the group membership. We present tables with adjusted sample sizes for a variety of scenarios that can be readily used by investigators at the study design stage. The proposed approach is motivated by a genetic study of cerebral malaria and a sleep apnea study.

*Key words*: Chi-square test, mean, power, proportion, two-sample t-test.

## 1. Introduction

In study design investigators often wish to determine the sample size necessary to provide a specified power. Clinical studies without a sufficient sample size can fail to detect a significant effect when it exists. This consideration must be balanced with the high cost of recruiting and evaluating large samples of subjects, thus making sample size (power) calculations a crucial step in designing clinical research studies. To conduct sample size calculations for the two-group case, an investigator needs to specify the group means (i.e., proportions for a binary response), the group variances, the desired power, the type I error rate,

the number of sides of the test, and the ratio of the two-group sizes (Bland, 2000). However, sometimes an individual's group membership is only available after the data have been collected. This may occur due to cost, time, and other considerations and is most common in non-randomized trials and observational studies.

For example, HIV studies often result in group membership that is subject to binomial variability. Often patients are recruited at STD or drug clinics and are categorized by their HIV status or CD4 count ($< 200, 200 - 500$, and $\geq 500$). Patients with and without HIV may be compared to gain insight on potential risk factors for HIV in different settings, and HIV-positive subjects may then be followed over time to assess their disease progression. Persons with lower CD4 counts are usually sicker, more at risk for opportunistic infection, and may be at an advanced stage of disease. However, the HIV tests must be sent to a laboratory making it difficult to recruit patients into the study before their status or CD4 count is known.

Sample surveys also often result in data where group membership may not be ascertained until after the sample is selected. For example, sample surveys are often conducted in sub-Sarahan Africa to assess the effectiveness of malaria prevention techniques in different settings, such as insecticide-treated bednets or indoor residual spraying, and the impediments to such use (Rowe *et al.*, 2006). In some cases households are selected from a census or surveillance list and in other cases households are randomly selected when the interviewers enter a village when no such list exists. In either case, bednet use or indoor residual spraying is not assessed until after the household has been selected and visited.

Here we describe two specific studies where group membership is subject to binomial variability, one with a continuous outcome and the other with a binary one. An example of a study with a continuous outcome is a malaria study that plans to test the role of a biological factor (BF) to protect against cerebral malaria. This study will attempt to determine if polymorphism in the BF gene influences BF levels and cases of cerebral malaria among malaria endemic populations. The genetic testing and measurement of BF levels will be conducted on stored blood samples from a previous study and will be costly and time consuming. We can assume that approximately 10% of the population has the genotype associated with low production of BF. The remaining 90% of the population has the genotype associated with normal or high production of BF. We are interested in testing the hypothesis that there is a statistically significant difference in mean BF levels between the two genotype groups.

Several studies among adults have shown that sleep-disordered breathing (SDB) is significantly associated with metabolic disorders (Punjabi and Polotski, 2005; Vgontzas, Bixler, and Chrousos, 2005). However, little is known about

the impact of SDB among children. The Penn State Children Cohort (PSCC) was designed as a population-based study of the prevalence and correlates of SDB among young children (i.e., kindergarten to fifth grade). Previous estimates indicate that approximately 25% of these children have a mild to moderate form of SDB, defined as an apnea/hypopnea index (AHI)$\geq$ 1. We are interested in testing the hypothesis that children with SDB have increased risk of insulin resistance, a major metabolic disorder. Following Cruz and Goran (2004), abnormal waist circumference (WC) was used in this study as a surrogate for increased insulin resistance. The percentages of abnormal WC in the pilot data study were 30% and 19% among those with and without SDB, respectively.

In the first phase of this study, questionnaires were sent to the parents to identify some of the signs and symptoms of their childrens' sleep disorder, such as snoring, breath cessation or difficulty breathing, restless sleep, daytime sleepiness, and school or behavior problems. However, it was not possible to confirm the presence of clinically diagnosed SDB based solely upon parental reports. Therefore, the study required a second phase which involved the selection of a subset of children from those parents who returned the questionnaire. These children then spent one night in the sleep laboratory, and only then could SDB status and WC be ascertained. The objective of the sample size calculation is to determine the number of children to participate in the phase II study.

In each of the above examples, preliminary sample size calculations may be misleading because group membership is subject to uncertainty and can only be ascertained only after the sample is collected. Methods for correcting sample sizes for testing group mean differences with random predictors or estimated population parameters have been proposed. The focus has been on a normally-distributed quantitative trait that can be described by an univariate linear model. Jayakar (1970) proposed tests for the detection of linkage between a marker locus and a locus influencing a quantitative trait. Methods for evaluating the power of F-tests based on an ANOVA model used for assessing the linkage have been presented (Soller and Genizi, 1978; Genizi and Soller, 1979). In this situation, group differences could assume a few discrete values with prescribed probabilities. The above authors note that there will be variation in the number of offspring with a particular genotype and that sample size calculations may have to be slightly inflated.

Fay *et al.* (2007) considered two sources of variability when conducting sample-size calculations for testing differences in means between two samples. The first source of variability is nonadherence (noncompliance). They assumed that the proportion of subjects who will adhere to their treatment regimen is not known before the study but is a stochastic variable with known distribution. They derive closed form sample size calculations based on asymptotic normality

taking this assumption into account. Secondly, the authors account for the variability in parameter estimates (i.e., effect size and variances) that are estimated from prior data by using a slightly larger nominal power in the usual sample size calculation. The first scenario of Fay *et al.* (2007) is similar to our problem in that the final sample size in each group may not be the same as initially planned.

In Section 2, we demonstrate how the preliminary sample size calculated in such circumstances may not achieve the power prespecified because of the binomial variability in the group sizes. We propose a method for sample size correction in these situations to account for this variability and investigate various scenarios for both continuous and binary responses where the method is required or not. We focus on the situation where group membership is ascertained after the entire sample has been collected. If group membership is being ascertained sequentially or in batches then a sequential or adaptive method might be used as suggested by a reviewer. We illustrate our approach with the above examples in Section 3. We conclude with a short discussion and make recommendations on when such sample size adjustments are necessary.

## 2. Methods

### 2.1 Overview

Let $w$ denote the weight for group 1, that is the probability that an individual will belong to group 1. Assume that $w$ is fixed throughout the study period. Further assume that $N$ individuals will be randomly selected from the study population. Usually investigators are able to select $n_1$ and $n_2$ individuals from groups 1 and 2, respectively. However, in our situation group membership is not ascertained until after sample selection, resulting in varying $n_1$ and $n_2$ group sizes. In the two examples previously described, individuals need to be enrolled prior to the group membership determination. Therefore, calculating $n_1$ and $n_2$ during the design stage requires an assumption concerning the proportion of the population classified in group 1. This leads to a potential problem in the sample size calculation; after the study is conducted and the group membership is ascertained for each individual, the actual size of each group may not be the same as what was prespecified because of binomial variability. As a result, the actual power based on the calculated sample size is different from that targeted. Thus, we distinguish between three types of power. *Targeted* power is the power specified for the study during the design stage ($1 - \beta$ where $\beta$ is the type II error rate), and is also referred to as *nominal* power (e.g., Fay *et al.*, 2007). *Initial-N* power is the power given the initially calculated sample size $N$. Here we define *expected* power as a weighted average of power estimates across the range of possible $n_1$ and $n_2$ values. This is the probability of rejecting the null hypothesis given the

study design that yields a total sample size of $N$ with the assumption that the probability of being in group 1 equals $w$.

Consider the following hypothetical example in which a study plans to recruit individuals with the assumption that 10% ($w$) of the individuals will belong to group 1 and 90% of the individuals will belong to group 2. We assume the outcome is normally distributed and that the two groups have a common standard deviation ($\sigma = 1$). The hypothesis of interest is: $H_0$: $\mu_1$- $\mu_2 = 0$ versus $H_1$: $\mu_1$- $\mu_2 \neq 0$. The total sample size of 50 is calculated based on a significance level of 0.05, a two-sided t-test with 90% power, and a mean group difference of 1.56. After recruitment group membership is ascertained; of the 50 recruited subjects, the number of individuals belonging to group 1 may not be exactly 5. Because of binomial variability, $n_1$ can range from 0 to 50, with the largest probabilities near 5. If $n_1 = 5$ or 45, the *initial-N* power is 90%, as planned. However, if $n_1 = 4$, then the *initial-N* power is 83%. In the extreme situation in which $n_1 = 0$ or 50, the *initial-N* power is 0% because for the two-sample comparison $n_1$ and $n_2$ must both be at least 1 (or 2 when the variances are estimated separately). Here we define *expected* power (EP) as the weighted average of power over all possible values of $n_1$ from 0 to $N$ with the weights being the corresponding probabilities obtained from the binomial distribution:

$$\mathrm{EP} = \sum_{y=0}^{N} \binom{N}{y} w^y (1-w)^{N-y} (\mathrm{Power}_{y,\,N}), \tag{1}$$

in which $\mathrm{Power}_{y,\,N}$ is the power given that $y$ individuals are in group 1 and $N - y$ in group 2. In Appendix 1, we present the power functions that we use for the $t$-test (O'Brien and Muller, 1993, pp. 297-344), the continuity-corrected Pearson chi-square test, and Fisher's exact test (Fisher, 1934). In particular, power for the two-sided two-sample $t$-test is based on the $F$ distribution. Power for the two-sided Pearsons chi-square test is computed by adding the lower- and upper-sided powers each with $\alpha/2$ size, where one-sided power is computed as suggested by Diegert and Diegert (1981). Then the sample sizes for the two-sided two-sample $t$-test and Pearsons chi-square test are obtained by numerically inverting the power formulas. For Fisher's exact test, power and sample size computations are based on the continuity-adjusted arcsine test (Walters, 1979; also see SAS PROC POWER, 2004). Because $n_1$ and $n_2$ are random variables during the planning stage, the *initial-N* power may be different from the *targeted* power and the *expected* power. In this example with $w = 0.1$ and $N = 50$ the *expected* power is 84.0%, which is less than the *targeted* 90%. In Appendix 2, we present the binomial probability for $n_1$ with $w = 0.1$ and $N = 50$, the *initial-N* power, and the cumulative expected power for all possible combinations of $n_1$ and $n_2$. Note that while $n_1$ can technically range from 1 to 49, the practical range of

values given the assumptions is from 1 to 16.

### 2.2 Comparison of Two Independent Groups

Now we propose a method to correct the sample size when comparing two independent means or two independent proportions. The correction is made to ensure that the expected power equals the targeted power.

**Step 1:** Perform the usual sample size calculation for comparing two groups. *(For example, the total sample size reported for the previous hypothetical example is 50, implying that $n_1$ and $n_2$ are 5 and 45, respectively.)*

**Step 2:** Calculate the expected power given $N, w$, and the other relevant parameters using formula (1) where power is calculated via the methods discussed in sections 2.4 and 2.5. *(The expected power for the example is 84.0%, which is less than the targeted power of 90%. See Appendix 2.)*

**Step 3:** If the *expected* power is less than the *targeted* power, increase the total sample size by 1.

**Step 4:** Repeat Steps 2 and 3 until the expected power achieves the targeted power. *(The required sample size is 61 to achieve 90% expected power.)*

In practice, investigators are likely to specify the ratio between $n_1$ and $n_2$ when calculating sample size following the assumption made of the two-group proportions in the population of interest. Therefore in a two-group design, all group sizes are adjusted to be multiples of the corresponding group weights. Here we advocate that, in the context of uncertain group membership at the design stage, the need to force the $n_1$ and $n_2$ ratio to be fixed should be relaxed because the ratio itself is an estimate as $n_1$ and $n_2$ are subject to binomial variability. Forcing the ratio of $n_1$ and $n_2$ to remain as specified can potentially result in a very conservative estimate of $N$ yielding more power than desired. Different software packages have different methods for specifying $n_1$ and $n_2$ or their ratio. For the purpose of illustration, in Appendix 3 we show how the NFRACTIONAL option in SAS PROC POWER can affect the sample size calculation using the previous example.

### 2.3 Sample Size Correction for Comparing Two Means

Define $N^*$ as the corrected sample size that takes into account the binomial variability of group membership, and define the correction factor ($CF$) as the ratio between the corrected sample size and the original calculated sample size, that is $CF = N^*/N$. When comparing two means, one can express the means

and variances together in terms of the standardized effect size (assuming equal variances). Here we focus on the relationships of the sample size correction factor and the difference $N^* - N$ with power and the standardized effect size, assuming the type I error rate is 0.05 with a two-sided test. We calculated the initial $N$ and followed Steps 2, 3, and 4 until the *expected* power was at least the targeted power for each scenario.

Table 1 shows the unadjusted sample size $N$, the difference between the adjusted $N^*$ and $N$, and the correction factor as a function of targeted power (90% and 80%) and standardized effect size (Std ES= 0.2 to 3) for different group 1 weights ($w = 0.05$, 0.1). In general, the CF is largest for small group 1 weights and large effect sizes. However, the additional sample size required is more appreciable when $w$ approaches 0.1 or less. For $w = 0.2$, 2 to 4 extra subjects are needed and no more than 2 extra subjects are needed for $w \geq 0.3$ across all effect sizes (results not shown). Interestingly, the difference between $N^*$ and $N$ remains similar across a wide range of standardized effect sizes. Furthermore, a larger standardized effect size or a one-sided test (results not shown) alone does not always correspond to a larger CF.

Table 1: The unadjusted sample size $N$, the additional sample required ($N^* - N$), and the correction factor (CF) as a function of targeted power, standardized effect size (Std ES) and group 1 weight ($w$). A two-sided two-sample $t$-test was used for hypothesis testing at the $\alpha = 0.05$ level.

| | **80% Targeted Power** | | | | | | **90% Targeted Power** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Std | $w = 0.05$ | | | $w = 0.1$ | | | $w = 0.05$ | | | $w = 0.1$ | | |
| ES | $N$ | $N^* - N$ | CF | $N$ | $N^* - N$ | CF | $N$ | $N^* - N$ | CF | $N$ | $N^* - N$ | CF |
| 0.2 | 4133 | 16 | 1.00 | 2183 | 7 | 1.00 | 5533 | 23 | 1.00 | 2921 | 10 | 1.00 |
| 0.4 | 1035 | 15 | 1.01 | 547 | 7 | 1.01 | 1385 | 23 | 1.02 | 732 | 10 | 1.01 |
| 0.6 | 461 | 16 | 1.03 | 245 | 7 | 1.03 | 617 | 23 | 1.04 | 327 | 10 | 1.03 |
| 0.8 | 261 | 15 | 1.06 | 139 | 7 | 1.05 | 348 | 23 | 1.07 | 185 | 10 | 1.05 |
| 1.0 | 168 | 15 | 1.09 | 90 | 7 | 1.08 | 224 | 23 | 1.10 | 119 | 11 | 1.09 |
| 1.2 | 117 | 16 | 1.14 | 63 | 7 | 1.11 | 156 | 24 | 1.15 | 84 | 10 | 1.12 |
| 1.4 | 87 | 15 | 1.17 | 47 | 7 | 1.15 | 115 | 24 | 1.21 | 62 | 11 | 1.18 |
| 1.6 | 67 | 16 | 1.24 | 37 | 7 | 1.19 | 89 | 24 | 1.27 | 48 | 11 | 1.23 |
| 1.8 | 53 | 16 | 1.30 | 29 | 7 | 1.24 | 71 | 24 | 1.34 | 39 | 10 | 1.26 |
| 2.0 | 44 | 16 | 1.36 | 24 | 7 | 1.29 | 58 | 24 | 1.41 | 32 | 11 | 1.34 |
| 2.2 | 37 | 16 | 1.43 | 21 | 7 | 1.33 | 48 | 25 | 1.52 | 27 | 11 | 1.41 |
| 2.4 | 31 | 17 | 1.55 | 18 | 7 | 1.39 | 41 | 25 | 1.61 | 23 | 11 | 1.48 |
| 2.6 | 27 | 17 | 1.63 | 16 | 7 | 1.44 | 35 | 26 | 1.74 | 20 | 11 | 1.55 |
| 2.8 | 24 | 17 | 1.71 | 14 | 7 | 1.50 | 31 | 26 | 1.84 | 18 | 11 | 1.61 |
| 3.0 | 21 | 17 | 1.81 | 12 | 8 | 1.67 | 27 | 27 | 2.00 | 16 | 11 | 1.69 |

We also investigated the above relationships when using the Satterthwaite t-

test assuming unequal variances for groups 1 and 2. Exact solutions for power for the two-sided case are presented by Moser, Stevens, and Watts (1989). To illustrate, we continue our hypothetical example assuming various $\sigma_1$ $(= 0.5, 1, 1.5)$ and $\sigma_2$ $(= 0.5, 1, 1.5)$ values. The results for power= 90% and $w = 0.1$ are presented in Table 2. Many of the observations noted in the equal variance case are applicable here (results for power= 80% and $w = 0.3$ and 0.5 not shown). Specifically, the relative increase in sample size caused by binomial variability (i.e., CF) is greater when the study's sample size is smaller and when a greater difference exists between the two groups' weights. In general, a larger *targeted* power also, but not always, corresponds to a larger CF. Given the same standard deviation for one group, the CF and $N^*-N$ decrease as the standard deviation for the other group increases. In most cases, a larger CF corresponds to the situation in which the smaller standard deviation is associated with the smaller group weight. For instance, with weights= 0.1 and 0.9 for groups 1 and 2, respectively, the CF is 1.72 when $\sigma_1 = 0.5$ and $\sigma_2 = 1$, as compared to 1.26 when $\sigma_1 = 1$ and $\sigma_2 = 0.5$.

Table 2: Sample Size Calculations for Testing $H_0$: $\mu_1 - \mu_2 = 0$ versus $H_1$: $\mu_1 - \mu_2 \neq 0$, assuming $\mu_1 - \mu_2 = 1.56$ with $w = 0.1$.

| Targeted Power | $\sigma_1$ | $\sigma_2$ | $N$ | Expected Power | $N^*$ | CF[a] | $N^* - N$ |
|---|---|---|---|---|---|---|---|
| 90% | 1.0 | 0.5 | 66 | 80.3% | 83 | 1.26 | 17 |
| | 1.0 | 1.0 | 67 | 81.1% | 83 | 1.24 | 16 |
| | 1.0 | 1.5 | 70 | 82.6% | 85 | 1.21 | 15 |
| | 0.5 | 1.0 | 32 | 59.6% | 55 | 1.72 | 23 |
| | 1.0 | 1.0 | 67 | 81.1% | 83 | 1.24 | 16 |
| | 1.5 | 1.0 | 122 | 86.4% | 135 | 1.11 | 13 |

Note:
Unequal variances are assumed with $\alpha = 0.05$ and a two-sided test.

[a] The correction factor is $N^*/N$.

### 2.4 Sample Size Correction for Comparing Two Proportions

Here we investigate the relationship between the sample size correction factor for a two-sided Pearson chi-square test with a significance level of 0.05, power of 80% and 90%, group 1 weight ($w = 0.05$ and 0.1), and the outcome proportions for groups 1 and 2, $P_1$ $(= 0.1$ to $0.5)$ and $P_2$ $(= 0.1$ to $0.9)$. We calculated an initial sample size $N$ for a test of a difference in proportions using Pearson's chi-square test, or Fisher's exact test when the assumption of expected cell counts greater or equal to 5 is violated. Then we followed Steps 2, 3, and 4 until the *expected* power achieved the *targeted* power. Results are presented in Table 3.

Table 3: The unadjusted sample size $N$, the additional sample required ($N^*$-$N$), and the correction factor (CF) as a function of targeted power, proportions of outcome of interest for groups 1 and 2 ($P_1$ and $P_2$), and group 1 weight ($w$). Pearson's chi-square test, or Fisher's exact test (bold) when the assumption of all expected cell counts $\geq 5$ does not hold, was used for hypothesis testing at the $\alpha = 0.05$ level.

| | | **80% Targeted Power** | | | | | | **90% Targeted Power** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $w = 0.05$ | | | $w = 0.1$ | | | $w = 0.05$ | | | $w = 0.1$ | | |
| $P_1$ | $P_2$ | $N$ | $N^* - N$ | CF | $N$ | $N^* - N$ | CF | $N$ | $N^* - N$ | CF | $N$ | $N^* - N$ | CF |
| 0.1 | 0.2 | 2251 | 18 | 1.01 | 1182 | 8 | 1.01 | 2875 | 26 | 1.01 | 1516 | 12 | 1.01 |
| | 0.3 | **746** | **10** | **1.01** | **393** | **0** | **1.00** | **958** | **19** | **1.02** | **505** | **3** | **1.01** |
| | 0.4 | **399** | **14** | **1.04** | **210** | **3** | **1.01** | **507** | **22** | **1.04** | **267** | **7** | **1.03** |
| | 0.5 | **257** | **15** | **1.06** | **135** | **5** | **1.04** | **324** | **24** | **1.07** | **171** | **8** | **1.05** |
| | 0.6 | **181** | **17** | **1.09** | **96** | **5** | **1.05** | **227** | **26** | **1.11** | **120** | **9** | **1.08** |
| | 0.7 | **134** | **17** | **1.13** | **71** | **6** | **1.08** | **167** | **27** | **1.16** | **88** | **11** | **1.13** |
| | 0.8 | **102** | **17** | **1.17** | **54** | **6** | **1.11** | **126** | **27** | **1.21** | **67** | **10** | **1.15** |
| | 0.9 | **77** | **17** | **1.22** | **41** | **6** | **1.15** | **94** | **28** | **1.30** | **50** | **11** | **1.22** |
| 0.2 | 0.1 | 1836 | 13 | 1.01 | 987 | 6 | 1.01 | 2584 | 20 | 1.01 | 1381 | 8 | 1.01 |
| | 0.3 | 3202 | 16 | 1.00 | 1685 | 7 | 1.00 | 4189 | 24 | 1.01 | 2209 | 11 | 1.00 |
| | 0.4 | 886 | 17 | 1.02 | 467 | 8 | 1.02 | 1146 | 25 | 1.02 | 606 | 11 | 1.02 |
| | 0.5 | **476** | **12** | **1.03** | **251** | **2** | **1.01** | **526** | **26** | **1.05** | **280** | **11** | **1.04** |
| | 0.6 | **289** | **14** | **1.05** | **152** | **4** | **1.03** | **368** | **23** | **1.06** | **194** | **8** | **1.04** |
| | 0.7 | **194** | **15** | **1.08** | **102** | **5** | **1.05** | **245** | **24** | **1.10** | **129** | **9** | **1.07** |
| | 0.8 | **136** | **16** | **1.12** | **72** | **5** | **1.07** | **171** | **25** | **1.15** | **90** | **10** | **1.11** |
| | 0.9 | **97** | **16** | **1.16** | **52** | **5** | **1.10** | **120** | **26** | **1.22** | **64** | **10** | **1.16** |
| 0.3 | 0.1 | 520 | 12 | 1.02 | 285 | 5 | 1.02 | 748 | 19 | 1.03 | 405 | 8 | 1.02 |
| | 0.2 | 2905 | 14 | 1.00 | 1546 | 6 | 1.00 | 3981 | 21 | 1.01 | 2112 | 10 | 1.00 |
| | 0.4 | 3810 | 16 | 1.00 | 2009 | 7 | 1.00 | 5042 | 24 | 1.00 | 2661 | 11 | 1.00 |
| | 0.5 | 984 | 16 | 1.02 | 521 | 7 | 1.01 | 1298 | 24 | 1.02 | 687 | 11 | 1.02 |
| | 0.6 | 429 | 15 | 1.03 | 228 | 8 | 1.04 | 566 | 24 | 1.04 | 302 | 10 | 1.03 |
| | 0.7 | **297** | **13** | **1.04** | **157** | **3** | **1.02** | **380** | **22** | **1.06** | **201** | **7** | **1.03** |
| | 0.8 | **189** | **15** | **1.08** | **100** | **4** | **1.04** | **240** | **24** | **1.10** | **127** | **8** | **1.06** |
| | 0.9 | **125** | **14** | **1.11** | **66** | **5** | **1.08** | **156** | **25** | **1.16** | **83** | **9** | **1.11** |
| 0.4 | 0.1 | **373** | **9** | **1.02** | 139 | 5 | 1.04 | 362 | 18 | 1.05 | 199 | 7 | 1.04 |
| | 0.2 | 767 | 14 | 1.02 | 412 | 6 | 1.01 | 1063 | 21 | 1.02 | 568 | 9 | 1.02 |
| | 0.3 | 3632 | 15 | 1.00 | 1925 | 7 | 1.00 | 4917 | 23 | 1.00 | 2603 | 10 | 1.00 |
| | 0.5 | 4084 | 15 | 1.00 | 2156 | 7 | 1.00 | 5447 | 23 | 1.00 | 2877 | 10 | 1.00 |
| | 0.6 | 997 | 16 | 1.02 | 529 | 7 | 1.01 | 1334 | 23 | 1.02 | 707 | 10 | 1.01 |
| | 0.7 | 409 | 15 | 1.04 | 219 | 7 | 1.03 | 553 | 22 | 1.04 | 295 | 10 | 1.03 |
| | 0.8 | **281** | **12** | **1.04** | **149** | **2** | **1.01** | 271 | 21 | 1.08 | 147 | 9 | 1.06 |
| | 0.9 | **167** | **13** | **1.08** | **89** | **3** | **1.03** | **211** | **23** | **1.11** | **112** | **7** | **1.06** |
| 0.5 | 0.1 | **237** | **12** | **1.05** | **126** | **2** | **1.02** | 212 | 18 | 1.08 | 118 | 8 | 1.07 |
| | 0.2 | 350 | 14 | 1.04 | 190 | 6 | 1.03 | 486 | 20 | 1.04 | 261 | 9 | 1.03 |
| | 0.3 | 925 | 15 | 1.02 | 493 | 7 | 1.01 | 1256 | 22 | 1.02 | 668 | 9 | 1.01 |
| | 0.4 | 4024 | 16 | 1.00 | 2129 | 7 | 1.00 | 5405 | 23 | 1.00 | 2857 | 10 | 1.00 |

In general, the CF is larger for smaller group 1 weights ($w$). As the difference between $P_1$ and $P_2$ increases, the sample size $N$ decreases and the CF increases. As with the two means comparison, the additional sample size required is more appreciable when $w$ approaches 0.1 or less. For $w = 0.2$ and 0.3, 0 to 4 extra

subjects are needed, and no more than 2 extra subjects are needed for $w > 0.3$ across all combinations of $P_1$ and $P_2$ (results not shown). The difference between $N^*$ and $N$ depends only slightly on the values of $P_1$and $P_2$ but more importantly on the value of $w$. In general use of Fisher's exact test in the sample size calculation will result in a larger total sample size, as expected. However, the need for sample size correction due to the binomial variability of the membership assignment is negligible for $w > 0.1$ and can be totally ignored for $w \geq 0.2$. Table 3 provides guidance on how many additional subjects are required. Similar to the two-sample means comparison, larger *targeted* power or a one-sided test (not shown) does not always correspond to a larger CF. In all the scenarios investigated in sections 2.3 and 2.4, the *expected* power was always less than the *targeted* power.

## 3. Examples

### 3.1 Cerebral Malaria Study

For the cerebral malaria example presented in the Introduction, we are interested in testing the null hypothesis that there is no statistically significant difference in mean BF levels between the two genotype groups. We assume that the two groups comprise 10% (low BF levels) and 90% (normal or high BF levels) of the population, respectively. We assume that the BF levels are normally distributed with means of 250 and 500 for the respective groups with a common standard deviation of 100. A $t$-test will be conducted to evaluate the null hypothesis. We will need a sample of 21 to achieve 90% power based on the usual sample size calculations without taking into account the binomial variability of the group sizes. However, the expected power is only 76% with this sample size. One would need a sample of size 32 to ensure that the expected power is at least 90%.

### 3.2 Children Sleep Disorder Study

The main hypothesis of the Penn State Children Cohort was that children with SDB, defined as an apnea/hypopnea index (AHI)$\geq$ 1 and estimated to be 25% ($w$) of the population, have increased risk of metabolic disorder. The abnormal waist circumference (WC) was used as a surrogate for insulin resistance, a major metabolic disorder (Cruz and Goran, 2004). The percentages of abnormal WC in the pilot data study were 30% and 19% among those with and without SDB, respectively. The objective of the sample size calculation is to determine the number of children needed to participate in the phase II study to detect an 11% difference in the abnormal WC rates between the SDB and no SDB groups that

was found in the pilot study.

The first row of Table 4 shows the original and adjusted sample sizes for the scenario under consideration. There was a little difference between $N$ and $N^*$. We also performed several sensitivity analyses for different values of $w$, $P_1$, and $P_2$ because the prevalence rates of SDB and WC were estimated based on minimal preliminary data. In this particular example, the study required a large phase 2 sample size $N$ in general, and it is quite sensitive to the specification of $w$, $P_1$, and $P_2$. This example illustrates a situation in which sample size correction is not needed because the expected power for the initial sample size is approximately the same as the targeted power for all the scenarios displayed in Table 4.

Table 4: Sample Size Calculations for the Penn State Children Cohort Study

| Targeted Power | $w^a$ | $P_1$ | $P_2$ | $N$ | Expected Power | $N^*$ | CF$^b$ |
|---|---|---|---|---|---|---|---|
| 90% | 0.25 | 0.3 | 0.19 | 835 | 89.9% | 837 | 1.002 |
| | 0.25 | 0.3 | 0.20 | 1028 | 89.9% | 1031 | 1.003 |
| | 0.25 | 0.3 | 0.15 | 415 | 89.9% | 418 | 1.007 |
| | 0.2 | 0.3 | 0.19 | 973 | 89.9% | 977 | 1.004 |
| | 0.3 | 0.3 | 0.19 | 749 | 89.9% | 751 | 1.003 |

[a] The proportion of subjects in group 1.
[b] The correction factor equal to $N^*/N$.

## 4. Discussion

Two SAS macros that calculate the minimum sample size needed to achieve an expected power greater than or equal to the targeted power for a two-group comparison with a continuous and a binary response, respectively, are available from the authors upon request. One must specify $w$ and $1 - w$ (the weights of the two groups), the targeted power, the type I error rate, and the number of sides of the test. For a continuous response, one must additionally specify the expected means $\mu_1$ and $\mu_2$ for groups 1 and 2 and their respective standard deviations (whether assumed equal or unequal). For the binary response, one must additionally specify the expected probabilities $P_1$and $P_2$ of the outcome of interest for groups 1 and 2. The macros output the unadjusted sample size $N$ without taking into account the uncertainty of the group membership, the initial-$N$ power, the expected power for $N$, the adjusted sample size $N^*$, the expected power for $N^*$, and the correction factor. The program will search for and return the required sample size without requiring the user to manually increase or decrease the sample size iteratively.

There may be instances when a simpler (or cheaper), although imperfect, test may be available for ascertaining group membership. As noted by a referee, if the probabilities of correctly identifying subjects from each group are known or can be estimated, then a study may achieve its specified power with a sample size less than the adjusted sample size $N^*$, or even the initial sample size $N$, by refusing more subjects with a higher probability of belonging to group 2. This new sample size will of course be related to the misclassification probabilities of the imperfect test.

To account for binomial variability in group sizes, we encourage investigators to calculate *expected* power when designing a study that compares two groups in which the group membership is not defined until the data have been collected. The tables and macros provided by the authors can be used by an investigator to verify (in any particular trial) if ignoring the sample size correction will impact the power of the study. Although the increased sample size that results from taking into account the binomial variability will ensure, on average, adequate power to detect an effect, it will not guarantee it because a great imbalance between the two group sizes is still possible. In general, the relative increase in sample size (CF) caused by binomial variability is greater for larger differences in the group weights for both continuous and binary outcomes. The CF also increases as the standardized effect size increases for continuous outcomes or as the difference in proportions increases for binary outcomes. More importantly, the difference between $N^*$ and $N$ depends only slightly on the values of standardized effect size (or $P_1$ and $P_2$ when the response is binary), but more noticeably on the value of $w$, the probability of group 1 membership. However, the difference should be negligible unless the group weights are fairly dissimilar (0.1 or less). These methods can be extended when there are more than two groups subject to uncertain group membership.

## Appendix 1. Power Functions for Comparing Two Means and Two Proportions

| Notation | |
|---|---|
| $\alpha$ | significance level |
| $N$ | total sample size |
| $w_i$ | allocation weight for $i^{th}$ group (standardized to sum to 1) |
| $\mu_{diff}$ | mean difference |
| $\mu_0$ | null mean |
| $\sigma$ | common standard deviation |
| $p_i$ | proportion of successes in group $i$ |
| $p_0$ | null proportion of successes |
| $z_p$ | the $p^{th}$ quantile from the standard normal distribution |
| $F(v_1, v_2, \lambda)$ | $F$ distribution with numerator d.f. $v_1$, denominator d.f. $v_2$, and noncentrality parameter $\lambda$ |
| $\Phi(\cdot)$ | cumulative distribution function of the standard normal distribution |

For the two-sample, two-sided $t$-test assuming equal variances, the exact power function is given by O'Brien and Muller, 1993:

$$\delta = N^{\frac{1}{2}}(w_1 w_2)^{\frac{1}{2}}\left(\frac{\mu_{diff}-\mu_0}{\sigma}\right),$$

$$power = P\left(F(1, N-2, \delta^2) \geq F_{1-\alpha}(1, N-2)\right).$$

Power function for the two-sample Pearson chi-square test for two proportions:

$$power = \Phi\left(\frac{(p_2-p_1-p_0)(Nw_1w_2)^{\frac{1}{2}} - z_{1-\frac{\alpha}{2}}\left[(w_1p_1+w_2p_2)(1-w_1p_1-w_2p_2)\right]^{\frac{1}{2}}}{\left[w_2p_1(1-p_1)+w_1p_2(1-p_2)\right]^{\frac{1}{2}}}\right)$$
$$+\Phi\left(\frac{-(p_2-p_1-p_0)(Nw_1w_2)^{\frac{1}{2}} - z_{1-\frac{\alpha}{2}}\left[(w_1p_1+w_2p_2)(1-w_1p_1-w_2p_2)\right]^{\frac{1}{2}}}{\left[w_2p_1(1-p_1)+w_1p_2(1-p_2)\right]^{\frac{1}{2}}}\right).$$

Power function for Fisher Exact Test for two proportions:

$$\delta = (4Nw_1w_2)^{\frac{1}{2}}\left[\arcsin\left(\left[p_2+\frac{1}{2Nw_2}(I_{\{p_2<p_1\}}-I_{\{p_2>p_1\}})\right]^{\frac{1}{2}}\right)\right.$$
$$\left.-\arcsin\left(\left[p_1+\frac{1}{2Nw_1}(I_{\{p_1<p_2\}}-I_{\{p_1>p_2\}})\right]^{\frac{1}{2}}\right)\right],$$

$$power = \Phi\left(\delta - z_{1-\frac{\alpha}{2}}\right)+\Phi\left(-\delta - z_{1-\frac{\alpha}{2}}\right).$$

**Appendix 2. Binomial Probability for $w = 0.1$ and $N = 50$, the Initial-N Power, and the Cumulative Expected Power**

| $n_1$ | $n_2$ | Binomial $(n_1, 0.1)$ pdf | Initial-$N$ Power given $(n_1, n_2)$ | Initial-$N$ Power ×Binomial pdf | Cumulative expected power |
|---|---|---|---|---|---|
| 0 | 50 | 0.00515 | 0.00000 | 0.00000 | 0.00000 |
| 1 | 49 | 0.02863 | 0.32787 | 0.00939 | 0.00939 |
| 2 | 48 | 0.07794 | 0.56290 | 0.04387 | 0.05326 |
| 3 | 47 | 0.13857 | 0.72808 | 0.10089 | 0.15415 |
| 4 | 46 | 0.18090 | 0.83454 | 0.15097 | 0.30512 |
| 5 | 45 | 0.18492 | 0.90015 | 0.16646 | 0.47158 |
| 6 | 44 | 0.15410 | 0.93967 | 0.14481 | 0.61639 |
| 7 | 43 | 0.10763 | 0.96327 | 0.10367 | 0.72006 |
| 8 | 42 | 0.06428 | 0.97735 | 0.06282 | 0.78288 |
| 9 | 41 | 0.03333 | 0.98580 | 0.03286 | 0.81574 |
| 10 | 40 | 0.01518 | 0.99092 | 0.01505 | 0.83079 |
| 11 | 39 | 0.00613 | 0.99407 | 0.00610 | 0.83688 |
| 12 | 38 | 0.00222 | 0.99603 | 0.00221 | 0.83909 |
| 13 | 37 | 0.00072 | 0.99728 | 0.00072 | 0.83981 |
| 14 | 36 | 0.00021 | 0.99808 | 0.00021 | 0.84002 |
| 15 | 35 | 0.00006 | 0.99861 | 0.00006 | 0.84007 |
| 16 | 34 | 0.00001 | 0.99897 | 0.00001 | 0.84009 |
| 17 | 33 | 0.00000 | 0.99921 | 0.00000 | 0.84009 |
| 18 | 32 | 0.00000 | 0.99937 | 0.00000 | 0.84009 |
| 19 | 31 | 0.00000 | 0.99949 | 0.00000 | 0.84009 |
| 20 | 30 | 0.00000 | 0.99957 | 0.00000 | 0.84009 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 49 | 1 | 0.00000 | 0.32787 | 0.00000 | 0.84009 |
| 50 | 0 | 0.00000 | 0.00000 | 0.00000 | 0.84009 |

**Appendix 3. NFRACTIONAL Option in SAS**

In SAS PROC POWER, there are two methods available to specify group weights. For the first method you may specify integer values for the two group weights with the GROUPWEIGHTS option. For the second method you may specify decimal values for group weights with the NFRACTIONAL option. When specifying integer values with the GROUPWEIGHTS option the total sample size must be a multiple of the sum of the two weight integers. However, when specifying decimal values in conjunction with the NFRACTIONAL option this no longer applies. For example, suppose our pre-specified group weights were 10% for group 1 and 90% for group 2. Using option 1 with group weights 1 and

9, a sample size of 50 yields a *initial-N* power of 90%. Using option 2 with group weights 0.1 and 0.9 and the NFRACTIONAL option we obtained the same result. However, if we slightly changed our pre-specified group weights to 11% for group 1 and 89% for group 2, then using option 1 a sample size of 100 yields an *initial-N* power of 0.998, which is much higher than the *targeted* power. Using option 2, a sample size of 47 yields an *initial-N* power of 0.906. Note that SAS only produces the total sample size, but not the sample sizes for each group. In the above example with *initial-N*= 47 and pre-specified group weight 11% for group 1, the rounded-down $n_1$is 5 and the rounded-up $n_1$ is 6. Note that in either case the percent of subjects in group 1 is not exactly equal to $w$, which is in accordance to our previous assumption of binomial variability.

**Acknowldegements**

Shannon McClintock's research is supported by an appointment at the Division of Parasitic Diseases, National Center for Zoonotic Vector-Borne and Enteric Diseases, Centers for Disease Control and Prevention, Atlanta, GA, and the support of Atlanta Research and Education Foundation, Decatur GA. The authors thank Venkatachalum Udhayakumar of the Centers for Disease Control and Prevention for critically reading the manuscript, and Dr. Edward O Bixler for sharing details of the Penn State Children Cohort Study.

## References

Bland, M. (2000). *An Introduction to Medical Statistics.* Oxford University Press, Oxford.

Cruz, M. L. and Goran, M. I. (2004). The metabolic syndrome in children and adolescents. *Current Diabetes Report* **4(1)**, 53-62.

Diegert, C. and Diegert, K. V. (1981). Note on inversion of Casagrande-Pike-Smith approximate sample-size formula for Fisher-Irwin test on 2 X 2 tables. *Biometrics* **37**, 595.

Fay, M. P., Halloran, M. E. and Follmann, D. A. (2007). Accounting for variability in sample size estimation with applications to nonadherence and estimation of variance and effect size. *Biometrics* **63**, 465–474.

Fisher, R. A. (1934). *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh.

Genizi, A. and Soller, M. (1979). Power derivation in an ANOVA model which is intermediate between the "fixed-effects" and the "random-effects" models. *Journal of Statistical Planning and Inference* **3**, 127-134.

Jayakar, S.D. (1970). On the detection and estimation of linkage between a locus influencing a quantitative character and a marker locus. *Biometrics* **26**, 451-464.

Moser, B. K., Stevens, G. R. and Watts, C. L. (1989). The two-sample T Test versus Satterthwaite's Approximate F Test. *Communications in Statistics A — Theory and Methods* **18**, 3963-3975.

O'Brien, R. G. and Muller, K. E. (1993). Unified power analysis for T-tests through multivariate hypothesis. *Applied Analysis of Variance in Behavioral Science*. Marcel Dekker, New York.

Punjabi, N. M. and Polotsky, V. Y. (2005). Disorders of glucose metabolism in sleep apnea. *Journal of Applied Physiology* **99**, 1998-2007.

Rowe, A. K., Rowe, S. Y., Snow, R. W., Korenromp, E. L., Schellenberg, J. R., Stein, C., Nahlen, B. L., Bryce, J., Black, R. E. and Steketee, R. W. (2006). The burden of malaria mortality among African children in the year 2000. *International Journal of Epidemiology* **35**, 691-704.

SAS Institute, Inc. (2004). *SAS/STAT 9.1 User's Guide*. Author, Cary, NC.

Soller, M. and Genizi, A. (1978). The efficiency of experimental designs for the detection of linkage between a marker locus and a locus affecting quantitative trait in segregating populations. *Biometrics* **34**, 47-55.

Vgontzas, A. N., Bixler, E. O. and Chrousos, G. P. (2005). Sleep apnea is a manifestation of the metabolic syndrome. *Sleep Medicine Review* **9**, 211-224.

Walters, D.E. (1979). In Defence of the Arc Sine Approximation. *The Statistician* **28**, 219 - 232.

Received July 20, 2009; accepted January 18, 2010.

Hung-Mo Lin
Department of Anesthesiology
Mount Sinai School of Medicine
One Gustave L. Levy Place, Box 1010, New York, NY 10029, U.S.A.
Hung-Mo.Lin@mountsinai.org

Shannon K. McClintock
Emory University School of Public Health
Department of Biostatistics and Bioinformatics
1518 Clifton Road Atlanta, GA 30322
skmcclintock@gmail.com

John M. Williamson
Division of Parasitic Diseases
National Center for Zoonotic, Vector-Borne and Enteric Diseases, Centers for Disease Control and Prevention
4770 Buford Highway, Atlanta, GA 30341, U.S.A.
jow5@cdc.gov