Published in final edited form as:

Stat Med. 2010 May 10; 29(10): 1137–1144. doi:10.1002/sim.3873.

Power for Studies with Random Group Sizes

Walter T. Ambrosius^{1,*} and Jonathan D. Mahnken²

¹ Wake Forest University School of Medicine, University of Kansas Medical Center, Kansas City, KS, 66160

² Department of Biostatistics, 3901 Rainbow Blvd., MSN 1026, University of Kansas Medical Center, Kansas City, KS, 66160

Summary

In any study it is essential to select the sample size carefully to ensure adequate power. For many studies this is simple: recruit a desired number of subjects within each group, conduct measurements, and perform the statistical test. In some studies (*e.g.*, observational studies), however, the group membership is unknown at recruitment. In this paper we examine the effect of random group sizes on power. Additionally, we consider the situation when the group proportions are unknown and specified by a prior distribution. The problem that initially motivated this research is presented (power for a 2-by-2 table), as are examples using continuous outcomes. We find that standard estimates of power using expected group sizes can over- or under-estimate power. SAS macros are available at <http://www.phs.wfubmc.edu/public/wambrosi/RandomPower>.

Keywords

Binomial/Multinomial; Beta/Dirichlet; ANOVA; chi-square test

1. Introduction

When planning any study it is necessary to determine the appropriate sample size. For many studies this is straightforward. For example, in a randomized comparison of treatments A and B it is easy to assign the desired number of participants to each experimental condition. This is also simple in many observational studies, for example comparing women to men or young to old, when the *exposure* is known at recruitment. Julious [1], for normal data, and Sahai and Khurshid [2], for proportions, describe how to calculate sample sizes for clinical trials (or observational studies with known group sizes). In some observational studies, however, the exposure is unknown until more detailed measurements are made. That is, the number of subjects in each exposure group is a random variable, even when the total sample size is fixed. Assume for a moment that we know the proportions within each of the k possible groups. We propose that averaging power over the distribution of possible group sizes provides a better estimate of power as compared with the usual practice of calculating power with the expected group sizes. We will then relax the often invalid assumption of known proportions by using a beta or Dirichlet prior to specify the distribution of possible group proportions.

Other researchers have examined similar problems related to the one we address here. Spiegelhalter *et al* describe a hybrid classical-Bayesian method for calculation of power

* Correspondence to: Walter T. Ambrosius, Department of Biostatistical Sciences, Division of Public Health Sciences, Wake Forest University School of Medicine, Medical Center Boulevard, Winston-Salem, NC 27157-1063, wambrosi@wfubmc.edu, 336-716-6281, 336-716-6427 (facsimile).

where a prior distribution on the effect size is used for the study design but classical methods will be used for calculating power [3]. They calculate power averaged over a prior distribution on the effect size. We will also specify a Bayesian prior on the design and then use classical methods for calculating power.

O'Hagan *et al* describe the concept of *assurance* as the “unconditional probability that a trial will lead to a specific outcome” [4]. Assurance is the estimated power averaged over a prior distribution on the assumed parameters. They argue that a Bayesian perspective at study design is appropriate as it can take into account all available information. One consequence of placing a prior distribution on the relevant parameters (as opposed to assuming fixed values) is that the estimated power may be lower after accounting for the prior. The usual power calculation is conditioned on assumptions made about the underlying distributions and the effect size. In a Bayesian formulation there can be reduced average power because some of the prior mass is on smaller effect sizes with lower power.

Fay *et al* examined the effects of non-adherence and estimation of variance on power [5]. A common adjustment for dropout is to increase the sample size such that the number of evaluable subjects is specified to achieve the desired assurance or power confidence (*i.e.*, if r is the expected rate of dropout and n is the required sample size, then $n/(1 - r)$ are recruited). Fay *et al* also demonstrated that the common adjustment underestimates the sample size. Fay *et al* demonstrated that using an estimated variance results in an overestimate of statistical power and that the overestimation is greatest when the variance is estimated with few degrees of freedom. This was also described by Julious [6] and Julious and Owen [7]. To make this calculation, Fay *et al* and Julious and Owen used the concept of averaging power over the distribution of unknown parameters, much like Spiegelhalter *et al* [3], O'Hagan *et al* [4], and we propose.

Finally, the current work is similar in spirit to a previous paper by Ambrosius, Lange, and Langefeld where they examined the impact of unknown genotype distributions on power for genetic association studies [8].

In this paper we examine the impact of design uncertainty in the calculation of power for observational studies. We will first describe the example that motivated this line of research in Section 2. In Section 3 we will describe the proposed method for calculation of power and present examples, including an example with random effect sizes (Subsection 3.5). We consider the implications of this research in Section 4.

2. Motivational Example

This research was prompted by a request for a sample size calculation for a study examining the relationship between the two binary factors of bitter taste status and susceptibility to nausea [9]. The hypothesis was that both are related to the body's ability to prevent ingestion of toxic substances. Although determination of bitter taste receptor status is much simpler than nausea susceptibility, it is not immediately obvious at recruitment. The hypothesized proportions of the four cells are presented in Table I. We will return to this example later in the paper but first we will discuss notation and some simpler examples.

3. Methods and Examples

When designing a study, we would like to know the probability that we will end up with a significant statistical test (*i.e.*, power of the study) given assumptions about the magnitude of the effect, variability, statistical test, and the significance level. In most studies, the sample sizes within each group are fixed at the design stage. As demonstrated by the example in the previous section this is not always the case and, in some experiments, the sample sizes

within each group will be random. If we consider that power is the *a priori* probability of a significant result, we believe that an appropriate estimate of power would account for the random group sizes by integrating over the possible set of sample sizes within groups.

3.1. Notation

We will start with simple notation for a two-group comparison. Assume that we have two groups, A and B, and that the population frequencies are p_1 and $p_2 = 1 - p_1$. We assume that we recruit a total sample of size N individuals, n_1 is the number of A, n_2 is the number of B, and $n_1 + n_2 = N$. In designing most observational studies we specify N and not the vector $\mathbf{n} \equiv (n_1, n_2)$.

If we know $\mathbf{n}$, the significance level of the test (α), the statistical test (T), the model (M), the null hypothesis (H_0), the alternative hypothesis space (H_1), and the assumed alternative effect (E), then we can calculate the statistical power using standard methods. Denote this power as $\pi(\mathbf{n}|\alpha, T, M, H_0, H_1, E)$. For brevity of notation we assume that α, T, M, H_0, H_1 , and E are fixed and we will simply denote power for a specific $\mathbf{n}$ as $\pi(\mathbf{n})$. We use proc power in SAS [10] to calculate the $\pi(\mathbf{n})$. We will construct several weighted averages of the $\pi(\mathbf{n})$ to calculate power. The weights will be denoted as W with subscripts specifying the type of weighting function.

3.2. Problem Statement and Proposed Solution

We will consider two types of uncertainty: the uncertainty in the relative frequencies of the two groups (denoted as p in the previous example) and the allocation of the total N to the groups given p . We will address the latter issue first.

To address the uncertainty in sample sizes, we propose to use a weighted average of $\pi(\mathbf{n})$ over the possible $\mathbf{n}$. Let $f(\mathbf{n}|p)$ be the binomial (N, p) probability mass function. We propose to use

$$\begin{aligned} E_n[\pi(\mathbf{n})|p] &\equiv \sum_{\mathbf{n}} f(\mathbf{n}|p) \pi(\mathbf{n}) \\ &= \sum_{n_1=0}^N \binom{N}{n_1} p^{n_1} (1-p)^{N-n_1} \pi(\mathbf{n}) \\ &= \sum_{n_1=0}^N W_{exp} \pi(\mathbf{n}) \end{aligned} \tag{1}$$

which is the power given p .

Many studies will have excellent data on the population proportion p . For example, we know that the proportion of females is approximately 0.5. In many scenarios, we know much less about the population proportions. A prior distribution could be used to specify what we believe the unknown parameters to be and how uncertain we are about this belief. We could integrate $E_n[\pi(\mathbf{n})|p]$ over the prior distribution of p , $f(p|\theta)$, to derive an estimate of power that incorporates our uncertainty in the population proportion p . This calculation can be done in closed form if we use a beta prior as it is the conjugate prior for the binomial distribution. We will assume $p \sim \text{beta}(\theta_1, \theta_2)$ with $E(p) = \theta_1/(\theta_1 + \theta_2)$ and $\text{Var}(p) = \theta_1\theta_2/[(\theta_1 + \theta_2)^2(\theta_1 + \theta_2 + 1)]$.

The power accounting for uncertainty in p and $\mathbf{n}$ is

$$\begin{aligned}\pi(N) &\equiv E_p[E_n[\pi(n)|p]] \\ &\equiv \int_0^1 E_n[\pi(n)|p] f(p|\theta) dp \\ &\equiv \int_0^1 \left\{ \sum_{n_1=0}^N f(n|p) \pi(n) \right\} f(p|\theta) dp.\end{aligned}\quad (2)$$

This can be simplified by exchanging integration and summation, collecting powers of p and $(1-p)$, and calculating the integrals in closed form:

$$\begin{aligned}\pi(N) &= \frac{\Gamma(\theta_1 + \theta_2)}{\Gamma(\theta_1) \Gamma(\theta_2)} \sum_{n_1=0}^N \frac{\Gamma(n_1 + \theta_1) \Gamma(N - n_1 + \theta_2) \Gamma(N + 1)}{\Gamma(N + \theta_1 + \theta_2) \Gamma(n_1 + 1) \Gamma(N - n_1 + 1)} \pi(n) \\ &= \sum_{n_1=0}^N W_{Bayes} \pi(n)\end{aligned}\quad (3)$$

which is another weighted average of $\pi(n)$.

We have now accounted for uncertainty in n by averaging over the sample size distribution and accounted for uncertainty in the proportion p by averaging over the prior for p .

3.3. Example: Two-sample t-test

We first consider a small illustrative example where we are designing a study with $N = 6$ subjects and two groups; later we will examine realistic examples. We assume that both

groups are equally represented in the population. That is, $p = P(\text{group A}) = \frac{1}{2}$. In this scenario, $n_1 \sim \text{binomial}(6, 0.5)$ and $n_2 = N - n_1$. As n_1 increases from zero to six, $\pi(n)$ will increase and then decrease. If $n_1 = 0$ or 6 , then $\pi(n) = 0$. Assume the data are normally distributed with means μ_A and μ_B , common variance σ^2 , and that we use a two-sided two-sample t -test at the $\alpha = 0.05$ level to compare the means of groups A and B. Set $\delta = \mu_A - \mu_B = 4$ and $\sigma^2 = 1$. Table II gives the binomial probability of each sample size, W_{exp} , and the power, $\pi(n)$, for this comparison for all possible sample sizes.

Power can be calculated at $E[n]$ and is denoted $\pi(E[n])$. In this example, $E[n] = (3, 3)$ and $\pi(E[n]) = 0.9479$. The weights applied to each n for a fixed sample size are W_{fixed} (column 2). Then $\pi(E[n])$ is simply the sum of $W_{fixed} \pi(n)$ as presented in column 6. It is not uncommon for $\pi(E[n])$ to be presented as the power even though there may be considerable uncertainty in the realized sample size (n) and power may vary substantially as a function of n . In this case, $E[n]$ consists of integers. Most sample size formulae work even with non-integer sample sizes. Alternatively, $E[n]$ could be rounded to the nearest integer-valued vector.

The seventh column of Table II contains the products $W_{exp} \pi(n)$ and their sum is 0.8753. This is the probability of a positive test given that $p = 0.5$. Notice that this is smaller than the power of 0.9479 we would have if we were lucky and had equal sample sizes. This is due to the fact that the power in this example is a concave function of n and, for concave functions, $E[f(x)] < f(E[x])$. Expressed another way, we are more likely than not to have a sample size with power lower than the maximum. In short, $E_n[\pi(n)|p] < \pi(E[n])$, at least in this case.

For this example, we will assume that $p \sim \text{beta}(10, 10)$ which is plotted in Figure 1. With this prior, the weights (W_{Bayes}) in Equation 3 are presented in the fourth column of Table II and $\pi(N)$ is presented in the eighth column. The value of $\pi(N)$ is 0.8454 which is smaller

than both $\pi(E[\mathbf{n}])$ and $E_{\mathbf{n}}[\pi(\mathbf{n})|p]$. As in the calculation of $E_{\mathbf{n}}[\pi(\mathbf{n})|p]$, we have constructed a weighted average that averages over the possible designs. The weights W_{Bayes} are more diffuse than W_{exp} and W_{fixed} with more weight being assigned to designs with lower power, resulting in $\pi(N) < E_{\mathbf{n}}[\pi(\mathbf{n})|p]$. If we make the prior more diffuse using $\theta_1 = \theta_2 = 8, 6, 4, 2$, and 1 the power decreases to 0.8379, 0.8255, 0.8011, 0.7329, and 0.6217.

The example using $N = 6$ was very simple to demonstrate the proposed method. We will now present more realistic examples using the two-sample t -test with $N = 50, 100$, and 200 assuming a two-sided test at $\alpha = 0.05$, $\delta = 7/\sqrt{N}$, and $\sigma^2 = 1$. Power is presented for a variety of prior distributions in Table III. Within each of the first three sets of three rows, the prior variances decrease. In all three sets, $\pi(N) < E_{\mathbf{n}}[\pi(\mathbf{n})|p] < \pi(E[\mathbf{n}])$ and $\pi(N)$ approaches $E_{\mathbf{n}}[\pi(\mathbf{n})|p]$ as the prior variance decreases. This last point can be seen in the final set where the prior variance was very small and the calculated values of $\pi(N)$ and $E_{\mathbf{n}}[\pi(\mathbf{n})|p]$ are equal to within roundoff error. The differences between the various power estimates are smaller for these realistic sample sizes than they were for $N = 6$. In this example, the design matrix is random but the effect size is fixed. We will shortly examine an example where the effect size is random, too.

3.4. Generalizations

The beta prior distribution and the binomial distribution work well when there are only two groups of interest. However, many interesting problems have more than two groups. Examples include ANOVA (with 3 or more groups) and R-by-C tables. To address such problems, we can simply use the Dirichlet and multinomial distributions which are the multivariate generalizations of the beta and binomial. We will assume k groups and use the Dirichlet ($\boldsymbol{\theta}$) where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ and the multinomial ($N, \mathbf{p}$) where $\mathbf{p} = (p_1, \dots, p_k)$. If $p \sim$

Dirichlet ($\boldsymbol{\theta}$) then $E[p_i|\boldsymbol{\theta}] = \theta_i/\Theta$ where $\Theta = \sum_{i=1}^k \theta_i$. Also, $\text{Var}[p_i|\boldsymbol{\theta}] = \theta_i(\Theta - \theta_i)/[\Theta^2(\Theta + 1)]$ and $\text{Cov}[p_i, p_j|\boldsymbol{\theta}] = -\theta_i\theta_j/[\Theta^2(\Theta + 1)]$ [11].

The calculation of $E_{\mathbf{n}}[\pi(\mathbf{n})|p]$ is very similar to that presented in Equation 1 with the summation taken over $\mathbf{n} = (n_1, \dots, n_k)$ such that $n_1 + \dots + n_k = N$. Because the Dirichlet is the conjugate prior distribution for the multinomial, the calculation of the power accounting for uncertainty in $\mathbf{p}$ as in Equation 2 can also be calculated in closed form as in Equation 3 as

$$\pi(N) = \frac{\Gamma(\Theta)\Gamma(N+1)}{\Gamma(N+\Theta) \left[\prod_{i=1}^k \Gamma(\theta_i) \right]} \sum_{n_1+\dots+n_k=N} \left[\prod_{i=1}^k \frac{\Gamma(n_i+\theta_i)}{\Gamma(n_i+1)} \right] \pi(\mathbf{n}). \quad (4)$$

In both calculations, the major difference is that the summation is over $k - 1$ dimensions. (There are $k - 1$ degrees of freedom in $n_1 + \dots + n_k = N$). The resulting computational complexity is thus $O(N^{k-1})$.

3.5. Example: 2-by-2 Table, Random Effect Size

Consider the association of two binary factors. Once an experiment has been conducted, we would have cell counts in row-major order as $\mathbf{n} = (n_{11}, n_{12}, n_{21}, n_{22})$. In Subsection 3.3, we presented an example for a two-group comparison where the effect size was fixed but the design matrix was random. In this example, the effect size and distribution of observations among cells (essentially the design matrix but that term is not used in this context) are the same. Unlike with a two-group comparison, there is nothing stochastic once $\mathbf{n} = (n_{11}, n_{12}, n_{21}, n_{22})$ is known. Either the test is significant or it is not. Likewise, $\pi(E[\mathbf{n}])$ is 0 or 1.

Let $\mathbf{p} = (p_{11}, p_{12}, p_{21}, p_{22})$ be the proportions within each cell, also listed in row-major order. Assume we wish to compare proportions within columns (comparing within rows would be equivalent). The conditional proportions are $p_{11}/p_{\bullet 1}$ and $p_{12}/p_{\bullet 2}$ with expected group sizes of $Np_{\bullet 1}$ and $Np_{\bullet 2}$. Denote the power given for comparing these proportions with these group sizes as $\pi(N, \mathbf{p})$. This denotes power at a fixed set of probabilities $\mathbf{p}$ rather than at a fixed $\mathbf{n}$ and represents what might be done naively. In this paper we will use the Pearson χ^2 test although any other test of association for a 2-by-2 table could be used.

Although we have specified the cell proportions as a multinomial distribution, it is interesting to note that this could also be specified using a binomial distribution to specify column followed by two conditional binomial distributions to specify row within each column. That is, let $C \sim B(N, p_1)$ be the number in column 1 (treatment 1) with $N - C$ in column 2 (treatment 2). Here, p_1 is the probability of receiving treatment 1. Then assume that $R_{11}|C \sim B(C, p_2)$ is the number in row 1 (diseased) within column 1 and $R_{12}|C \sim B(N - C, p_3)$ is the number in row 1 (diseased) within column 2. The cell probabilities can then be expressed (in row-major order) as $p_1 p_2$, $(1 - p_1)p_3$, $p_1(1 - p_2)$, and $(1 - p_1)(1 - p_3)$. Using p_1 , p_2 , and p_3 is just an alternative way of parameterizing the four probabilities comprising $\mathbf{p}$ that must sum to 1.

We will specify our belief in $\mathbf{p}$ with a Dirichlet prior with parameter $\boldsymbol{\theta} = (\theta_{11}, \theta_{21}, \theta_{12}, \theta_{22})$. The odds ratio, $OR = (\theta_{11}\theta_{22}) / (\theta_{12}\theta_{21})$, can be used to express the strength of an association in a 2-by-2 table. A simple illustrative example for $N = 5$, $\boldsymbol{\theta} = (10, 5, 5, 10)$, and $\alpha = 0.05$ using a Pearson χ^2 test is presented in Table IV. As in Table II, W_{exp} and W_{Bayes} are the weights using the multinomial and Dirichlet distributions. When calculating Equations 1 and 4, we only have to sum over $\mathbf{n}$ where $\pi(\mathbf{n}) = 1$.

For more common designs with larger N , power is presented for a variety of N and $\boldsymbol{\theta}$ in Table V. As previously presented in Table III, $\pi(N)$ approaches $E_{\mathbf{n}}[\pi(\mathbf{n})|\mathbf{p}]$ as the prior variances decrease as can be seen within each of the first three sets of results. In the final set of results, the prior variances are very small and $E_{\mathbf{n}}[\pi(\mathbf{n})|\mathbf{p}]$ and $\pi(N)$ are virtually identical. Unlike as in Table III, the ordering of $\pi(N, \mathbf{p})$, $E_{\mathbf{n}}[\pi(\mathbf{n})|\mathbf{p}]$, and $\pi(N)$ is no longer as strict. Although this may be counterintuitive at first, it makes sense. In the standard calculation of power for the two-sample t -test, the *effect size* is fixed. However, in the calculation of $\pi(N)$ for a 2-by-2 table, the effect size is allowed to vary. When calculating Equation 2, we give some weight to $\boldsymbol{\theta}$ which specify large effect sizes. This is especially true when the Dirichlet prior is diffuse (elements of $\boldsymbol{\theta}$ are small).

3.6. Motivational Example Redux

We now return to the motivational example described in Section 2. When we first provided an answer to our colleague, we provided power accounting for the uncertainty in $\mathbf{n}$ as in $E_{\mathbf{n}}[\pi(\mathbf{n})|\mathbf{p}]$. We subsequently extended the results to account for the uncertainty in $\mathbf{p}$ as in $\pi(N)$. Power for a few sample sizes and prior parameters is presented in Table VI. Power was calculated using a two-sided χ^2 test at $\alpha = 0.05$. On the basis of these power calculations, we suggested $N = 162$ as this is the minimum sample size such that $E_{\mathbf{n}}[\pi(\mathbf{n})|\mathbf{p}] \geq 0.8$. To achieve power of 0.9, $N = 216$ is needed.

However, the choice of $\boldsymbol{\theta}$ has an impact on $\pi(N)$. Had we accounted for uncertainty in $\mathbf{p}$, we would likely have suggested a larger sample size. For example, using $\boldsymbol{\theta} = (3, 2, 4, 1)$, we would need $N = 351$ to have $\pi(N) \geq 0.8$ and $N = 1419$ to have $\pi(N) \geq 0.9$. These are many times what we calculated using the naive estimate $\pi(N, \mathbf{p})$ or even accounting for uncertainty in $\mathbf{n}$ using $E_{\mathbf{n}}[\pi(\mathbf{n})|\mathbf{p}]$. To be clear, the choice of $\boldsymbol{\theta} = (3, 2, 4, 1)$ is arbitrary and not driven by any pilot data. However, this choice reflects the expected relative distributions of the cell counts as presented in Table I.

3.7. Alternative Ways to Choose N

In the previous subsection, we suggested choosing N such that $\pi(N)$ was sufficiently large. Three other approaches might be considered. First, we could choose N to ensure that a design with high unconditional power is chosen with high probability. For example, pick N such that $\psi(0.9) = \Pr(\pi(\mathbf{n}) \geq 0.9) \geq 0.95$ where the probability is taken with respect to the prior distribution. We consider the role of 0.95 to be similar to that of the confidence coefficient in a confidence interval. That is, we want to assure that we have high power (e.g., 0.9) with high certainty (e.g., 0.95). For R-by-C tables, this would simply require that $\pi(N) \geq 0.95$ as $\pi(\mathbf{n})$ is either 0 or 1. For methods with continuous outcomes, $\psi(0.9)$ can be easily calculated by summing W_{Bayes} when $\pi(\mathbf{n}) \geq 0.9$. This approach could result in an increase in the required sample size, as demonstrated below. Consider, for example, a 3-group ANOVA with $\theta = (10, 10, 10)$, $\mu = (1, 2, 3)$, and $\sigma^2 = 1$. $N = 26$ will ensure $\pi(N) \geq 0.9$ but $\psi(0.9)$ is only 0.6970. To ensure $\psi(0.9) \geq 0.95$, we would need $N = 34$ which yields $\pi(N) = 0.9715$.

A second approach might be to sequentially recruit participants and ascertain group membership which defines $\mathbf{n}$. Without looking at the statistical test, recruitment would end when the power conditional on $\mathbf{n}$, $\pi(\mathbf{n})$, is sufficiently large. We do not pursue this here because we believe this approach to be less useful. Any study where group membership is ascertained in batch mode after recruitment could not use this approach. Additionally, this approach would require the continued involvement of a statistician to calculate $\pi(\mathbf{n})$ repeatedly. Finally, the temptation to “peek” at the results would be too tempting for many investigators leading to Type I error inflation.

A third approach would be to consider sample size re-estimation. For example, it may become apparent while collecting data that one group (e.g., males) is overrepresented. The sample size could then be adjusted accordingly.

3.8. Use of Pilot Data to Choose θ

In order to calculate $\pi(N)$, we will need an appropriate θ . In the absence of information, we can use $\theta = (c, \dots, c)$ for $0 \leq c \leq 1$ which are commonly used non-informative priors [12]. The uniform distribution for θ is specified when $c = 1$ and the uniform distribution for $\log(\theta)$

is specified when $c = 0$ [12]. The Jeffreys prior is specified when $c = \frac{1}{2}$ [13] and is the reference prior recommended by Berger and Bernardo for typical use [14]. A comprehensive review of choice of prior distributions was done by Kass and Wasserman [15].

Pilot data can be used to update a noninformative prior. For each observation in group i , increment θ_i by 1. For example, assume we have a pilot study with 10 subjects – 7 are observed with exposure and 3 without. We might then use $\text{beta}(0.5 + 7, 0.5 + 3)$ as the prior for calculation of power. If we had conducted a pilot study of 2-by-2 data with $N = 20$ and observed group sizes of 9, 5, 4, and 2 we might use a $\text{Dirichlet}(0.5 + 9, 0.5 + 5, 0.5 + 4, 0.5 + 2)$ as the prior. In practice, the choice of non-informative prior will not have a large impact if there is sufficient pilot data. For example, with $N = 10$, we might use prior distributions of $\theta = (7, 3)$, $(7.5, 3.5)$, or $(8, 4)$ which would result in $\pi(N) = 0.9068, 0.9236$, or 0.9355 (assuming a test at the 5% level, $N = 70$, $\mu_A - \mu_B = 1$, and $\sigma^2 = 1$). If the pilot study had twice as many subjects and the same proportions within the two groups, we might use $\theta = (14, 6)$, $(14.5, 6.5)$, or $(15, 7)$ which would result in $\pi(N) = 0.9348, 0.9410$, or 0.9461 . The estimates are more similar in the second set than in the first, as would be expected with more pilot data. We recommend using the middle estimates as they correspond to the Jeffreys prior recommended by Berger and Bernardo.

3.9. Implementation Details

SAS macros were used to perform the calculations. These macros simply calculate Equations 1, 3, and 4 by iterating over all possible designs and weighting $\pi(\mathbf{n})$ by W_{exp} and W_{Bayes} . Macros have been written for two- and three-group ANOVA, the χ^2 test, and Fisher's exact test; they are freely available at <http://www.phs.wfubmc.edu/public/wambrosi/RandomPower>. All examples in this paper took less than 7 seconds, except Table VI which took 31 minutes, on a moderately-priced 2009 laptop with a clock speed of 2.93 GHz running Windows XP.

4. Discussion

The techniques presented here are useful for observational studies and supplementary analyses for randomized trials when the group sizes cannot be specified *a priori*. They would not be useful for randomized trials as the group sizes are fixed. (Note, one could consider a randomized trial as a special case of the method proposed here by treating the fixed distribution to groups as a degenerate random vector.) A study comparing a grouping that is easily discernable (*e.g.*, gender) could use a recruitment strategy that ensured sufficiently many subjects within each group.

In this paper we described examples using a two-sample *t*-test and a χ^2 test for a 2-by-2 table. The same approach can be used for other designs and/or statistical tests. Further, other prior distributions could be used to calculate $\pi(N)$ but Equation ?? might need to be evaluated using simulation or numerical integration.

The differences between $\pi(E[\mathbf{n}])$ and $E_{\mathbf{n}}[\pi[\mathbf{n}]|p]$ tend to be greater for smaller N (approximately $N \leq 40$). For larger N , the law of large numbers ensures that $\mathbf{n} \approx E[\mathbf{n}]$ and, hence, the difference should be small. (Generally true for $N \geq 100$.) Note that this is not the case for $\pi(N)$ where a larger sample size does not necessarily provide protection against uncertainty.

We believe that many investigators have the calculation of ψ in mind when they choose N sufficiently large to ensure that $\pi(\mathbf{n})$ is high even under the “worst-case scenario.” However, this may result in substantially greater samples than would be required by using $\pi(N)$. When calculating power, we want to know the probability of a significant test statistic at the end of a study given an effect size, variability, size of test, etc. This is $\pi(N)$, the *a priori* probability of a statistically significant test. It is for this reason we believe that $\pi(N)$ is the quantity of true interest when designing a study. Using ψ may result in a larger $\pi(N)$. If this is needed, one could simply pick N to ensure that $\pi(N)$ is as high as desired.

Others have previously presented the idea of averaging power over a distribution of other unknown parameters [3,4,5,8]. In this paper, we have used this approach to account for random designs in observational studies. In the examples presented, ignoring uncertainty in sample sizes and group proportions can lead to incorrect power. For two-group comparisons, power is overestimated. For categorical analyses, power can be over or underestimated. The latter observation was made by others when calculating average power using a prior on the effect size (*e.g.*, [3,4,5,8]). Given that these calculations are easy to perform and can be done very quickly using freely available software, we strongly believe that the uncertainty in sample sizes and group proportions should be included in power calculations when the group sizes are random. This is similar in spirit to the commonly used practice of performing sensitivity analyses on the effect of unknown parameters (*e.g.*, the variance) on the calculation of power. A further extension of our method would be to also account for uncertainty in effect sizes with prior distributions.

Acknowledgments

We would like to thank Richard B. Weinberg, M.D., Section on Gastroenterology, Department of Internal Medicine, Wake Forest University School of Medicine, for the example which motivated this work and Timothy M. Morgan, Ph.D., Department of Biostatistical Sciences, Division of Public Health Sciences, Wake Forest University School of Medicine, for his helpful comments. We acknowledge the financial support of National Institutes of Health grant M01-RR07122 (WTA). Finally, we would like to thank the associate editor and reviewers whose many suggestions resulted in a substantially improved manuscript.

Contract/grant sponsor: NIH; contract/grant number: M01-RR07122

References

1. Julious SA. Tutorial in Biostatistics: Sample sizes for clinical trials with Normal data. *Statistics in Medicine*. 2004; 23:1921–1986.10.1002/sim.1783 [PubMed: 15195324]
2. Sahai H, Khurshid A. Formulae and Tables for the Determination of Sample Sizes and Power in Clinical Trials for Testing Differences in Proportions for the Two-Sample Design: A Review. *Statistics in Medicine*. 1996; 15:1–21.10.1002/(SICI)1097-0258(19960115)15:1<1::AID-SIM134>3.0.CO;2-E [PubMed: 8614740]
3. Spiegelhalter, DJ.; Abrams, KR.; Myles, JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley; Chichester: 2004. p. 193-194.
4. O'Hagan A, Stevens JW, Campbell MJ. Assurance in Clinical Trial Design. *Pharmaceutical Statistics*. 2005; 4:187–201.10.1002/pst.175
5. Fay MP, Halloran ME, Follmann DA. Accounting for Variability in Sample Size Estimation with Applications to Nonadherence and Estimation of Variance and Effect Size. *Biometrics*. 2007; 63:465–474.10.1111/j.1541-0420.2006.00703.x [PubMed: 17688499]
6. Julious S. Letter to the Editors of *Biometrics*. *Biometrics*. 2004; 60:284–285. [PubMed: 15032802]
7. Julious SA, Owen RJ. Sample size calculations for clinical studies allowing for uncertainty about the variance. *Pharmaceutical Statistics*. 2006; 5:29–37.10.1002/pst.197 [PubMed: 17080926]
8. Ambrosius WT, Lange EM, Langefeld CD. Power for Genetic Association Studies with Random Allele Frequencies and Genotype Distributions. *American Journal of Human Genetics*. 2004; 74:683–693. [PubMed: 15024689]
9. Benson PW, Hooker J, Weinberg RB. Bitter taster status predicts susceptibility to vection-induced nausea. *Neurogastroenterology & Motility*. 2009; 21(Supplement 1):15–16.10.1111/j.1365-2982.2009.01343_2.x
10. SAS Institute, Inc.. *SAS/STAT® 9.2 User's Guide*. SAS Institute; Cary, NC: 2008. p. 4809-5039.
11. Kotz, S.; Balakrishnan, N.; Johnson, NL. *Continuous Multivariate Distributions, Volume 1: Models and Applications*. Second. Wiley; New York: 2000. p. 488
12. Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. *Bayesian Data Analysis*. Chapman & Hall/CRC; Boca Raton, FL: 1995. p. 399p. 76and
13. Agresti, A. *Categorical Data Analysis*. Second. Wiley; Hoboken, NJ: 2002. p. 607
14. Berger JO, Bernardo JM. Ordered Group Reference Priors with Application to the Multinomial Problem. *Biometrika*. 1992; 79:25–37.
15. Kass RE, Wasserman L. The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association*. 1996; 91:1343–1370.

Figure 1.
Beta(10, 10) distribution.

Table I

Hypothesized proportions of people for combination of nausea susceptibility and ability to taste bitter tastes from example in Section 2.

	Bitter Taste Receptor Status	
	Yes	No
Susceptible to Nausea–Yes	0.3	0.2
No	0.4	0.1

Table II

Table of possible sample sizes for the first example in Subsection 3.3 with $\delta = 4$, $\sigma^2 = 1$, and $\alpha = 0.05$.

n	W_{fixed}	W_{exp}	W_{Bayes}	$\pi(n)$	$W_{\text{fixed}}\pi(n)$	$W_{\text{exp}}\pi(n)$	$W_{\text{Bayes}}\pi(n)$
(0,6)	0	0.0156	0.0283	0	0	0	0
(1,5)	0	0.0938	0.1130	0.7778	0	0.0729	0.0879
(2,4)	0	0.2344	0.2220	0.9243	0	0.2166	0.2052
(3,3)	1	0.3125	0.2733	0.9479	0.9479	0.2962	0.2591
(4,2)	0	0.2344	0.2220	0.9243	0	0.2166	0.2052
(5,1)	0	0.0938	0.1130	0.7778	0	0.0729	0.0879
(6,0)	0	0.0156	0.0283	0	0	0	0
					0.9479	0.8753	0.8454

Power calculations for the second example described in Subsection 3.3 with $\delta=7/\sqrt{N}$, $\sigma^2=1$, and $\alpha = 0.05$.

Table III

Parameters						
N	θ	p	E [n]	π (E [n])	$E_n [\pi (n) p]$	π (N)
50	(1,1)	0.5	(25, 25)	0.9291	0.9240	0.7163
50	(5,5)	0.5	(25, 25)	0.9291	0.9240	0.8889
50	(10,10)	0.5	(25, 25)	0.9291	0.9240	0.9082
100	(1,1)	0.5	(50, 50)	0.9339	0.9316	0.7305
100	(5,5)	0.5	(50, 50)	0.9339	0.9316	0.8994
100	(10,10)	0.5	(50, 50)	0.9339	0.9316	0.9173
200	(1,1)	0.5	(100, 100)	0.9361	0.9350	0.7375
200	(5,5)	0.5	(100, 100)	0.9361	0.9350	0.9043
200	(10,10)	0.5	(100, 100)	0.9361	0.9350	0.9215
50	(10 ⁴ , 10 ⁴)	0.5	(25, 25)	0.9291	0.9240	0.9240
100	(10 ⁴ , 10 ⁴)	0.5	(50, 50)	0.9339	0.9316	0.9316
200	(10 ⁴ , 10 ⁴)	0.5	(100, 100)	0.9316	0.9350	0.9350

Table IV
Selected rows demonstrating power calculations for simple 2-by-2 table described in Subsection 3.5.

n								
n ₁₁	n ₁₂	n ₂₁	n ₂₂	π (n)	W _{exp}	W _{Bayes}	W _{exp} π (n)	W _{Bayes} π (n)
0	0	0	5	0	0.0041	0.0072	0	0
0	0	1	4	0	0.0103	0.0128	0	0
0	0	2	3	0	0.0103	0.0119	0	0
⋮								
0	1	4	0	1	0.0006	0.0013	0.0006	0.0013
⋮								
2	0	0	3	1	0.0412	0.0435	0.0412	0.0435
⋮								
							0.1273	0.1447

Table V

Power calculations for example described in Subsection 3.5.

Parameters					
N	θ	OR	π (N, p)	$E_n [\pi(n) p]$	π (N)
35	$(16, 1, 1, 2) \times 1$	32	0.9058	0.8968	0.7802
35	$(16, 1, 1, 2) \times 2$	32	0.9058	0.8968	0.8288
35	$(16, 1, 1, 2) \times 4$	32	0.9058	0.8968	0.8590
90	$(7, 1, 1, 1) \times 1$	7	0.9126	0.9109	0.6853
90	$(7, 1, 1, 1) \times 2$	7	0.9126	0.9109	0.7270
90	$(7, 1, 1, 1) \times 4$	7	0.9126	0.9109	0.7778
130	$(4, 1, 1, 1) \times 1$	4	0.9182	0.9158	0.7130
130	$(4, 1, 1, 1) \times 2$	4	0.9182	0.9158	0.7126
130	$(4, 1, 1, 1) \times 4$	4	0.9182	0.9158	0.7357
35	$(16, 1, 1, 2) \times 10^4$	32	0.9058	0.8968	0.8968
90	$(7, 1, 1, 1) \times 10^4$	7	0.9126	0.9109	0.9108
130	$(4, 1, 1, 1) \times 10^4$	4	0.9182	0.9158	0.9157

Table VI

Power calculations for the motivational example, Subsection 3.6.

Parameters					
N	θ	OR	π (N, p)	$E_n [\pi(n) p]$	π (N)
161	(3,2,4,1)	0.375	0.7987	0.7985	0.7082
162	(3,2,4,1)	0.375	0.8012	0.8011	0.7090
215	(3,2,4,1)	0.375	0.9002	0.8992	0.7461
216	(3,2,4,1)	0.375	0.9015	0.9010	0.7468
350	(3,2,4,1)	0.375	0.9859	0.9853	0.7999
351	(3,2,4,1)	0.375	0.9862	0.9855	0.8002
351	(3, 2, 4, 1) $\times$ 10	0.375	0.9862	0.9855	0.8538
351	(3, 2, 4, 1) $\times$ 100	0.375	0.9862	0.9855	0.9699
1418	(3, 2, 4, 1)	0.375	1.0000	1.0000	0.8999
1419	(3, 2, 4, 1)	0.375	1.0000	1.0000	0.9000