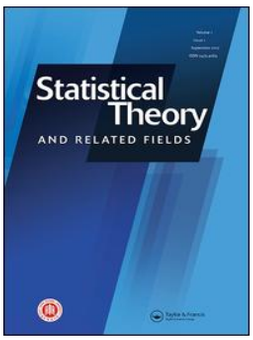

Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Power analysis, sample size calculation for testing the largest binomial probability

Thuan Nguyen & Jiming Jiang

To cite this article: Thuan Nguyen & Jiming Jiang (2019): Power analysis, sample size calculation for testing the largest binomial probability, Statistical Theory and Related Fields, DOI: 10.1080/24754269.2019.1586283

To link to this article: https://doi.org/10.1080/24754269.2019.1586283

Published online: 15 Mar 2019.

STATISTICAL THEORY AND RELATED FIELDS
https://doi.org/10.1080/24754269.2019.1586283

Taylor & Francis
Taylor & Francis Group

Power analysis, sample size calculation for testing the largest binomial probability

Thuan Nguyen[a,b] and Jiming Jiang[a,b]

[a]Oregon Health and Science University, Portland, OR, USA; [b]University of California, Davis, CA, USA

ABSTRACT

A procedure is developed for power analysis and sample size calculation for a class of complex testing problems regarding the largest binomial probability under a combination of treatments. It is shown that the asymptotic null distribution of the likelihood-ratio statistic is not parameter-free, but χ^2_1 is a conservative asymptotic null distribution. A nonlinear Gauss-Seidel algorithm is proposed to uniquely determine the alternative for the power and sample size calculation given the baseline binomial probability. An example from an animal clinical trial is discussed.

ARTICLE HISTORY
Received 31 October 2018
Accepted 20 February 2019

KEYWORDS
Asymptotic null distribution; binomial probability; complex hypotheses; Gauss-Seidel; logistic regression; power; sample size; tests

1. Introduction

In biological and medical research, it is often necessary to perform power analysis, or determine the sample size required for binomial trials involving multiple treatments. For example, such an analysis/calculation is required by the National Institutes of Health (NIH) for research grant applications. One particular type of questions that need to be answered is regarding the largest binomail probability under these treatments. Suppose that there are r treatments, denoted by $1, \ldots, r$. Here the 'treatment' can be a real treatment (e.g., drug), or a factor that has two levels (e.g., male/female). Let x_q be the indicator for the qth treatment (0 or 1), $1 \le q \le r$. Suppose that n independent trials are run under each combination of the treatments, $x_1, \ldots, x_r$, resulting a binary outcome for each trial (1 – success, 0 – failure). Let $N(x_1, \ldots, x_r)$ be the total number of successes under $x_1, \ldots, x_r$. It is assumed that $N(x_1, \ldots, x_r)$ has a Binomia$\{n, p(x_1, \ldots, x_r)\}$ distribution, where $p(x_1, \ldots, x_r)$ is the probability of success in a single trial under $x_1, \ldots, x_r$. It is believed that all the treatments have at least nonnegative effects, so $p(0, \ldots, 0)$ is the smallest probability and $p(1, \ldots, 1)$ is the largest. The question is to determine the sample size, n, so that one has at least the power $1 - \gamma$ to prove the case that $p(1, \ldots, 1)$ is higher than the rest of the probabilities, if it is indeed higher by at least as much as δ. In some cases, the researcher has a target sample size (e.g., the maximum under the budget constraint). The question then is to perform the power analysis for the given sample size. We illustrate with an example.

Example 1.1: Researchers at the Oregon Health & Science University (OHSU) were preparing to meet the deadline of a grant submission. One of the research aims of the grant proposal had to do with comparative transplantations via animal clinical trials (i.e., mice) to determine the best overall protocol, which can then be further optimised. Successful liver repopulation is defined as greater than 70% cell replacement and successful blood reconstitution is defined as greater than 50% human cells in the bone marrow. From previous studies, it was known that these levels of liver repopulation could be achieved in 20–80% of transplanted mice with a good hepatocyte donor, the average being 50–60%. For cord blood transplants into neonates, about 75% of the mice reach the desired human repopulation levels, again with a range of 50–90% between experiments. Using these numbers from the single cell type transplants, we made the assumption that a successful protocol would yield about 30% success in double-repopulation. Here the protocol involved a treatment high and low dose levels, and a control factor of youth and grown-up mice. In other words, there are two treatments, $x_1 = 1$ for high dose and $x_1 = 0$ for low dose; $x_2 = 1$ for youth mice and $x_2 = 0$ for grown-up one. The trials were to be carried out independently on n different mice, resulting a binomial proportion, under each protocol. It was determined that n should be no more than 30 due to the budget constraint. An initial request was made to perform a power analysis in order to detect a 10% difference that separates the success rates of the 'optimal' protocol (i.e., with high dose and youth mice) and the rest.

CONTACT Jiming Jiang jimjiang@ucdavis.edu

A naive approach to the power analysis/sample size calculation would be to compare $p(1,\ldots,1)$ with each of the other probabilities (actually, only those with exactly one of the treatment indicators equal to zero), and perform a two-sample t-test for the difference in proportions. However, this approach is low power, and often results in a sample size that the researcher cannot afford. The inefficiency of the naive approach is not surprising, because only a (small) portion of the data are involved in the two-sample t-test (for each comparison). One can do better by utilising the entire data in the analysis. To do so we need to assume a model for the binomial probability. Suppose that

$$p(x_1,\ldots,x_r) = h(\beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r), \qquad (1)$$

where $h(\cdot)$ satisfies $0 < h(x) < 1$ and is strictly increasing. A well-known example of h is the logistic function, $h(x) = \mathrm{e}^x/(1+\mathrm{e}^x)$. Under the assumed model, the problem of interest can be expressed more precisely as testing the hypothesis

$$\mathrm{H}_0 : \text{at least one of the } \beta_j, \quad 1 \le j \le r, \text{ is } \le 0 \qquad (2)$$

versus $\mathrm{H}_1 : \beta_j > 0,\ 1 \le j \le r$. Natually, the likelihood-ratio test is considered. The latter is based on the test statistic

$$\mathcal{L} = 2(\hat{l} - \hat{l}_0), \qquad (3)$$

where $\hat{l}$ is the maximised log-likelihood and $\hat{l}_0$ the maximised log-likelihood under H_0. A first step for the likelihood-ratio test (LRT) would be to determine the critical value, c_α, corresponding to the given level of significance α, such that

$$\sup_{\beta\in\mathrm{H}_0} \mathrm{P}_\beta(\mathcal{L} > c_\alpha) \le \alpha, \qquad (4)$$

where $\beta = (\beta_0,\ldots,\beta_r)'$ and P_β is the probability distribution given that β is the true vector of parameters. If the log-likelihood function is log-concave, which is the case, for example, for the logistic regression, the event inside the probability of (4) implies that the maximum for $\hat{l}_0$ must take place on the boundary of H_0, provided that $c_\alpha > 0$. Therefore, the critical value is computed by considering the boundary of H_0, which is a subset of

$$\tilde{\mathrm{H}}_0 : \beta_j = 0 \quad \text{for some } 1 \le j \le r. \qquad (5)$$

Unfortunately, even with the latest simplification, the asymptotic null distribution of the LRT is not parameter-free. This is shown in the next section. On the other hand, the arguments also show that a conservative asymptotic null distribution (CAND) for the LRT is χ^2_1, which is parameter-free. Here the CAND is in the sense that

$$\sup_{\beta\in\tilde{\mathrm{H}}_0} \limsup_{n\to\infty} \mathrm{P}_\beta(\mathcal{L} > \chi^2_{1,\alpha}) = \alpha. \qquad (6)$$

The main objective of the current paper is to determine the sample size, n, for the LRT so that the test will have the designated power, or to obtain the power of the LRT, under a given sample size. Although standard power and sample size problems in logistic regression are well studied (e.g., Alam, Rao, & Cheng, 2010; Borenstein, Rothstein, & Cohen, 2001; Demidenko, 2007; Hsieh, Bloch, & Larsen, 1998; Novikov, Fund, & Freedman, 2010; Whittemore, 1981), to the best of our knowledge, the kind of problems that we are dealing with have not been addressed. Clearly, the power of the test depends on the alternative, and there are infinitely many possible alternatives to (2), that is, in H_1. On the other hand, a practitioner would prefer a 'short answer', as opposed to something that is case-by-case. This issue is addressed in Section 3, where a unique alternative, based on a reasonable argument, is determined. A simple Gauss-Seidel type algorithm is proposed to compute the alternative. A Monte Carlo procedure is then proposed to compute the power or sample size. A real-life application is considered in Section 4. Technical details are deferred to Appendix.

2. Asymptotic null distribution

In the standard situation, the LRT is known to have an asymptotic χ^2 distribution, under the null hypothesis, with a certain degrees of freedom that does not depend on the parameter under the null. For example, if, instead of (2), one were to test

$$\mathrm{H}_0^j : \beta_j = 0 \qquad (7)$$

versus $\mathrm{H}_1^j : \beta_j \ne 0$ for a fixed $1 \le j \le r$, then the asymptotic distribution of the LRT is χ^2_1, regardless of the value of the true $\beta_k, k \ne j$, as long as the true β_j is zero. In other words, the asymptotic null distribution is parameter-free. However, this is not the case for testing (2). We show this for the case of logistic regression with $r = 2$.

Write the log-likelihood function as $l(\beta_0,\beta_1,\beta_2)$. Let $(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2)'$, $(\hat{\beta}_0^{[2]},\hat{\beta}_1^{[2]},0)'$, and $(\hat{\beta}_0^{[1]},0,\hat{\beta}_2^{[1]})'$ be the maximiser of l without constraint, over $\mathrm{H}_0^2 = \{(\beta_0,\beta_1,0) : \beta_0,\beta_1 \in R\}$, and over $\mathrm{H}_0^1 = \{(\beta_0,0,\beta_2) : \beta_0,\beta_2 \in R\}$, respectively. Suppose that the true parameter vector is $(\beta_0,\beta_1,0) \in \mathrm{H}_0^2$, where $\beta_1 \ne 0$. By the standard asymptotic theory, we have $\hat{\beta}_j^{[2]} \xrightarrow{\mathrm{P}} \beta_j, j = 0,1$, as $n \to \infty$. On the other hand, by White (1982), we have $\hat{\beta}_j^{[1]} \xrightarrow{\mathrm{P}} \beta_j^{[1]}, j = 0,2$, as $n \to \infty$ for some $\beta_0^{[1]}$ and $\beta_2^{[1]}$. Thus, by the Taylor expansion, we have

$$\frac{l(\hat{\beta}_0^{[2]},\hat{\beta}_1^{[2]},0)}{n} = \frac{l(\beta_0,\beta_1,0)}{n} + \frac{1}{n}\sum_{j=0,1}\left\{\frac{\partial l}{\partial\beta_j}\bigg|_{\tilde{\beta}^{(2)}}\right\}\left(\hat{\beta}_j^{[2]} - \beta_j\right), \qquad (8)$$

$$\frac{l(\hat{\beta}_0^{[1]},0,\hat{\beta}_2^{[1]})}{n} = \frac{l(\beta_0^{[1]},0,\beta_2^{[1]})}{n}$$

$$+\frac{1}{n}\sum_{j=0,2}\left\{\left.\frac{\partial l}{\partial\beta_j}\right|_{\tilde\beta^{(1)}}\right\}\left(\hat\beta_j^{[1]}-\beta_j^{[1]}\right), \qquad (9)$$

for some $\tilde\beta^{(j)}$, $j=1,2$. It is easy to show that the partial derivatives are uniformly bounded when divided by n, so the second terms on the right sides of (8) and (9) are $o_{\mathrm{P}}(1)$. As for the first terms, it is easy to show that, for any β, we have

$$l(\beta)=c+\sum_{x_1,x_2=0,1}\{N(x_1,x_2)(\beta_0+\beta_1x_1+\beta_2x_2) -n\log(1+\mathrm{e}^{\beta_0+\beta_1x_1+\beta_2x_2})\},$$

where c does not depend on the parameter. Thus, we have

$$\frac{l(\beta_0,\beta_1,0)-l(\beta_0^{[1]},0,\beta_2^{[1]})}{n} =\sum_{x_1,x_2=0,1}\left[\frac{N(x_1,x_2)}{n}\{(\beta_0+\beta_1x_1)-(\beta_0^{[1]} +\beta_2^{[1]}x_2)\}-\log\left(\frac{1+\mathrm{e}^{\beta_0+\beta_1x_1}}{1+\mathrm{e}^{\beta_0^1+\beta_2^1x_2}}\right)\right].$$

By the weak law of large numbers, we have $n^{-1}N(x_1,x_2)\xrightarrow{\mathrm{P}}p(x_1,x_2)=h(\beta_0+\beta_1x_1)$, with $h(x)=\mathrm{e}^x/(1+\mathrm{e}^x)$, $x_1,x_2=0,1$. Thus, we have

$$\frac{l(\beta_0,\beta_1,0)-l(\beta_0^{[1]},0,\beta_2^{[1]})}{n} \xrightarrow{\mathrm{P}}\sum_{x_1,x_2=0,1}\left[\frac{\mathrm{e}^{\beta_0+\beta_1x_1}}{1+\mathrm{e}^{\beta_0+\beta_1x_1}}\{(\beta_0+\beta_1x_1) -(\beta_0^{[1]}+\beta_2^{[1]}x_2)\}-\log\left(\frac{1+\mathrm{e}^{\beta_0+\beta_1x_1}}{1+\mathrm{e}^{\beta_0^{[1]}+\beta_2^{[1]}x_2}}\right)\right]. \qquad (10)$$

For fixed $x_1,x_2\in\{0,1\}$, write $p_0=h(\beta_0+\beta_1x_1)$ and $p_1=h\left(\beta_0^{[1]}+\beta_2^{[1]}x_2\right)$. Then, by the inequality in Appendix A.1, we have

$$\frac{\mathrm{e}^{\beta_0+\beta_1x_1}}{1+\mathrm{e}^{\beta_0+\beta_1x_1}}\{(\beta_0+\beta_1x_1)-(\beta_0^{[1]}+\beta_2^{[1]}x_2)\} -\log\left(\frac{1+\mathrm{e}^{\beta_0+\beta_1x_1}}{1+\mathrm{e}^{\beta_0^{[1]}+\beta_2^{[1]}x_2}}\right) =p_0\log\left(\frac{p_0}{p_1}\right)+(1-p_0)\log\left(\frac{1-p_0}{1-p_1}\right)\geq 0,$$

with the equality holding if and only if $p_0=p_1$. It follows that the right side of (10) is positive unless $p_0=p_1$ for all $x_1,x_2=0,1$. Because the latter implies $\beta_1=0$, a contradiction, the right side of (10) must be positive. Therefore, in conclusion, we have with probability tending to one that $l(\hat\beta_0^{[2]},\hat\beta_1^{[2]},0)>l(\hat\beta_0^{[1]},0,\hat\beta_2^{[1]})$, hence $\mathcal{L}=2\{l(\hat\beta_0,\hat\beta_1,\hat\beta_2)-l(\hat\beta_0^{[2]},\hat\beta_1^{[2]},0)\}\xrightarrow{\mathrm{d}}\chi_1^2$, by the standard asymptotic result [see the note below (7)].

Now suppose that the true parameter vector is $(\beta_0,0,0)$. Then, it can be shown (see Appendix A.2) that $\mathcal{L}\xrightarrow{\mathrm{d}}\eta-\eta_1\vee\eta_2$, as $n\to\infty$, where

$$\eta=\begin{pmatrix}\xi_0\\ \xi_1\\ \xi_2\end{pmatrix}'\begin{pmatrix}4&2&2\\2&2&1\\2&1&2\end{pmatrix}^{-1}\begin{pmatrix}\xi_0\\ \xi_1\\ \xi_2\end{pmatrix},$$

$$\begin{pmatrix}\xi_0\\ \xi_1\\ \xi_2\end{pmatrix}\sim N\left[\begin{pmatrix}0\\0\\0\end{pmatrix},\begin{pmatrix}4&2&2\\2&2&1\\2&1&2\end{pmatrix}\right],$$

$$\eta_j=\begin{pmatrix}\xi_0\\ \xi_j\end{pmatrix}'\begin{pmatrix}4&2\\2&2\end{pmatrix}^{-1}\begin{pmatrix}\xi_0\\ \xi_j\end{pmatrix},\quad j=1,2.$$

Note that $\eta-\eta_1\vee\eta_2$ is not distributed as χ_1^2. To see this, note that $\eta_1<\eta_2$ if and only if $(\xi_1-\xi_0/2)^2<(\xi_2-\xi_0/2)^2$. Because the pdf of ξ is positive everywhere, there is a positive probability that $\eta_1<\eta_2$, hence $\eta_1\vee\eta_2>\eta_1$. On the other hand, one always has $\eta_1\vee\eta_2\geq\eta_1$. It follows that $\mathrm{E}(\eta_1\vee\eta_2)>\mathrm{E}(\eta_1)$, hence $\mathrm{E}(\eta-\eta_1\vee\eta_2)<\mathrm{E}(\eta)-\mathrm{E}(\eta_1)=3-2=1$.

The results so far in this section have shown that the asymptotic null distribution of $\mathcal{L}$ depends on the values of the true parameters, namely, if $\beta=(\beta_0,\beta_1,0)$, where $\beta_1\neq0$ [or $\beta=(\beta_0,0,\beta_2)$, where $\beta_2\neq0$, by a similar argument], the asymptotic null distribution is χ_1^2; if $\beta=(\beta_0,0,0)$, the asymptotic null distribution is not χ_1^2. Thus, the asymptotic null distribution is not parameter-free.

Nevertheless, χ_1^2 is, in general (not restricted to logistic and $r=2$), a CAND in the sense of (6), provided that the log-likelihood is log-concave. This can be shown with a simple argument. For any $\beta\in\tilde{\mathrm{H}}_0$, there is a $1\leq j\leq r$ such that $\beta_j=0$. Due to the log-concavity, the event $\mathcal{L}>\chi_{1,\alpha}^2$ is the same as the event $2(\hat l-\tilde l_0)>\chi_{1,\alpha}^2$, where $\tilde l_0$ is the maximised log-likelihood over $\tilde{\mathrm{H}}_0$ [see the note above (5)]. On the other hand, we have $2(\hat l-\tilde l_0)\leq2(\hat l-\hat l_{0j})$, where $\hat l_{0j}$ is the maximised log-likelihood over $\tilde{\mathrm{H}}_{0j}=\{\beta:\beta_j=0\}$. Therefore, we have $\mathrm{P}_\beta(\mathcal{L}>\chi_{1,\alpha}^2)=\mathrm{P}_\beta\{2(\hat l-\tilde l_0)>\chi_{1,\alpha}^2\}\leq\mathrm{P}_\beta\{2(\hat l-\hat l_{0j})>\chi_{1,\alpha}^2\}\to\alpha$, by the standard asymptotic result. Therefore, we have $\limsup_{n\to\infty}\mathrm{P}_\beta(\mathcal{L}>\chi_{1,\alpha}^2)\leq\alpha$.

On the other hand, if $\beta\in\tilde{\mathrm{H}}_0$ such that $\beta_j=0$ while $\beta_k\neq0,k\neq j$, by a similar argument as the above for the special case of logistic regression with $r=2$, it can be shown that, with P_β tending to one, we have $\mathcal{L}=2(\hat l-\hat l_{0j})\xrightarrow{\mathrm{d}}\chi_1^2$, hence $\lim_{n\to\infty}\mathrm{P}_\beta(\mathcal{L}>\chi_{1,\alpha}^2)=\alpha$. Because $\lim_{n\to\infty}\mathrm{P}_\beta(\mathcal{L}>\chi_{1,\alpha}^2)$ achieves its supremum at $\beta_j=0$ while $\beta_k\neq0,k\neq j$, (6) must hold. Note that, by the definition of $\tilde H_0$, which has $j\geq1$, all the indexes j,k mentioned in this paragraph are assumed to be ≥1; in other words, β_0 is not involved.

3. Power and sample size calculation

By the results of the previous section, we can use $\chi^2_{1,\alpha}$ as the critical value of the LRT. As for the power calculation, although there are infinitely many alternatives that influence the power, it is often reasonable to assume, in practice, that the baseline probability is known. In other words, $h(\beta_0)$ is known according to (1); therefore, β_0 is known.

In addition, the minimum probabilistic increase of the largest probability over the other probabilities, δ, is given. In other words, we consider all the alternatives such that

$$p(1,\ldots,1) \geq p(x_1,\ldots,x_r) + \delta \tag{11}$$

for all $(x_1,\ldots,x_r) \neq (1,\ldots,1)$. It follows, under (1), that all of the $\beta_j, 1 \leq j \leq r$ must be positive, and that (11) is equivalent to

$$h\left(\beta_0 + \sum_{j=1}^{r} \beta_j\right) \geq h\left(\beta_0 + \sum_{1\leq j\leq r, j\neq k} \beta_j\right) + \delta, \quad 1 \leq k \leq r. \tag{12}$$

The minimum amount of increase of the left sides of (12) over the right sides takes place when the equalities hold in all of the inequalities, that is, when

$$h\left(\beta_0 + \sum_{j=1}^{r} \beta_j\right) = h\left(\beta_0 + \sum_{1\leq j\leq r, j\neq k} \beta_j\right) + \delta, \quad 1 \leq k \leq r. \tag{13}$$

This results in r equations, from which we can uniquely determine the alternative. Note that (13) is a nonlinear equation system; however, it can be solved conveniently by utilising a Gauss-Seidel type algorithm (e.g., Jiang, 2000). Namely, from (13) we have

$$\begin{aligned}\beta_k &= h^{-1}\left\{h\left(\beta_0 + \sum_{1\leq j\leq r, j\neq k} \beta_j\right) + \delta\right\} \\ &\quad - \beta_0 - \sum_{1\leq j\leq r, j\neq k} \beta_j \\ &= g\left(\beta_0 + \sum_{1\leq j\leq r, j\neq k} \beta_j\right),\end{aligned} \tag{14}$$

where $g(x) = h^{-1}\{h(x) + \delta\} - x$ (the inverse of h exists because h is assumed to be strictly increasing). Thus, given the initial values $\beta_j^{(0)}, 1 \leq j \leq r-1$ (e.g., all zero), we have

$$\beta_k^{(l)} = g\left\{\beta_0 + \sum_{j=1}^{k-1} \beta_j^{(l-1)} + \sum_{j=k+1}^{r} \beta_j^{(l)}\right\}, \quad k = r,\ldots,1 \tag{15}$$

Table 1. Convergence of Gauss-Seidel algorithm.

	Initial value: $\beta_1 = 0$		Initial value: $\beta_1 = 1$	
Iteration	β_1	β_2	β_1	β_2
0–1	0.0000000	0.4418328	1.0000000	0.4144315
1–2	0.4054651	0.4070466	0.4066332	0.4069919
2–3	0.4069726	0.4069760	0.4069751	0.4069759
3–4	0.4069759	0.4069759	0.4069759	0.4069759
4–5	0.4069759	0.4069759	0.4069759	0.4069759

for $l = 1, 2, \ldots$. The convergence is guaranteed, and fast. We illustrate with an example.

Example 3.1 (continued): In the animal clinical trial example, the researchers suggested a baseline probability of 0.3. Using the logistic regression, we have $\beta_0 = \text{logit}(0.3) = -0.8472979$. Furthermore, the minimum probability increase was set (again by the researchers) as $\delta = 0.1$. Recall that $r = 2$ in this case. The Gauss-Seidel algorithm converged within three iterations. Table 1 shows the R outputs of the first five iterations.

Given the target sample size, the power of the LRT at the alternative can be computed by a Monte Carlo method. Consider, for example, the case of logistic regression. Under the alternative $\beta = (\beta_0, \beta_1, \ldots, \beta_r)'$, one can simulate data, $N(x_1,\ldots,x_r), x_1,\ldots,x_r \in \{0,1\}$, under the logistic regression. For each simulated data set, $r+1$ logistic regressions are fit. The first one is under the full model, the next one under the model without $x_1,\ldots,$ and the last one under the model without x_r. Let $\hat{l}, \hat{l}_{01},\ldots,\hat{l}_{0r}$ denote the maximised log-likelihoods as results of these logistic regressions. We compute $\mathcal{L} = 2(\hat{l} - \max_{1\leq j\leq r} \hat{l}_{0j})$ for the simulated data set. This is repeated B times, resulting $\mathcal{L}^{(b)}, b = 1,\ldots,B$. The power of the LRT is then approximated by $B^{-1}\#\{1 \leq b \leq B : \mathcal{L}^{(b)} > \chi^2_{1,\alpha}\}$.

If, instead, the goal is to determine the sample size so that the LRT has a designated power, say, γ, we can use the following bisection procedure to speed up the search for the minimum sample size. First pick a couple of initial sample sizes, n_0 and n_1, and compute the power under n_0 and n_1 using the above procedure. Suppose that the power under n_0 is less than γ, and the power under n_1 is greater than γ. We then let $n_2 = (n_0 + n_1)/2$ (take the integer part, if necessary), and compute the power under n_2 using the above procedure. If the power under n_2 is greater than γ, let $n_3 = (n_0 + n_2)/2$; otherwise, let $n_3 = (n_2 + n_1)/2$, and so on. The procedure should converge quickly to a single integer, n_*, so that either n_* or $n_* + 1$ is the minimum sample size to have the power greater than or equal to γ.

4. Animal clinical trial revisited

Let us go back to Example 1.1 of Section 1. Recall the initial request was to make a power analysis based on the sample size $n = 30$ for detecting a 10% difference

between $p(1,1)$ and the rest of the binomial probabilities, and the baseline probability was set as 30%. We considered a logistic regression with $r=2$ for this case. Here $p_0 = 0.3$ and $\delta = 0.1$. The alternative was computed by the Gauss-Seidel algorithm [see Example 1.1 (continued) in Section 3] as $\beta_0 = -0.8472979$ and $\beta_1 = \beta_2 = 0.4069759$. The corresponding probabilities are $p(0,0) = 0.3, p(0,1) = p(1,0) = 0.3916643$, and $p(1,1) = 0.4916643$. As we can see, the minimum difference between $p(1,1)$ and the rest of the p's is 0.1. For this alternative, the power of the LRT at 5% level of significance was computed, using the Monte-Carlo method described in Section 3 with $B=1000$, as approximately 88%.

As the 80% power was considered satisfactory by the researchers, it appeared that the sample size might be reduced a little. However, when the same procedure was applied to $n=20$, the power was computed as approximately 78%. In communicating with the lead researcher, the researcher suggested that he would rather sacrifice a little regarding the minimum probabilistic difference in exchange for a reduced sample size (i.e., $n=20$). Thus, the new δ was set as 0.15. The new alternative was computed by the Gauss-Seidel algorithm (which, again, converged in three iterations) as $\beta_0 = -0.8472979$ and $\beta_1 = \beta_2 = 0.6051085$. The corresponding probabilities are $p(0,0) = 0.3, p(0,1) = p(1,0) = 0.4397469$, and $p(1,1) = 0.5897469$. As we can see, the minimum probabilistic increase of $p(1,1)$ over the rest of the p's is 0.15. For the new alternative, the power of the LRT at 5% level of significance was computed, again using the Monte-Carlo method with $B=1000$, as approximately 86%. The researcher was satisfied with the result.

Acknowledgments

The authors are grateful to Dr. Markus Grompe of the Doernbecher Children's Hospital of the Oregon Health & Science University for presenting the problem from their research, and for information and helpful discussions. The authors also wish to thank a reviewer for helpful comments.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The authors' research is partially supported by the National Institutes of Health (NIH) grant R01-GM085205A1. In addition, Thuan Nguyen's research is partially supported by the National Science Foundation (NSF) grant SES-1118469; Jiming Jiang's research is partially supported by the National Science Foundation (NSF) grant SES-1121794.

Notes on contributors

Thuan Nguyen is Associate Professor, Department of Public Health and Preventive Medicine, Oregon Health and Science University, USA.

Jiming Jiang is Professor, Department of Statistics, University of California, Davis, USA.

References

Alam, M. K., Rao, M. B., & Cheng, F.-C. (2010). Sample size determination in logistic regression. *Sankhyā B*, *72*, 58–75.

Borenstein, M., Rothstein, H., & Cohen, J. (2001). *Power and precision*. Englewood, US: Biostat Inc.

Demidenko, E. (2007). Sample size determination for logistic regression revisited. *Statistics in Medicine*, *26*, 3385–3397.

Hsieh, F. Y., Bloch, D. A., & Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine*, *17*, 1623–1634.

Jiang, J. (2000). A nonlinear Gauss-Seidel algorithm for inference about GLMM. *Computational Statistics*, *15*, 229–241.

Jiang, J. (2010). *Large sample techniques for statistics*. New York: Springer.

Novikov, I., Fund, N., & Freedman, L. S. (2010). A modified approach to estimating sample size for simple logistic regression with one continuous covariate. *Statistics in Medicine*, *29*, 97–105.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25.

Whittemore, A. (1981). Sample size for logistic regression with small response probability. *Journal of the American Statistical Association*, *76*, 27–32.

Appendix

A.1 An inequality

Consider the function $g(p) = p^{p_0}(1-p)^{1-p_0}$ and $h(p) = \log\{g(p)\}, 0 < p < 1$. We have $h'(p) = \{p(1-p)\}^{-1}(p_0 - p)$. Thus, $h'(p) > 0, = 0$, or < 0 depending on $p < p_0$, $p = p_0$, or $p > p_0$. It follows that $h(\cdot)$, hence $g(\cdot)$, has a unique maximum at $p = p_0$, and $g(p) < g(p_0)$ for any $p \neq p_0$. Thus, for any $0 < p_1 < 1, p_1 \neq p_0$, we have

$$\left(\frac{p_0}{p_1}\right)^{p_0}\left(\frac{1-p_0}{1-p_1}\right)^{1-p_0} = \frac{g(p_0)}{g(p_1)} > 1.$$

A.2 Some derivation in Section 2

We show that $\mathcal{L} \xrightarrow{\mathrm{d}} \eta - \eta_1 \vee \eta_2$ as $n \to \infty$, where the η's are defined in Section 2, if $(\beta_0, 0, 0)$ is the true parameter vector. With the notation introduced in Section 2, we have, by the Taylor expansion,

$$\begin{aligned} l(\beta_0,0,0) &= l(\hat\beta_0^{[2]}, \hat\beta_1^{[2]}, 0) \\ &\quad + \frac{1}{2}\begin{pmatrix} \beta_0 - \hat\beta_0^{[2]} \\ -\hat\beta_1^{[2]} \end{pmatrix}' \\ &\quad \times \begin{pmatrix} \partial^2 l/\partial\beta_0^2 & \partial^2 l/\partial\beta_0\partial\beta_1 \\ \partial^2 l/\partial\beta_1\partial\beta_0 & \partial^2 l/\partial\beta_1^2 \end{pmatrix}\Bigg|_{(\hat\beta_0^{[2]}, \hat\beta_1^{[2]}, 0)} \\ &\quad \times \begin{pmatrix} \beta_0 - \hat\beta_0^{[2]} \\ -\hat\beta_1^{[2]} \end{pmatrix} \end{aligned}$$

$+o_{\mathrm{P}}(1)$, implying

$$\begin{aligned} l(\hat\beta_0^{[2]}, \hat\beta_1^{[2]}, 0) &= l(\beta_0, 0, 0) \\ &\quad - \frac{1}{2}\begin{pmatrix} \hat\beta_0^{[2]} - \beta_0 \\ \hat\beta_1^{[2]} \end{pmatrix}' \end{aligned}$$

$$\mathrm{E}\left\{\begin{pmatrix}\partial^2 l/\partial\beta_0^2 & \partial^2 l/\partial\beta_0\partial\beta_1\\ \partial^2 l/\partial\beta_1\partial\beta_0 & \partial^2 l/\partial\beta_1^2\end{pmatrix}\Big|_{(\beta_0,0,0)}\right\}\begin{pmatrix}\hat\beta_0^{[2]}-\beta_0\\ \hat\beta_1^{[2]}\end{pmatrix}$$

$+o_{\mathrm{P}}(1)$. Also, by the standard asymptotic expansion (e.g., Jiang, 2010, Ch. 4), we have

$$\begin{pmatrix}\hat\beta_0^{[2]}-\beta_0\\ \hat\beta_1^{[2]}\end{pmatrix}=\left[\mathrm{E}\left\{\begin{pmatrix}\partial^2 l/\partial\beta_0^2 & \partial^2 l/\partial\beta_0\partial\beta_1\\ \partial^2 l/\partial\beta_1\partial\beta_0 & \partial^2 l/\partial\beta_1^2\end{pmatrix}\Big|_{(\beta_0,0,0)}\right\}\right]^{-1}\times\begin{pmatrix}\partial l/\partial\beta_0\\ \partial l/\partial\beta_1\end{pmatrix}\Big|_{(\beta_0,0,0)}+O_{\mathrm{P}}(n^{-1}).$$

It follows that

$$\begin{aligned}l(\hat\beta_0^{[2]},\hat\beta_1^{[2]},0)&=l(\beta_0,0,0)\\&\quad-\frac{1}{2}\begin{pmatrix}\partial l/\partial\beta_0\\ \partial l/\partial\beta_1\end{pmatrix}\Big|'_{(\beta_0,0,0)}\\&\quad\left[\mathrm{E}\left\{\begin{pmatrix}\partial^2 l/\partial\beta_0^2 & \partial^2 l/\partial\beta_0\partial\beta_1\\ \partial^2 l/\partial\beta_1\partial\beta_0 & \partial^2 l/\partial\beta_1^2\end{pmatrix}\Big|_{(\beta_0,0,0)}\right\}\right]^{-1}\\&\quad\times\begin{pmatrix}\partial l/\partial\beta_0\\ \partial l/\partial\beta_1\end{pmatrix}\Big|_{(\beta_0,0,0)}+o_{\mathrm{P}}(1).\end{aligned}$$

Similarly, we have

$$\begin{aligned}l(\hat\beta_0^{[1]},0,\hat\beta_2^{[1]})&=l(\beta_0,0,0)\\&\quad-\frac{1}{2}\begin{pmatrix}\partial l/\partial\beta_0\\ \partial l/\partial\beta_2\end{pmatrix}\Big|'_{(\beta_0,0,0)}\\&\quad\left[\mathrm{E}\left\{\begin{pmatrix}\partial^2 l/\partial\beta_0^2 & \partial^2 l/\partial\beta_0\partial\beta_2\\ \partial^2 l/\partial\beta_2\partial\beta_0 & \partial^2 l/\partial\beta_2^2\end{pmatrix}\Big|_{(\beta_0,0,0)}\right\}\right]^{-1}\\&\quad\times\begin{pmatrix}\partial l/\partial\beta_0\\ \partial l/\partial\beta_2\end{pmatrix}\Big|_{(\beta_0,0,0)}+o_{\mathrm{P}}(1),\end{aligned}$$

and

$$\begin{aligned}l(\hat\beta_0,\hat\beta_1,\hat\beta_2)&=l(\beta_0,0,0)\\&\quad-\frac{1}{2}\frac{\partial l}{\partial\beta}\Big|'_{(\beta_0,0,0)}\left\{\mathrm{E}\left(\frac{\partial^2 l}{\partial\beta\partial\beta'}\Big|_{(\beta_0,0,0)}\right)\right\}^{-1}\frac{\partial l}{\partial\beta}\Big|_{(\beta_0,0,0)}\\&\quad+o_{\mathrm{P}}(1).\end{aligned}$$

Furthermore, we have

$$\mathrm{E}\left(\frac{\partial^2 l}{\partial\beta\partial\beta'}\Big|_{(\beta_0,0,0)}\right)=-nh'(\beta_0)\begin{pmatrix}4&2&2\\2&2&1\\2&1&2\end{pmatrix},\qquad(\mathrm{A1})$$

and, by the standard asymptotic result,

$$\frac{1}{\sqrt{n}}\frac{\partial l}{\partial\beta}\Big|_{(\beta_0,0,0)}\xrightarrow{\mathrm{d}}N\left[\begin{pmatrix}0\\0\\0\end{pmatrix},\ h'(\beta_0)\begin{pmatrix}4&2&2\\2&2&1\\2&1&2\end{pmatrix}\right],\qquad(\mathrm{A2})$$

Let $\xi_n=(\xi_{n,0},\xi_{n,1},\xi_{n,2})'$ denote the left side of (A2) divided by $\sqrt{h'(\beta_0)}$, and A denote the 3×3 matrix in (A1). For $a=(a_0,a_1,a_2)'$ and $A=(a_{st})_{s,t=0,1,2}$, let $a[0,j]$ and $A[0,j]$ denote the subvector $(a_0,a_j)'$ and submatrix $(a_{st})_{s,t=0,j}$, respectively, $j=1,2$. Then, by the above expressions, it is easy to show that

$$\mathcal{L}=\xi_n'A^{-1}\xi_n-\max_{j=1,2}\{\xi_n[0,j]'A[0,j]^{-1}\xi_n[0,j]\}+o_{\mathrm{P}}(1).$$

Because $\xi_n\xrightarrow{\mathrm{d}}\xi=(\xi_0,\xi_1,\xi_2)'\sim N(0,A)$, as $n\to\infty$, by the continuous mapping theorem (e.g., Jiang, 2010, p.30), we have $\mathcal{L}\xrightarrow{\mathrm{d}}\eta-\eta_1\vee\eta_2$, where $\eta=\xi'A\xi$ and $\eta_j=\xi[0,j]'A[0,j]^{-1}\xi[0,j]$, $j=1,2$.