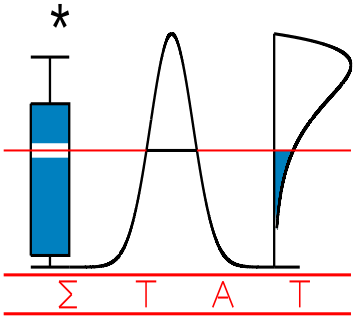
T E C H N I C A L

R E P O R T

0505

POWER AND SAMPLE SIZE CALCULATIONS FOR DISCRETE BOUNDED OUTCOME SCORES

TSONAKA, R., RIZOPOULOS, D. and E. LESAFFRE



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Power and Sample Size Calculations for Discrete Bounded Outcome Scores

Roula Tsonaka,[1,*] **Dimitris Rizopoulos**[1] **and Emmanuel Lesaffre**[1]

[1]Biostatistical Centre, Catholic University Leuven,

U.Z. St. Rafaël, Kapucijnenvoer 35, B–3000 Leuven, Belgium

June 29, 2006

Summary. We consider power and sample size calculations for a randomized controlled clinical trial with a bounded outcome score as primary response adjusted for a priori chosen covariates. A bounded outcome score (BOS) is a random variable that is restricted to a finite interval. Examples of a BOS are found in health-related quality of life research, e.g., the SF-36 being the most commonly used score or the Barthel-index often used in stroke trials. There are two popular ways to analyze such data: as a numeric score or as an ordinal variable. We will take the first approach here, and treat it as a special case of grouped or coarse data. Often such scores have J- or U-shaped distributions hindering traditional parametric methods of analysis. When no adjustment for covariates is needed, a non-parametric test could be chosen. However, there is still a problem with calculating the power and sample size with classical approaches since the common location-shift alternative does not hold in general for a BOS. In this paper, we consider a parametric approach and assume that the observed BOS is a coarsened version of a true BOS, which has a logit-normal distribution in each treatment group allowing correction for covariates. A two-step procedure is used to calculate the power function. First, the power function is calculated conditionally on the realized covariate values. Secondly, the marginal power is obtained by averaging the conditional power with respect to an assumed distribution for the covariates using Monte Carlo integration. This procedure provides also a practical method for sample size calculations. A simulation study evaluates the performance of

* *email:* spyridoula.tsonaka@med.kuleuven.be

our method. When the BOS is regarded as ordinal, ordinal logistic regression (OLR) is a valid and popular approach to compare the efficacy of two treatments while accounting for covariates. We will indicate briefly the connection between our approach and OLR. But, more importantly we will indicate that our approach can also be implemented with some minor modifications to calculate the power and sample size for OLR when continuous covariates are involved. Finally, as an illustration of our method, we perform a re-sample size calculation on the primary outcome (Barthel index) of the ECASS-1 study, a stroke trial designed to compare the effect of placebo and a thrombolytic drug on patients with an acute ischemic stroke.

KEY WORDS: bounded outcome scores; grouped data; Barthel index; power; sample size calculations; Wald statistic

1. Introduction

A bounded outcome score (BOS) is a random variable that is restricted to a finite interval and can be continuous, discrete or a mix of a continuous and a discrete variable. In this paper we will concentrate on discrete bounded outcome scores and treat them as a special case of the grouped or coarse data framework as it has been formalized by (Heitjan, 1989; Heitjan and Rubin, 1991; Heitjan, 1993). Grouped data arise from a variable whose true values are known only up to subsets of the sample space and this can happen in many ways, e.g., rounding, interval censoring or censoring of continuous variables into categories. We note, in passing, that the coarsening mechanism not only can be quite general but can also differ between subjects. Examples of discrete BOSs include the various Quality of Life indexes like the Barthel-index, the SF-36 score, etc. Here we will exemplify our approach on the Barthel-index, which is an Activity on Daily Living (ADL) scale with (in one version) a minimal value of 0 (death or completely immobilized) and a maximal value of 100 (able to perform all daily activities independent) jumping with steps of 5. This scale is often used in stroke trials to measure the recovery of a patient in practical terms after an acute stroke. The visual analogue scale is another important example of a BOS. Formally, it is a continuous variable on (0,1) which attains

the boundary values 0 and 1 with non-zero probability. However, coarsened or rounded versions of this scale are often reported and in this case our proposal for power and sample size calculations for discrete BOSs will also be adequate. For reasons of simplicity we will assume that a discrete BOS takes values in the closed interval $[0, 1]$.

An important aspect in setting up a clinical trial is the estimation of the power (and sample size) to detect a clinically relevant treatment effect. A BOS can have a peculiar shape of distribution, i.e., J- or U-shaped and thus non-parametric tests such as the Wilcoxon test are likely to be used in this context. If covariate adjustment is aimed at and the covariate is categorical, then the van Elteren test (van Elteren, 1960) could be used, but for a continuous covariate it is not clear how to handle adjustment in a non-parametric way. In addition, when pilot data are available the bootstrap method (Efron and Tibshirani, 1993) provides a valuable non-parametric alternative that easily allows for covariate adjustments (Collings and Hamilton, 1988; Hamilton and Collings, 1991; Walters and Campbell, 2005). Moreover, for large sample sizes the Central Limit Theorem can be used and the power calculations can be performed with conventional methods (Walters and Campbell, 2005). However, in the BOS context and for skewed distributions the boundaries might be crossed under the alternative and thus the appropriateness of the conventional methods might be questionable. Recently, (Lesaffre et al., 2006) have proposed to use the logistic transformation to model BOSs and assume a parametric expression for the distribution of a latent BOS on (0,1) which is coarsely measured giving rise to the observed BOS on [0,1]. This approach allows easily for covariate adjustment. This is important as the power of detecting a treatment effect will be increased considerably if the baseline covariates are well chosen, i.e., correlate well with the response.

One of the traditional methods for calculating the power of a statistical test assume the location shift alternative (LSA). LSA assumes that under the alternative hypothesis the distribution of the primary endpoint is shifted with a fixed amount Δ for, say, the active treatment with respect to the control treatment. However, the LSA assumption is not very appealing for a BOS. Indeed, since the score is restricted to, say, the interval

$[0, 1]$, the LSA assumption could imply that under the alternative hypothesis one of the distributions has to cross the boundaries which is not valid. This has also been recognized by (Lesaffre et al., 1993; Lesaffre and de Klerk, 2000; Tsodikov et al., 1998) who proposed alternative approaches for power and sample size calculations but none of the proposed methods accounts for covariate adjustment. A typical alternative to the LSA is to assume that the values of one group tend to be larger than the other, such as in the Wilcoxon test (Noether, 1987; Walters and Campbell, 2005). However, quantifying such an alternative may not often be a trivial task.

In this paper we propose a parametric method for power and sample size calculations under LSA, in a transformed scale, in discrete bounded responses when baseline covariates are included in the model. The paper is organized as follows. In Section 2 we briefly review the idea of using the logistic transformation to model a BOS. In Section 3 we examine the properties of the Wald test for detecting a treatment effect in this context. In Section 4 we present formulas for power and sample size calculation. The power function is calculated in two steps: in the first step the power function is calculated conditionally on the covariate values and in the second stage this power function is averaged over an assumed distribution for the covariates using Monte Carlo integration. For sample size calculations, Monte Carlo sampling may render the computations time consuming. Thus, an approximate method is proposed using a Taylor series expansion to get a good initial estimate for the sample size. This initial estimate is then used to provide a more narrower search area to the algorithm used resulting in this way to much faster sample size calculations. In Section 5 a few simulation studies are described, which were set up to (i) explore the distributional properties for the Wald statistic; (ii) demonstrate the increase in power by adjusting for covariates; (iii) show the sensitivity of the marginal power calculations to mis-specification of the covariate distribution and (iv) evaluate the proposed methods under a number of scenarios. In Section 6 we relate our approach to OLR and show that, while based on a slightly different philosophy, our method of power and sample size calculation can easily be adapted to OLR. Moreover, we illustrate that the current approach already gives a precise estimate of the power and sample size for

OLR when allowing for covariates. In Section 7 we apply our approach to the ECASS-1 study, which is an early placebo-controlled randomized clinical trial evaluating the effect of a thrombolytic drug on patients with an acute ischemic stroke. Finally, in Section 8 we summarize our findings and discuss further research in calculating power (and sample size) with bounded outcomes.

2. The logistic transformation for discrete bounded scores

The logistic transformation to handle a discrete BOS $Y \in [0, 1]$ was suggested by (Lesaffre et al., 2006) following the approach of (Aitchison and Shen, 1980). The advantage of this approach is its flexibility in capturing various shapes of the distribution of the observed data (unimodal, J- or U-shaped) while allowing the incorporation of baseline covariates. The key idea is that a continuous latent variable U on $(0, 1)$ is coarsely measured giving rise to the observed discrete outcomes Y on $[0, 1]$. This latent variable is assumed to have, on the logit scale, a classical distribution; that is $Z = \text{logit}(U)$ can be normal, Student's-t or a logistic distribution. Here we focus on the normal case and in particular we assume that $\text{logit}(U) \sim N(\mu, \sigma^2)$ and denote that $U \sim LN(\mu, \sigma^2)$, where $LN(\mu, \sigma^2)$ is the logit-normal distribution with parameters μ and σ^2. Depending on the choice of μ, σ^2 we achieve different shapes of the distribution.

More specifically, the observed Y arises when the continuous latent variable U lies in one of the disjoint intervals of the form $[(a_l, b_l)]$ (a_l, b_l are real numbers in $[0, 1]$ and $[(.,.)]$ denotes the four possible kinds of intervals) with $\bigcup_l [(a_l, b_l)] = [0, 1]$. Different procedures generating the subintervals $[(a_l, b_l)]$ imply different types of coarsening mechanisms (see e.g., (Heitjan, 1993)). Hence, the probability for the ith individual ($i = 1, \ldots, n$) to have a score y_i equals:

$$P(Y_i = y_i; \boldsymbol{\theta}) = \int_{a_i}^{b_i} g(u_i; \boldsymbol{\theta})\, du_i = \Phi\left(\frac{z_i^{(u)} - \mu}{\sigma}\right) - \Phi\left(\frac{z_i^{(l)} - \mu}{\sigma}\right) \tag{1}$$

where $g(.; \boldsymbol{\theta})$ is the probability density function of the logit-normal distribution with parameters $\boldsymbol{\theta} = (\mu, \sigma)^T$, $\Phi(.)$ is the distribution function of the standard normal distribution and $z_i^{(l)} = \text{logit}(a_i)$, $z_i^{(u)} = \text{logit}(b_i)$. For more details on this approach for analyzing a BOS we refer to (Lesaffre et al., 2006). Moreover, it is obvious that (1) is in fact a

special case of a grouped data likelihood where $g(.;\boldsymbol{\theta})$ can be any density function and $z_i^{(l)} = h_i(a_i)$ and $z_i^{(u)} = h_i(b_i)$ with $h_i(.)$ any monotonic differential function. As a result, our proposals described in detail in Section 4 are still valid under different coarsening mechanisms for the different subjects in the same study.

Under this framework, the LSA assumption can be applied on the logit scale by assuming that the distribution of logit (U) under H_0 is $N(\mu, \sigma^2)$, while under the alternative it is $N(\mu + \Delta, \sigma^2)$ (see Figure 1). Thereby, the boundary restriction of the Y responses can be effectively handled. If logit (U) follows a logistic distribution, this assumption implies that the log-odds ratio in the original u-scale is constant and equal to the effect size Δ/σ (for more on the assumption of the logistic distribution see Section 6). In the other cases we can at least say that $\Delta_1 < \Delta_2$ implies that on the original scale the distribution under the first alternative hypothesis is stochastically smaller than under the second alternative hypothesis. Additionally, for the transformed normal distributions, the proportion of individuals better off with the new treatment than with the control treatment is equal to $P_\Delta = \Phi\left(\frac{\Delta}{\sigma\sqrt{2}}\right)$ and because the logistic transformation is monotone the same property holds on the original scale.

[Figure 1 about here.]

Finally, when covariate adjustment is envisaged, the mean parameter μ can be replaced by the linear predictor

$$\eta_i = \gamma_0 + \Delta t_i + \boldsymbol{\gamma}_1^T \boldsymbol{x}_i, \tag{2}$$

where γ_0 denotes the intercept, t_i the treatment indicator with treatment effect measured by Δ and $\boldsymbol{x}_i$ the extra baseline covariates with corresponding coefficients $\boldsymbol{\gamma}_1$.

3. Test Statistic

Under model (1) and assuming the linear predictor (2), the Wald test is used to detect a relevant treatment effect while correcting for the covariates $\boldsymbol{x}_i$. The Wald statistic to test the null hypothesis $H_0 : \Delta = \Delta_0 (= 0)$ versus the alternative hypothesis $H_a : \Delta =$

Δ_a, where Δ_a is the hypothesized value of the treatment effect under the alternative hypothesis, is defined as:

$$W = \frac{\hat{\Delta} - \Delta_0}{\hat{\sigma}_{\hat{\Delta}}}, \tag{3}$$

where $\hat{\Delta}$ is the MLE of the treatment effect and $\hat{\sigma}_{\hat{\Delta}}$ is an estimate of the standard error of $\hat{\Delta}$ (based on either the observed information or Fisher information matrix of model (1)). Usually, the asymptotic normality of the MLEs is used to claim the normality of the Wald statistic. However, since in model (1) the scale parameter σ is unknown, we assume a Student's-t approximation for W. A simulation study (described in Section 5.1) indicates that, even for relatively small sample sizes, W is well approximated by a Student's-t distribution with $n - p$ degrees of freedom under the null hypothesis (where $n = n_1 + n_2$ is the sample size and p is the rank of the design matrix under the assumed linear predictor), while under the alternative, the distribution of W is often well approximated by a non-central Student's-t distribution with $n - p$ degrees of freedom and non-centrality parameter $\delta = \Delta_a/\sigma_{\hat{\Delta}}$, where $\sigma_{\hat{\Delta}}$ is the standard error of $\hat{\Delta}$. In addition, the properties and the robustness of the Wald test has been explored in the context of ordinal responses arising from quality of life indexes (Heeren and D'Agostino, 1987; Sullivan and D'Agostino, 2003). Moreover, the beneficial effect of covariate adjustments on the power of the Wald statistic is explored in Section 5.2. Finally, the method proposed in this paper can be easily applied for statistics other than the Wald (i.e., score, likelihood ratio statistic, etc.) with small modifications.

4. Calculating power and sample size

4.1 *Conditional power*

A two-stage procedure is used to derive the (marginal) power function of the Wald statistic defined in Section 3 under model (1). Initially we calculate the conditional power of the statistic, namely the power of the statistic given the realized values of the covariates. Thus we first assume that the design matrix $X = (1, t, x_1, \ldots, x_{p-2})$, where $x_1, \ldots, x_{p-2}$ are possibly additional baseline covariates, is known. This corresponds to

a randomized experiment where the results are known and interest lies in the power of detecting a treatment effect of known magnitude given the realized covariate values. The calculation of the conditional power serves as a basis for the calculation of the marginal power in Section 4.2. For notational simplicity we consider here, in addition to the treatment indicator, only one covariate x (i.e., $X = (1, t, x)$ is of rank $p = 3$). However, the inclusion of more than one baseline covariate is straightforward.

Under the Student's-t approximation discussed in Section 3, the power of the Wald test for detecting a treatment effect equal to Δ_a while taking into account additional baseline covariates is calculated by the expression (Armitage and Berry, 1987):

$$\begin{aligned} 1 - \beta_C \equiv 1 - \beta_C(X) &= P(|W| > t_{\nu,1-\alpha/2} \mid H_a; X) \\ &= 1 - F_{t_{\nu,\delta}}(t_{\nu,1-\alpha/2} \mid H_a; X) + F_{t_{\nu,\delta}}(t_{\nu,\alpha/2} \mid H_a; X), \end{aligned} \tag{4}$$

where $F_{t_{\nu,\delta}}$ is the distribution function of the non-central Student's-t distribution with $\nu = n - p$ degrees of freedom and non-centrality parameter $\delta = \Delta_a/\sigma_{\hat{\Delta}}$, $t_{\nu,\alpha/2}$ is the $\alpha/2$ quantile of the central Student's-t distribution with ν degrees of freedom, α is the two-sided type I error and β_C is the conditional (on the observed covariate values) probability of making a type II error.

The Fisher Information matrix of the logit-normal model (1) delivers an estimate of $\sigma_{\hat{\Delta}}$. For alternatives that do not differ too much from the null hypothesis, $\sigma_{\hat{\Delta}}$ is assumed to be practically the same under the null and the alternative hypothesis (Noether, 1987). The Fisher Information matrix is needed here since we have to average over the possible response values. Lengthy calculations lead to the following compact expression for the Fisher Information matrix under model (1):

$$I = E_{Y|X}\left(-\begin{bmatrix} \frac{\partial^2 \ell}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}} & \frac{\partial^2 \ell}{\partial \boldsymbol{\gamma} \partial \sigma} \\ \frac{\partial^2 \ell}{\partial \sigma \partial \boldsymbol{\gamma}} & \frac{\partial^2 \ell}{\partial \sigma \partial \sigma} \end{bmatrix}\right) = \begin{bmatrix} X^T W^{(1)} X & X^T W^{(2)} \\ X^T W^{(2)} & \sum_{i=1}^n w_i^{(3)} \end{bmatrix},$$

where $W^{(1)}$, $w^{(2)}$ and $w_i^{(3)}$ are defined in Appendix A, and X^T denotes the transpose of X.

4.2 *Marginal power*

At the design stage of a study the values of the independent variables are usually not available. Only pilot or historical data may be available from which reasonable distributional assumptions for the covariates can be made. In this case, we suggest to calculate the marginal power as the expected conditional power with respect to the distribution of the baseline covariates (i.e., here the treatment indicator and one extra baseline covariate). Thus the marginal power is given by:

$$1 - \beta_M = 1 - \sum_{\mathcal{T}} \int_{\mathcal{X}} \beta_C(X) dH(X), \tag{5}$$

with $H(X)$ denoting the cdf of the joint distribution of the covariates. Under the setup considered here the density $h(X)$ of $H(X)$ is written as $h(X) = p(t)f(x)$ where $f(x)$ is the density function of the covariate x, $p(t)$ is a Bernoulli(π) representing the allocation rate in a randomized trial with two treatment groups and $\mathcal{X}$, $\mathcal{T}$ are their corresponding sample spaces. For more than one additional baseline covariate, the unidimensional integral in (5) becomes a multidimensional integral or a sum. In general, the expectation in (5) does not have an analytic solution since the independent variables appear in the standard error $\sigma_{\hat{\Delta}}$ of the denominator of the non-centrality parameter δ. However, an approximation of the above expectation is easily accomplished using Monte Carlo integration. In particular, (5) can be approximated by:

$$p_M = 1 - \beta_M \approx 1 - \frac{1}{B}\sum_{b=1}^{B} F_{t_{\nu,\delta}}(t_{\nu,1-\alpha/2} \mid H_a; X^{(b)}) + \frac{1}{B}\sum_{b=1}^{B} F_{t_{\nu,\delta}}(t_{\nu,\alpha/2} \mid H_a; X^{(b)}), \tag{6}$$

where B in the number of the Monte Carlo simulations and $X^{(b)}$ is a realization of the vectors t and x from Bernoulli(π) and $N(\mu_X, \sigma_X^2)$, respectively. The precision of the estimated marginal power equals $\sigma_{\hat{p}_M} = \sigma_{p_C}/\sqrt{B}$, where σ_{p_C} is the standard error in the conditional power calculation. The desired level of this precision determines the choice for B and thus it is advisable to make an initial choice for B (e.g., 50) and then adjust it accordingly.

Finally, note that the (marginal) power depends not only on the hypothesized regression coefficients $\boldsymbol{\gamma} = (\gamma_0, \Delta, \gamma_1)^T$, but also on the residual standard deviation, $\sigma_{U|X}$, for

the linear regression of the latent variable U on X. In some cases it might be difficult to postulate a value for $\sigma_{U|X}$ whereas postulations for the values of the variance of the response U and the covariate X in the two treatment groups might be more feasible. In such a case, based on linear regression theory (Armitage and Berry, 1987) we get the formulas:

$$\sigma^2_{U|X} = \sigma^2_{U|t,x} = \frac{1}{(n_1 + n_2 - p)} \left(n_1\sigma^2_{u_1} + n_2\sigma^2_{u_2} - \frac{(n_1\sigma_{ux_1} + n_2\sigma_{ux_2})^2}{n_1\sigma^2_{x_1} + n_2\sigma^2_{u_2}} \right) \tag{7}$$

and

$$\gamma_1 = \frac{n_1\sigma_{Ux_1} + n_2\sigma_{Ux_2}}{n_1\sigma^2_{x_1} + n_2\sigma^2_{x_2}}, \tag{8}$$

where for any two continuous random variables Q and R and any binary random variable S, $\sigma^2_{Q_i}$ and σ_{QR_i} denote the variance of Q and the covariance between Q and R respectively at the ith level of S.

4.3 *Sample size calculation*

The necessary sample size to detect a treatment effect with a given marginal power (nominal power) can be estimated using expression (6). Namely, the sample size n needed for a power p^0_M is the root of the following function:

$$f(n) = p_M(n) - p^0_M, \tag{9}$$

with respect to n, where $n = n_1 + n_2$ is the total sample size and $p_M(n)$ is in fact (6) as a function of n.

To solve expression (9) for n, repeated evaluations of $p_M(n)$ are needed and hence also repeated Monte Carlo integrations, which becomes quite involved especially when the target interval for the optimal n is too wide. However, a rough initial estimate of n could be obtained using an approximation to the likelihood thereby narrowing the target interval. The likelihood under the logit-normal model is given by:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \int_{a_i}^{b_i} g\left(u; \boldsymbol{\theta}\right) du = \prod_{i=1}^{n} \left[G\left(b_i; \boldsymbol{\theta}\right) - G\left(a_i; \boldsymbol{\theta}\right) \right], \tag{10}$$

where $G(.)$ is the distribution function of the logit-normal distribution with parameters $\mu = \boldsymbol{\gamma}^T \boldsymbol{x}_i$ and σ^2. A first order Taylor series expansion in both $G(a_i; \boldsymbol{\theta})$ and $G(b_i; \boldsymbol{\theta})$ around the mid-point of the ith coarsening interval, i.e., $\zeta_i = \frac{a_i + b_i}{2}$, gives:

$$G(a_i; \boldsymbol{\theta}) \approx G(\zeta_i; \boldsymbol{\theta}) + (a_i - \zeta_i)\, g(\zeta_i; \boldsymbol{\theta}),$$

$$G(b_i; \boldsymbol{\theta}) \approx G(\zeta_i; \boldsymbol{\theta}) + (b_i - \zeta_i)\, g(\zeta_i; \boldsymbol{\theta}).$$

Thus, likelihood (10) can be approximated by $L(\boldsymbol{\theta}) \approx \prod_{i=1}^{n} (b_i - a_i)\, g(\zeta_i; \boldsymbol{\theta})$ with a corresponding log-likelihood equal to:

$$\ell(\boldsymbol{\theta}) \approx \sum_{i=1}^{n} \left[\log(b_i - a_i) - \log(\zeta_i(1 - \zeta_i)) + \log(\phi(\zeta_i; \boldsymbol{\theta}))\right], \tag{11}$$

where $\phi(.)$ is the density of the normal distribution with mean $\boldsymbol{x}_i^T \boldsymbol{\gamma}$ and variance σ^2.

From (11) it is obvious that the approximate log-likelihood is proportional to the log-likelihood function of the simple linear model implying a much easier to compute Fisher Information matrix (given in Appendix B) at expense of loss of accuracy. Therefore, this approximation is only useful as an initial sample size calculation.

The methods for power and sample calculations that have been described in this section have been implemented in `R` (R Development Core Team, 2005) in the package `grouped` which has been written by the first two authors and it is available from CRAN (`htpp://cran.r-project.org`). In particular, power calculations for both the conditional (4) and the marginal case (6) for various distributional assumptions for the covariates (e.g., normal, gamma, beta, chi-square, uniform and bernoulli) are available. For the sample size calculations an estimate of the initial search area for the algorithm is provided using the approximate log-likelihood (11). This package allows also for fitting the model (1) under three link functions (i.e., identity, log, logit) and three distributional assumptions for the underlying transformed latent variable (i.e., normal, logistic, Student's-t).

A simulation study presented in Section 5 evaluates the performance of both the conditional and marginal power and sample size formulas under (1) as they have been presented in this section.

5. Simulation study

In this section we will describe the results of a set of simulation studies designed with the following objectives. Firstly, to explore the distribution of the Wald statistic (3) under both the null and alternative hypothesis. Secondly, to show the increase in the power of the Wald statistic when adjusting for an important covariate. Thirdly, to investigate the sensitivity of the power estimate to mis-specifications for the covariate distribution and finally to assess the validity of the proposed formulas for sample size calculations and (conditional and marginal) power calculations under a number of scenarios. In all simulation studies we considered model (1) with the linear predictor (2) assuming additional to the treatment covariate one continuous covariate and equal sample sizes for the two treatment groups (i.e., $n_1 = n_2$). Various shapes for the response distribution have been considered, namely unimodal ($\gamma_0 = 0$, $\sigma = 1$), U-shaped ($\gamma_0 = 0$, $\sigma = 4$) and J-shaped ($\gamma_0 = 2$, $\sigma = 1$) with $m = 11$ and $m = 21$ categories. The treatment effect size and the impact of the continuous covariate were taken to be small, moderate and strong with values $\Delta/\sigma = 0.2, 0.5, 0.7$ and $\gamma_1/\sigma = 0.2, 0.7, 0.9$, respectively. Finally, in order to evaluate our proposals we calculated the (conditional and marginal) empirical power of the Wald test for every scenario and compared it with the (conditional and marginal) power estimated by our formulas. The conditional empirical power is calculated as follows: using the known model design matrix and the assumed parameter values of each scenario 1,000 datasets are simulated, then model (1) is fitted to each of them and the empirical power is then defined as the percentage of times the null hypothesis was rejected. The marginal empirical power is calculated as follows: 100 design matrices are simulated (including both the treatment indicator and the continuous covariate). Based on each of the 100 design matrices, 1,000 datasets are simulated using the assumed parameter values of each scenario. Then model (1) is fitted to each of them and the empirical power is again defined as the percentage of times the null hypothesis was rejected.

5.1 *Distributional assumptions for the Wald statistic*

Usually, the asymptotic normality of the MLEs is used to claim the normality of the Wald test. However, simulations indicate that a Student's-t distribution is more appropriate in this case. Namely, we simulated 10,000 data sets under various scenarios regarding the shape of the response distribution, the sample size, the treatment effect size and the covariate effect size as described in the beginning of Section 5. The computation of the standard errors is based on the observed information matrix. Our findings were checked graphically by comparing the empirical and the hypothesized cumulative distribution function of the normal and the Student's-t distribution. Under the null hypothesis and even for a small sample size (i.e., $n_1 = n_2 = 20$) the Wald test statistic follows a central Student's-t distribution with $n - 3$ degrees of freedom (where 3 is the rank of the design matrix here). Moreover, the two-sided type I error of the Wald test under model (1) with significance level equal to 0.05 was shown through simulations in (Lesaffre et al., 2006) to be close to its nominal level for various shapes of the observed distribution. Besides its robustness has also been studied by (Heeren and D'Agostino, 1987; Sullivan and D'Agostino, 2003). Under the alternative hypothesis, the distribution of the Wald test statistic is often well approximated by a non-central t distribution with $n-3$ degrees of freedom and non-centrality parameter $\delta = \Delta_a/\sigma_{\hat{\Delta}}$. However, for a small sample size ($n_1 = n_2 = 10$) and a very skewed distribution (i.e, $\gamma_0 = 3$ and $\sigma = 1$ or $\gamma_0 = 3$ and $\sigma = 4$) the distribution of the Wald test statistic deviated under both the null and alternative hypothesis from the Student's-t distribution. In such cases, probably another test statistic is required.

5.2 *The impact of covariate adjustment on the power of the Wald test*

The impact of covariate adjustment on the power of the Wald statistic is illustrated in Table 1. It was found that even a weak covariate effect increased the power of the statistic.

[Table 1 about here.]

5.3 *Mis-specification of the distributional assumptions of the continuous covariate*

In this section the sensitivity of the marginal power calculations to mis-specification of the covariate distribution is explored. So far we have assumed that the continuous covariate x has a normal distribution. However, distributional assumptions for the covariates are usually difficult to make and thus prone to mis-specification. To evaluate the sensitivity of the marginal power function to mis-specification of the covariate distribution we have considered two cases for the true covariate distribution, namely a Gamma(4,1) distribution and the mixture distribution $0.6 \times N(-2, 1) + 0.4 \times N(2, 1)$ to allow for more general shapes. We then performed power calculation and assumed for the covariate a $N(4, 4)$ and a $N(-0.4, \sqrt{5})$, respectively. Thus, we assumed that the first two moments are correctly specified. For the covariate effect γ_1/σ we have taken it equal to 0.7. Under this assumption, various scenarios were simulated and we calculated the empirical power of the Wald statistic based on 1,000 data sets. This limited simulation study shows that when the first two moments for the covariate distribution are correctly specified the marginal power of the Wald statistic calculated by (6) is close to the empirical power. However, deviations were observed for small sample sizes and extreme shapes for the observed data distribution. The results for all the scenarios are given in Table 2 and 3.

[Table 2 about here.]

[Table 3 about here.]

5.4 *Evaluation of the determination of the power and sample size*

5.4.1 Description of the simulation study. In this simulation study the performance of the formulas for the conditional and marginal power and sample size are evaluated. For the conditional power, we generated a single design matrix obtained from the realized treatment indicator having a Bernoulli(0.5) distribution and a realized continuous covariate having a standard normal distribution. For various choices for the treatment effect size, the shape of the response distribution and the sample size, 1,000 data sets

were simulated and the empirical power of the Wald test was calculated. For the marginal power and the determination of the sample size, we generated 100 design matrices as described above. For each design matrix 1,000 data sets were generated and that for various choices of the treatment effect size, covariate effect size, model standard deviation, intercept and sample size. The number of Monte Carlo iterations was chosen to be 500 in order to achieve a precision level of 10^{-5}. To evaluate the procedure for sample size calculation we set the marginal power at 0.90 (nominal power) and calculated the sample size by using (9) employing the two-stage procedure explained in Section 4.3. The difference between the empirical marginal power obtained for the estimated sample size and the nominal marginal power was used to evaluate the performance of expression (9).

A simple but ad hoc approach to handle discrete bounded responses and thus to calculate the power and sample size is to treat them as continuous and apply the logit transformation after adding and/or subtracting a small quantity (e.g., 10^{-5}) from the values at the boundaries to achieve a normal distribution. This is done in practice when the number of realized discrete values is relatively high, say about 10 or more. Consequently, in the same simulation study we compared the performance of the proposed expressions using this ad hoc method.

5.4.2 Simulation results. The results for the conditional power are summarized in Tables 4. In Table 4, based on a BOS with 21 categories, we observe that for both sample sizes the conditional power calculation (LN power) is close to the empirical power. A larger deviation is seen for the ad hoc method but especially for small sample sizes. In addition, based on a BOS with 11 categories, essentially the same results were obtained but the deviation between the empirical and the ad hoc power increased. As expected, the power increases with the sample size for both methods.

[Table 4 about here.]

Similar results were obtained for the marginal power given in Table 5. In addition, we show the precision (expressed by the standard error) of the estimated marginal power

(LN power) based on Monte Carlo integration. In the majority of the cases the empirical power is included or is very close to the corresponding 95% confidence intervals implying a good performance of the proposed expressions.

[Table 5 about here.]

Table 6 presents the calculated sample sizes, using expression (9), and the corresponding empirical power of the Wald test for various scenarios under model (1) (LN power) and under the ad hoc method. We observe that the sample size calculation based on the LN power expression gives quite accurate results. Moreover, in the majority of the scenarios the ad hoc method fails to give a good estimate of the sample size required to get the marginal power 90%.

[Table 6 about here.]

In conclusion, it has been shown that the LN method is accurate enough for power or sample size calculations in practice, since it explicitly takes into account the rounding.

6. Relation to the OLR model

An alternative approach to compare two treatments based on a discrete bounded response in the presence of covariates is ordinal logistic regression (McCullagh, 1980). To see the approximate relationship between our approach and OLR in the case where the coarsening mechanism is the same for all subjects (i.e., all subjects have the same set of cut points), we replace $\left[\Phi\left(\frac{z_i^{(u)}-\eta_i}{\sigma}\right)-\Phi\left(\frac{z_i^{(l)}-\eta_i}{\sigma}\right)\right]$ in model (1) by $\left[F_L\left(\theta_j-\eta_i\right)-F_L\left(\theta_{(j-1)}-\eta_i\right)\right]$, with F_L representing the standard cumulative logistic distribution and θ_j the jth unknown cut point if the ith value of the BOS lies in the jth interval. In this way we obtain the ith contribution to the OLR likelihood. It is known that it is hard to distinguish between a normal and a logistic distribution function (McCullagh and Nelder, 1989). So, in the simple case where the coarsening mechanism is the same for all subjects, the two likelihoods mainly differ in the way the cut points are specified. In the OLR likelihood the cut points are estimated by maximizing the OLR with respect to the regression parameters and θ_j $(j = 1, \ldots, J)$ under the assumption that $\theta_{(j-1)} \leq \theta_j$

$(j = 1, \ldots, J)$, while in the likelihood based on model (1) the cut points are fixed up to the scale parameter σ. Of course, the regression parameters in the two models also differ by this scale parameter. This implies that BOSs following a simple coarsening mechanism with a logistic distribution in the transformed scale can also be analyzed using OLR.

Given the similarity of the two models (grouped logit-normal regression model and OLR model), we now pose the question whether our approach to calculate the power and necessary sample size could also be used for OLR, despite the (slightly) different philosophies. To this end we compared our calculated powers for the different scenarios described in the previous section with the empirical powers obtained by analyzing the data generated from these scenarios using OLR. This is an important question since there is no procedure available to perform power and/or sample size calculations in the presence of continuous covariates.

In Figure 2 we have put on the X-axis the calculated LN power of the scenarios described in the previous section and on the Y-axis the corresponding empirical power obtained each time from 1000 fitted OLR models.

[Figure 2 about here.]

The simulation results reveal that the power of OLR is basically the same as assuming a logit-normal distribution for the latent variable in both treatment groups. More importantly, though, it can be concluded that our approach to calculate the power and sample size for BOSs approximates quite well the true power and necessary sample size corresponding to OLR. However, to some extent, it is somewhat surprising that there is no loss of power for OLR despite that J cut points need to be estimated instead of 1 scale parameter for the logit-normal approach. This result deserves some further mathematical exploration.

7. Application

The Barthel index is an ADL scale commonly used in stroke trials to measure the ability of the patients to perform their daily activities after an acute stroke. The Barthel index

at 3 months after the acute stroke, was the primary endpoint of the ECASS-1 study (Dávalos et al., 1999) comparing a thrombolytic drug with placebo. In this section we perform power and sample size calculations for a future stroke study with a design similar to the ECASS-1 study and employing the historical data of the ECASS-1 study. The analysis revealed that the treatment effect using the Wald statistic is not significant at 5% (p-value=0.210) while age is an important covariate (p-value < 0.001). Consequently, we performed a sample size calculation to achieve a marginal power equal to 0.80 to detect a relevant treatment effect while adjusting for age based on our proposed methods described above. However, calculations without adjusting for age will also be performed to show the effect of covariate adjustment. In particular, we will consider two models with the following linear predictors:

$$\begin{aligned} \text{Model 1: } \eta_i &= \gamma_0 + \Delta t_i \\ \text{Model 2: } \eta_i &= \gamma_0 + \Delta t_i + \gamma_1 x_i \end{aligned}$$

where t_i denotes the treatment indicator and x_i represents age. Further, we standardized age to have zero mean and standard deviation 1. Under Model 1 the estimated linear predictor is $\hat{\eta}_i = 2.296 + 0.509 t_i$ and $\hat{\sigma} = 4.96$, while under Model 2 the estimated linear predictor is $\hat{\eta}_i = 2.227 + 0.543 t_i - 0.128 x_{1i}$ and $\hat{\sigma} = 4.71$. The estimated coefficients for treatment and age were assumed to be the same in a future study and thus sample size calculations were performed to detect a treatment effect of size $\Delta/\sigma = 0.1$ adjusting for age with effect size equal to $\gamma/\sigma = 0.03$. The historical data from the ECASS-1 study were used to make distributional assumptions for age. We assumed that in the future study there is a 1:1 allocation rate to the two arms of the study. Hence, we assumed a Bernoulli(0.5) distribution for the treatment indicator and a $N(0, 1)$ distribution for the standardized age.

Under these assumptions we calculated the necessary sample size based on the Wald test to obtain a marginal power of 80%. We obtained a sample size equal to 3670 when no covariate adjustment for age was performed (Model 1). When adjustment for age was performed (Model 2) the necessary sample size decreased to 2887. Hence, adjusting for age implied an important reduction in sample size. Given the results of the previous

section we conclude that the same reduction would be obtained if the Wilcoxon test is replaced by OLR. In addition, calculation of the sample size to achieve 80% using the ad hoc method while adjusting for covariates gives an estimate equal to 2366. However, the empirical power for this sample size equals 70.48%. This remark is in line with the conclusions of Section 5.4.2, in that our method explicitly takes the rounding into account.

8. Concluding Remarks

In this paper we have proposed expressions for power and sample size calculations for detecting treatment effects when the primary response is a BOS and adjustments for additional covariates are envisaged. Our methods are based on the parametric approach proposed by (Lesaffre et al., 2006) for handling BOS that assumes a logistic transformation on the latent scale. This family of models is quite general and can capture various shapes varying from unimodal to U or J.

The main features of our approach are the following: power calculations are performed under the LSA assumption on the logit scale that effectively deals with the bounded nature of the responses especially when skewed distributions are considered. In addition, we are assuming location in means that can be easily quantified in contrast to alternatives that involve probabilities or odds ratios that may be difficult to interpret. Moreover, the parametric model that is assumed for power calculations allows for covariate adjustments that can considerably increase the power. Furthermore, even though the focus of this paper is on BOSs, the proposed formulas are also applicable in the more general grouped data context, allowing for various coarsening mechanisms. Besides, our method for power calculations can be also applied to OLR models due to their asymptotic equivalence. Moreover, the extension to unequal variances between the treatment groups can be easily handled which may prove important in practice (Lesaffre et al., 2006). Finally, the practicality of our proposal can be investigated using the `R` package `grouped`, available from `http://cran.r-project.org`.

However, there are some limitations in our method and some issues that need further exploration. In particular, even though the family of distributions we consider is quite

general, some shapes may not be well approximated. Thus, formal tools for checking the goodness-of-fit of the model are required when pilot or historical data are available. Informally, there are two methods that can be considered. Firstly, the logit-normal distribution can be relaxed to a logit-t distribution and its appropriateness can be checked through the likelihood ratio test. Our formulas, in fact can be easily adjusted for the logit-t case, by replacing the Normal pdf and cdf by the Student's-t analogues. Secondly, in the absence of covariates we can graphically check the appropriateness of the logit-normal distribution using a kernel density estimation for the treatment groups. Another issue that needs further investigation is the effect of misspecifying the covariates' distribution in the calculation of the marginal power. Our limited simulation study showed that accurate power estimates are obtained when the first two moments are correctly specified. Finally, we have not discussed the inclusion of a BOS as a baseline covariate which is rather common in practice. However, such an extension is not straightforward and further investigation is required.

Acknowledgements

The authors acknowledge support from the Interuniversity Attraction Poles Program − Belgian State − Federal Office for Scientific, Technical and Cultural Affairs P5/24

Appendix A: Fisher Information matrix for Section 4

According to (1) the log-likelihood for the logit-normal model is:

$$\ell(\boldsymbol{\theta}; y) = \sum_{i=1}^{n} \log \{\Delta\Phi_i\},$$

where $\Delta\Phi_i = \left[\Phi\left(\frac{z_i^{(u)}-\mu_i}{\sigma}\right) - \Phi\left(\frac{z_i^{(l)}-\mu_i}{\sigma}\right)\right]$.

The elements of the Fisher Information matrix corresponding to the parameter vector $\boldsymbol{\gamma}$ are given as follows:

$$E_{Y|X}\left(-\frac{\partial^2 \ell}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T}\right) = X^T W^{(1)} X$$

where X is the $n \times p$ model design matrix and $W^{(1)}$ is a $n \times n$ diagonal matrix with

elements $w_i^{(1)}$:

$$w_i^{(1)} = \sum_{j=0}^{m} \left[\frac{(z_{ij}^{(u)} - \mu_i)\phi^{(u)} - (z_{ij}^{(l)} - \mu_i)\phi^{(l)}}{\sigma^3} + \frac{\left(\phi^{(u)} - \phi^{(l)}\right)^2}{\sigma^2 \Delta\Phi_{ij}} \right]$$

where $\mu_i = \boldsymbol{x}_i^T \boldsymbol{\gamma}$, $\phi^{(u)} = \phi\left(\frac{z_{ij}^{(u)} - \mu_i}{\sigma}\right)$, $\phi^{(l)} = \phi\left(\frac{z_{ij}^{(l)} - \mu_i}{\sigma}\right)$ with $\phi(.)$ denoting the density of the standard normal distribution.

The expected value of the minus second order derivative of the parameter vector $\boldsymbol{\gamma}$ with respect to σ are given as follows:

$$E_{Y|X}\left(-\frac{\partial^2 \ell}{\partial \boldsymbol{\gamma} \partial \sigma}\right) = X^T w^{(2)}$$

where the elements of the vector $w^{(2)}$ are given as follows:

$$\begin{aligned} w_i^{(2)} = \sum_{j=0}^{m} -\frac{\phi^{(u)} - \phi^{(l)}}{\sigma^2} &+ \frac{(z_{ij}^{(u)} - \mu_i)^2\phi^{(u)} - (z_{ij}^{(l)} - \mu_i)^2\phi^{(l)}}{\sigma^4} \\ &+ \frac{\left[(z_{ij}^{(u)} - \mu_i)\phi^{(u)} - (z_{ij}^{(l)} - \mu_i)\phi^{(l)}\right]\left(\phi^{(u)} - \phi^{(l)}\right)}{\sigma^3 \Delta\Phi_{ij}} \end{aligned}$$

Finally, the expected value of the minus second order derivative for σ is:

$$E_{Y|X}\left(-\frac{\partial^2 \ell}{\partial \sigma \partial \sigma}\right) = \sum_{i=1}^{n} w_i^{(3)}$$

where

$$\begin{aligned} w_i^{(3)} = \sum_{j=0}^{m} -\frac{2\left((z_{ij}^{(u)} - \mu_i)\phi^{(u)} - (z_{ij}^{(l)} - \mu_i)\phi^{(l)}\right)}{\sigma^3} &+ \frac{(z_{ij}^{(u)} - \mu_i)^3\phi^{(u)} - (z_{ij}^{(l)} - \mu_i)^3\phi^{(l)}}{\sigma^5} \\ &+ \frac{\left[(z_{ij}^{(u)} - \mu_i)\phi^{(u)} - (z_{ij}^{(l)} - \mu_i)\phi^{(l)}\right]^2}{\sigma^4 \Delta\Phi_{ij}} \end{aligned}$$

Appendix B: Fisher Information matrix based on the approximate likelihood

According to (11) the approximated log-likelihood for the logit-normal model is:

$$\ell(\boldsymbol{\theta}; y) = \sum_{i=1}^{n} \left\{ -\log(\pi\sigma) - \frac{1}{2}\left(\frac{\text{logit}(\zeta_i) - \mu_i}{\sigma}\right) - \log(1 - \zeta_i) \right\}$$

where $\zeta_i = \frac{a_i + b_i}{2}$.

The elements of the Fisher Information matrix corresponding to the parameter vector $\boldsymbol{\gamma}$ are those of the simple linear model:

$$E_{Y|X}\left(-\frac{\partial^2 \ell}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T}\right) = \frac{1}{\sigma^2} X^T X.$$

The expected value of the minus second order derivative of the parameter vector $\boldsymbol{\gamma}$ with respect to σ are given as follows:

$$E_{Y|X}\left(-\frac{\partial^2 \ell}{\partial \boldsymbol{\gamma} \partial \sigma}\right) = X^T \lambda^{(1)}$$

where the elements of the vector $\lambda^{(1)}$ are given as follows:

$$\lambda_i^{(1)} = \frac{2}{\sigma^3} \sum_{j=0}^{m} \left(\text{logit}\left(\zeta_{ij}\right) - \mu_i\right) \Delta\Phi_{ij}$$

Finally, the expected value of the minus second order derivative for σ is:

$$E_{Y|X}\left(-\frac{\partial^2 \ell}{\partial \sigma \partial \sigma}\right) = -\frac{n}{\sigma^2} + \sum_{i=1}^{n} \lambda_i^{(2)}$$

where

$$\lambda_i^{(2)} = \frac{3}{\sigma^4} \sum_{j=0}^{m} \left(\text{logit}\left(\zeta_{ij}\right) - \mu_i\right)^2 \Delta\Phi_{ij}$$

References

Aitchison, J. and Shen, S. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika* **67**, 261–272.

Armitage, R. and Berry, G. (1987). *Statistical Methods in Medical Research*. Blackwell Scientific Publications, second edition.

Collings, B. J. and Hamilton, M. A. (1988). Estimating the power of the two-sample wilcoxon test for location shift. *Biometrics* **44**, 847–860.

Dávalos, A., Toni, D., Iweins, F., Lesaffre, E., Bastianello, S. and Castillo, J. (1999). Neurological deterioration in acute ischemic stroke: Potential predistrors and associated factors in the european cooperative acute stroke study (ecass) i. *Stroke* **30**, 2631–2636.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, New York.

Hamilton, M. A. and Collings, B. J. (1991). Determining the appropriate sample size for nonparametric tests for location shift. *Technometrics* **33**, 327–337.

Heeren, T. and D'Agostino, R. B. (1987). Robustness of the two independent samples t-test when applied to ordinal scaled data. *Statistics in Medicine* **6**, 79–90.

Heitjan, D. (1989). Inference from grouped continuous data: A review. *Statistical Science* **4**, 164–183.

Heitjan, D. (1993). Ignorability and coarse data: Some biomedical examples. *Biometrics* **49**, 1099–1109.

Heitjan, D. and Rubin, D. (1991). Ignorability and coarse data. *Annals of Statistics* **19**, 2244–2253.

Lesaffre, E. and de Klerk, E. (2000). Estimating the power of compliance - improving methods. *Control Clinical Trials* **21**, 540–551.

Lesaffre, E., Rizopoulos, D. and Tsonaka, S. (2006). The logistic-transformation for bounded outcome scores. To appear.

Lesaffre, E., Scheys, I., Fröhlich, J. and Bluhmki, E. (1993). Calculation of power and sample size with bounded outcome scores. *Statistics in Medicine* **12**, 1063–1078.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B* **42**, 109–142.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London, second edition.

Noether, G. E. (1987). Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association* **82**, 645–647.

R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. 3-900051-07-0.

Sullivan, I. and D'Agostino, R. (2003). Robustness and power of analysis of covariance applied to ordinal scaled data as arising in randomized controlled trials. *Statistics in Medicine* **22**, 1317–1334.

Tsodikov, A., Hasenclever, D. and Loeffler, D. (1998). Regression with bounded outcome score: Evaluation of power by Bootstrap and Simulation in a chronic myelogenous leukaemia

clinical trial. *Statistics in Medicine* **17**, 1909–1922.

van Elteren, P. (1960). On the combination of independent two-sample tests of wilcoxon. *Bulletin of the International Statistical Institute* **37**, 351–361.

Walters, S. and Campbell, M. (2005). The use of bootstrap methods for estimating sample size and analysing health-related quality of life outcomes. *Statistics in Medicine* **24**, 1075–1102.

Figure 1. Correspondence of the LSA on the transformed scale and the treatment effect on the original scale.

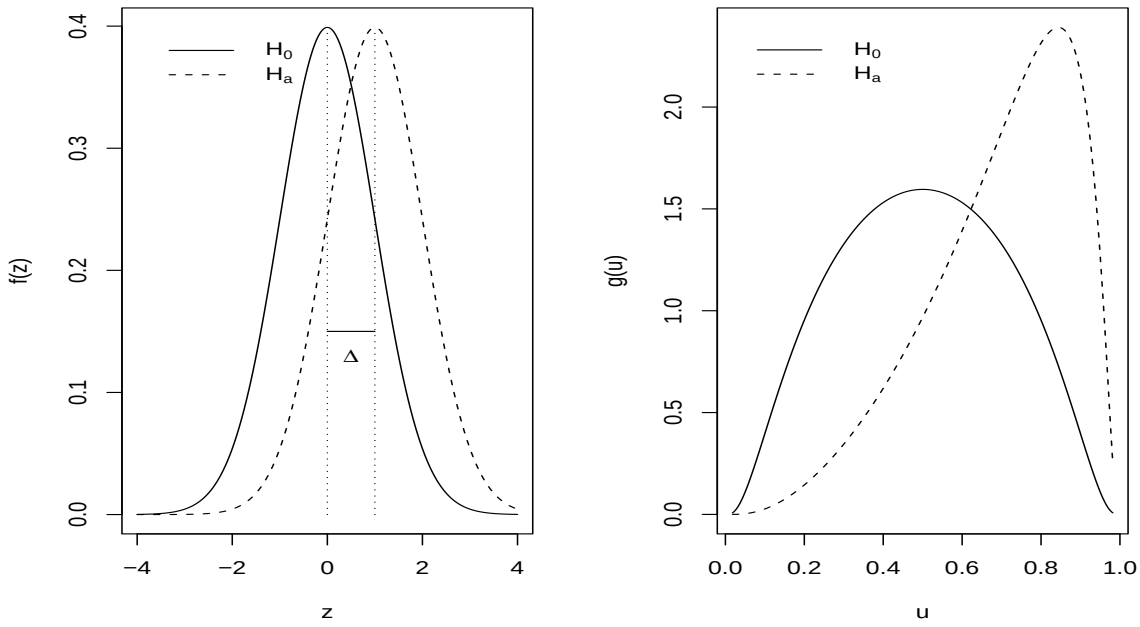


Figure 2. Comparison of the LN power to the empirical OLR power determined for the scenarios of Section 5.

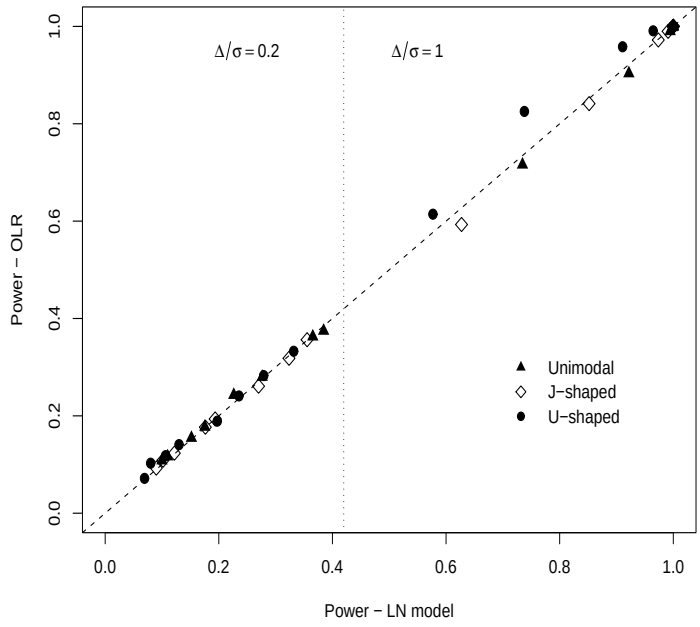


Table 1

Simulation study: Power of the Wald statistic for a BOS with 21 categories. For each scenario 1,000 data sets were simulated and the empirical power was calculated after adjusting (Adjusted) or not (Non-Adjusted) for the additional covariate (max. standard error 0.01575 and 0.01577, respectively).

Sample Size	γ_0	σ	Effect Size		Empirical Power	
$n_1 = n_2$			Δ/σ	γ_1/σ	Adjusted	Non-Adjusted
50	0	1	0.5	0.2	0.716	0.690
				0.9	0.705	0.463
50	2	1	0.5	0.2	0.672	0.649
				0.9	0.640	0.417
50	0	4	0.5	0.2	0.622	0.618
				0.9	0.581	0.380
50	2	4	0.5	0.2	0.597	0.574
				0.9	0.545	0.364
20	0	1	0.5	0.2	0.392	0.380
				0.9	0.371	0.225
20	2	1	0.5	0.2	0.326	0.305
				0.9	0.326	0.208
20	0	4	0.5	0.2	0.295	0.277
				0.9	0.247	0.140
20	2	4	0.5	0.2	0.250	0.236
				0.9	0.237	0.138

Table 2

Simulation study results for sensitivity of the marginal power function to covariate mispecification. The marginal power was evaluated using expression (6) with 500 Monte Carlo iterations. The covariate distribution was assumed to be $N(4,4)$ for a true Gamma(4,1). The empirical power (max. standard error 0.01452) is obtained via simulation using 1,000 datasets.

Sample Size	Max Score	γ_0	σ	Effect Size		Power	
$n_1 = n_2$	m			Δ/σ	γ_1/σ	Empirical	Marginal
50	20	0	1	0.2	0.7	0.168	0.164
				1.0	0.7	0.999	0.998
50	20	2	1	0.2	0.7	0.163	0.155
				1.0	0.7	0.996	0.993
50	20	0	4	0.2	0.7	0.144	0.144
				1.0	0.7	0.994	0.978
50	20	2	4	0.2	0.7	0.143	0.138
				1.0	0.7	0.991	0.954
20	20	0	1	0.2	0.7	0.113	0.092
				1.0	0.7	0.883	0.846
20	20	2	1	0.2	0.7	0.087	0.089
				1.0	0.7	0.822	0.773
20	20	0	4	0.2	0.7	0.084	0.085
				1.0	0.7	0.751	0.684
20	20	2	4	0.2	0.7	0.071	0.082
				1.0	0.7	0.698	0.609

Table 3

Simulation study results for sensitivity of the marginal power function to covariate mispecification. The marginal power was evaluated using expression (6) with 500 Monte Carlo iterations. The covariate distribution was assumed to be $N(-0.4, \sqrt{5})$ *for a true* $0.6 \times N(-2, 1) + 0.4 \times N(2, 1)$ *. The empirical power (max. standard error 0.01558) is obtained via simulation using 1,000 datasets.*

Sample Size	Max Score	γ_0	σ	Effect Size		Power	
$n_1 = n_2$	m			Δ/σ	γ_1/σ	Empirical	Marginal
50	20	0	1	0.2	0.7	0.184	0.164
				1.0	0.7	0.998	0.998
50	20	2	1	0.2	0.7	0.155	0.155
				1.0	0.7	0.993	0.993
50	20	0	4	0.2	0.7	0.147	0.143
				1.0	0.7	0.986	0.978
50	20	2	4	0.2	0.7	0.129	0.138
				1.0	0.7	0.975	0.954
20	20	0	1	0.2	0.7	0.097	0.092
				1.0	0.7	0.847	0.846
20	20	2	1	0.2	0.7	0.107	0.088
				1.0	0.7	0.831	0.773
20	20	0	4	0.2	0.7	0.091	0.084
				1.0	0.7	0.753	0.687
20	20	2	4	0.2	0.7	0.069	0.082
				1.0	0.7	0.586	0.608

Table 4

Simulation study: Evaluation of the conditional power expression (4) (LN power) for a BOS with 21 categories. For each scenario 1,000 data sets were generated from which the empirical (max. standard error 0.01572 and 0.01580) and the ad hoc power were calculated.

Sample Size	γ_0	σ	Effect Size		Conditional Power					
					Maximum score 20			Maximum score 10		
$n_1 = n_2$			Δ/σ	γ_1/σ	Empirical	LN	Ad hoc	Empirical	LN	Ad hoc
50	0	1	0.2	0.7	0.173	0.166	0.129	0.171	0.164	0.137
			0.5	0.7	0.723	0.686	0.553	0.699	0.682	0.457
			1.0	0.7	0.998	0.998	0.940	0.997	0.997	0.941
50	2	1	0.2	0.7	0.173	0.155	0.120	0.154	0.145	0.129
			0.5	0.7	0.633	0.629	0.403	0.582	0.566	0.432
			1.0	0.7	0.993	0.994	0.976	0.989	0.976	0.967
50	0	4	0.2	0.7	0.139	0.143	0.136	0.141	0.139	0.160
			0.5	0.7	0.615	0.591	0.607	0.571	0.558	0.525
			1.0	0.7	0.995	0.973	0.991	0.986	0.963	0.988
50	2	4	0.2	0.7	0.129	0.138	0.130	0.114	0.133	0.134
			0.5	0.7	0.553	0.546	0.532	0.477	0.494	0.426
			1.0	0.7	0.988	0.960	0.988	0.964	0.917	0.974
20	0	1	0.2	0.7	0.112	0.070	0.074	0.091	0.069	0.036
			0.5	0.7	0.362	0.308	0.295	0.340	0.324	0.172
			1.0	0.7	0.603	0.552	0.414	0.589	0.521	0.378
20	2	1	0.2	0.7	0.097	0.068	0.074	0.066	0.065	0.061
			0.5	0.7	0.315	0.299	0.210	0.271	0.255	0.181
			1.0	0.7	0.570	0.501	0.299	0.428	0.392	0.356
20	0	4	0.2	0.7	0.062	0.066	0.068	0.047	0.065	0.069
			0.5	0.7	0.272	0.270	0.288	0.208	0.251	0.318
			1.0	0.7	0.412	0.414	0.482	0.255	0.351	0.411
20	2	4	0.2	0.7	0.056	0.065	0.064	0.043	0.063	0.074
			0.5	0.7	0.228	0.232	0.299	0.188	0.229	0.237
			1.0	0.7	0.337	0.373	0.395	0.227	0.331	0.504

Table 5

Simulation study: Evaluation of the marginal power expression (6) (LN power) for a BOS with 21 and 11 categories. The empirical (max. standard error 0.01570 and 0.01580) and the ad hoc power are obtained from sampling 100 sets of covariates and 1,000 data sets for each one. The standard errors (s.e.) are based on 500 Monte Carlo simulated values.

Sample Size	γ_0	σ	Effect Size		Maximum score 20 Marginal Power			Maximum score 10 Marginal Power		
$n_1 = n_2$			Δ/σ	γ_1/σ	Empirical	LN (s.e.)	Ad hoc	Empirical	LN (s.e.)	Ad hoc
50	0	1	0.2	0.7	0.172	$0.165(6.71{\cdot}10^{-5})$	0.120	0.169	$0.163(6.71{\cdot}10^{-5})$	0.111
			0.5	0.7	0.700	$0.688(7.29{\cdot}10^{-5})$	0.510	0.687	$0.678(6.90{\cdot}10^{-5})$	0.433
			1.0	0.7	0.999	$0.998(< 10^{-5})$	0.948	0.998	$0.998(1.12{\cdot}10^{-5})$	0.946
50	2	1	0.2	0.7	0.161	$0.156(8.94{\cdot}10^{-5})$	0.116	0.144	$0.142(1.01{\cdot}10^{-4})$	0.121
			0.5	0.7	0.651	$0.640(8.10{\cdot}10^{-5})$	0.476	0.583	$0.567(12.60{\cdot}10^{-4})$	0.488
			1.0	0.7	0.996	$0.993(4.47{\cdot}10^{-5})$	0.971	0.985	$0.972(2.24{\cdot}10^{-4})$	0.965
50	0	4	0.2	0.7	0.143	$0.144(6.71{\cdot}10^{-5})$	0.142	0.132	$0.138(7.83{\cdot}10^{-5})$	0.136
			0.5	0.7	0.602	$0.584(7.87{\cdot}10^{-4})$	0.588	0.563	$0.550(8.31{\cdot}10^{-4})$	0.564
			1.0	0.7	0.992	$0.979(1.23{\cdot}10^{-4})$	0.990	0.987	$0.960(2.01{\cdot}10^{-4})$	0.985
50	2	4	0.2	0.7	0.136	$0.139(7.83{\cdot}10^{-5})$	0.134	0.132	$0.138(7.83{\cdot}10^{-5})$	0.136
			0.5	0.7	0.560	$0.545(11.4{\cdot}10^{-4})$	0.545	0.516	$0.505(11.22{\cdot}10^{-4})$	0.516
			1.0	0.7	0.982	$0.955(2.91{\cdot}10^{-4})$	0.978	0.970	$0.919(4.58{\cdot}10^{-4})$	0.967
20	0	1	0.2	0.7	0.107	$0.093(6.71{\cdot}10^{-5})$	0.069	0.104	$0.092(5.59{\cdot}10^{-5})$	0.058
			0.5	0.7	0.352	$0.328(8.93{\cdot}10^{-4})$	0.232	0.345	$0.322(8.47{\cdot}10^{-4})$	0.194
			1.0	0.7	0.870	$0.855(5.25{\cdot}10^{-4})$	0.672	0.862	$0.843(5.59{\cdot}10^{-4})$	0.606
20	2	1	0.2	0.7	0.099	$0.089(6.71{\cdot}10^{-5})$	0.074	0.090	$0.085(6.71{\cdot}10^{-5})$	0.079
			0.5	0.7	0.320	$0.300(9.44{\cdot}10^{-4})$	0.217	0.274	$0.260(11.68{\cdot}10^{-4})$	0.224
			1.0	0.7	0.815	$0.783(8.61{\cdot}10^{-4})$	0.667	0.720	$0.668(1.29{\cdot}10^{-3})$	0.651
20	0	4	0.2	0.7	0.081	$0.085(5.59{\cdot}10^{-5})$	0.086	0.069	$0.083(5.59{\cdot}10^{-5})$	0.083
			0.5	0.7	0.268	$0.270(8.57{\cdot}10^{-4})$	0.273	0.233	$0.252(7.69{\cdot}10^{-4})$	0.261
			1.0	0.7	0.753	$0.695(8.16{\cdot}10^{-4})$	0.755	0.688	$0.631(9.84{\cdot}10^{-4})$	0.728
20	2	4	0.2	0.7	0.078	$0.083(5.59{\cdot}10^{-5})$	0.083	0.065	$0.081(5.59{\cdot}10^{-5})$	0.082
			0.5	0.7	0.242	$0.249(11.54{\cdot}10^{-4})$	0.252	0.201	$0.231(1.16{\cdot}10^{-4})$	0.237
			1.0	0.7	0.670	$0.621(1.19{\cdot}10^{-3})$	0.689	0.571	$0.551(9.89{\cdot}10^{-3})$	0.647

Table 6

Simulation study: Evaluation of the calculated sample size using expression (9) (LN sample) for BOSs with 21 categories and marginal power equal to 90%. The empirical power (LN and Ad hoc) is obtained via simulating 100 sets of covariates and 1,000 data sets for each.

γ_0	σ	Effect Size		Sample Size		Empirical Power	
		Δ/σ	γ_1/σ	LN	Ad hoc	LN	Ad hoc
0	1	0.2	0.7	1063	1053	0.900	0.697
		0.5	0.7	172	170	0.901	0.701
		1.0	0.7	45	44	0.913	0.719
2	1	0.2	0.7	1154	1053	0.902	0.694
		0.5	0.7	192	171	0.904	0.701
		1.0	0.7	54	45	0.915	0.719
0	4	0.2	0.7	1293	1053	0.902	0.819
		0.5	0.7	218	170	0.910	0.815
		1.0	0.7	67	45	0.942	0.805
2	4	0.2	0.7	1374	1053	0.902	0.794
		0.5	0.7	240	171	0.912	0.778
		1.0	0.7	78	46	0.943	0.752