Cundill and Alexander *BMC Medical Research Methodology* (2015) 15:28
DOI 10.1186/s12874-015-0023-0

BMC
Medical Research Methodology

RESEARCH ARTICLE Open Access

Sample size calculations for skewed distributions

Bonnie Cundill and Neal DE Alexander*

Abstract

Background: Sample size calculations should correspond to the intended method of analysis. Nevertheless, for non-normal distributions, they are often done on the basis of normal approximations, even when the data are to be analysed using generalized linear models (GLMs).

Methods: For the case of comparison of two means, we use GLM theory to derive sample size formulae, with particular cases being the negative binomial, Poisson, binomial, and gamma families. By simulation we estimate the performance of normal approximations, which, via the identity link, are special cases of our approach, and for common link functions such as the log. The negative binomial and gamma scenarios are motivated by examples in hookworm vaccine trials and insecticide-treated materials, respectively.

Results: Calculations on the link function (log) scale work well for the negative binomial and gamma scenarios examined and are often superior to the normal approximations. However, they have little advantage for the Poisson and binomial distributions.

Conclusions: The proposed method is suitable for sample size calculations for comparisons of means of highly skewed outcome variables.

Keywords: Sample size, Generalized linear models, Power, Berry-Esséen theorem

Background

Sample size calculations estimate the required number of patients to meet a study's objective(s). The method used to analyse the subsequent data will affect the actual power, although this dependence is often ignored in practice. Sample size calculations are often based on normal approximation, such as those described by Lachin [1], even for data which are not Gaussian and which are analysed using generalized linear models (GLMs) [2-6]. Some medical statistics textbooks which cover Poisson regression still obtain sample sizes for rates via a normal approximation [7-10]. Using a statistical method which does not correspond to that used for the sample size may result in the actual power differing from the nominal value.

Methods have been proposed for the specific cases of logistic [11-14] or Poisson [15] models, or both [16], or for the negative binomial [17], and for generalized linear models [18,19]. The more general methods concentrate on single or multiple continuous predictor variables and can be somewhat complex to use. In particular, not all of them yield an explicit formula for sample size. In the current paper we consider a comparison of two means, i.e. a dichotomous predictor variable. We obtain a general formula which encompasses, for example, the Poisson and binomial distributions, but concentrate on the negative binomial and gamma — partly replicating Zhu and Lakkis for the former [17] — because these can be used to model skewed data, for which normal approximations are less likely to be satisfactory. We apply these methods to examples based on actual studies, including the negative binomial distribution for hookworm egg counts, a potential vaccine trial endpoint, and the gamma distribution for concentrations of insecticide on bednets.

Methods

We examine the magnitude of errors in normal approximations for discrete probability distributions. Then, using GLM theory, we then derive sample size formulae which are assessed using worked examples and simulations Additional file 1.

* Correspondence: neal.alexander@lshtm.ac.uk
MRC Tropical Epidemiology Group, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

Assessing the magnitude of error using normal approximations

The central limit theorem guarantees that, for a sufficiently large sample size, the sample mean has a distribution which is arbitrarily close to normal (Gaussian). To evaluate the adequacy of the normal approximation under specific circumstances, in terms of cumulative distribution functions, we used a) the Berry-Esséen theorem and b) computation of the specific distributions. All computing was done using R, version 2.15 or higher.

Berry-Esséen theorem

Let $R_1,R_2,...,R_n$ be independent and identically distributed (iid) zero-mean random variables with positive variance σ^2. Defining $S_n = \sum_{k=1}^{n} R_k/\sigma\sqrt{n}$ as the standardised mean of the random variables, $F_n(y)$ as the cumulative distribution function (CDF) of S_n, and Φ as the CDF of the standard normal distribution, the Berry-Esséen theorem [20] states

$$|F_n(y) - \Phi(y)| \leq \frac{C\rho}{\sigma^3\sqrt{n}} \quad (1)$$

where C is a distribution-independent positive constant, and $\rho < \infty$ is the absolute third central moment, $E(|R - E(R)|^3)$, which equals $E(|R|^3)$ thanks to the specification of zero mean. Values of C have decreased markedly from Esséen's original bound of 7.59 [20] to 0.4690 obtained by Shevtsova in 2013 [21]. For Poisson sums, including the Poisson itself, and the negative binomial as a mixture of Poissons, this can be replaced by 0.3051 [22]. More precise values are also available for the special cases of the binomial distributions with parameter 0.5 [23] or with denominator 1 [24], although the latter is applicable only to sample sizes of at least 200.

The Berry-Esséen approach can be used even when direct calculation from the distribution is not feasible. The bound can be expressed in terms of the third non-absolute central moment and a finite sum (see Additional file 2). Such bounds are one way to assess the adequacy of the normal distribution assumptions implicit in common sample size methods. In the following section we describe a potentially more robust sample size approach.

Sample sizes from generalized linear model theory

Generalized linear models are for vectors of independent responses, $Y_i (i = 1,...,N)$, arising from an exponential family distribution. Such distributions include the Poisson, binomial and gamma, as well as the negative binomial if its k parameter is assumed fixed [25,26]. Covariates x_{ij} enter the model as linear combinations with unknown regression coefficients β_j and can be written as

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$$

where η_i is related to μ_i, the mean of Y_i, via the link function $g{:}\eta_i = g(\mu_i)$.

The sample size for a hypothesis related to the mean of such a distribution can be calculated from the variance of its maximum likelihood estimate (MLE), on the scale of the link function. The covariance matrix of the parameter estimates for GLMs is approximately

$$\left(X^T W X\right)^{-1} \quad (2)$$

where X is the design matrix and W is the diagonal matrix of weights [27]. We need to know how the sample size affects the variance of the parameter estimate. When comparing the means of two groups of size N_0 and N_1 (with $N_0 + N_1 = N$), X has two columns and N rows. The first column, corresponding to the intercept, is all 1′s, and the second column is N_0 zeros and N_1 1′s. W is defined by

$$W = \frac{\left(\frac{d\mu}{d\eta}\right)^2}{V(\mu)} \quad (3)$$

where $V(\mu)$ is the variance function relating the mean and variance of Y [27]. The diagonal of W is composed of N_0 copies of w_0 and N_1 copies of w_1, in an obvious notation. To compare the two means, we are interested in the second diagonal element of the 2 × 2 matrix given by equation (2). Some basic matrix algebra shows that this element is $(N_0 w_0)^{-1} + (N_1 w_1)^{-1}$.

For the sample size of this comparison, we apply principles outlined by Lachin [1]. His notation uses subscripts 0 and 1 for the null and alternative hypotheses, which here we will change to O and A, using 0 and 1 instead to refer to the two groups being compared: 0 for reference or control, and 1 for intervention. We will also use λ rather than μ as a generic parameter, using the latter to denote the mean. We will also use a different subscript notation for standard normal deviates, so that z_p means the standard normal deviate for lower tail area p. Our statistic (X in Lachin's notation) is the estimate of

the difference in transformed means obtained by GLM. The transformation is typically log, or logit for binomial. The mean of this statistic is λ_O under the null hypothesis and λ_A under the alternative hypothesis, with the standard deviation being Σ_O and Σ_A. Lachin's equation 1 then becomes

$$|\lambda_A - \lambda_O| = z_{1-\frac{\alpha}{2}}\Sigma_O - z_{1-\beta}\Sigma_A \tag{4}$$

Following Lachin again, we will denote the proportions in the groups by $Q_0 = N_0/N$ and $Q_1 = =N_1/N$. Our approach is to apply a normal approximation on the scale of the link function. This is often the log, although, with the identity link, more familiar equations are obtained. We consider two approaches for estimating the variance under the null hypothesis. One is to use the reference value in both groups: following Zhu and Lakkis [17], we call this method 1. Using the above matrix algebra, Σ_O equals

$$\sqrt{\frac{1}{Q_1 N}\frac{V(\mu_0)}{(d\mu/d\eta|_{\mu=\mu_0})^2} + \frac{1}{Q_0 N}\frac{V(\mu_0)}{(d\mu/d\eta|_{\mu=\mu_0})^2}} = \sqrt{\frac{1}{N}\frac{V(\mu_0)}{(d\mu/d\eta|_{\mu=\mu_0})^2}\left(\frac{1}{Q_1}+\frac{1}{Q_0}\right)}$$

and Σ_A equals

$$\sqrt{\frac{1}{Q_1 N}\frac{V(\mu_1)}{(d\mu/d\eta|_{\mu=\mu_1})^2} + \frac{1}{Q_0 N}\frac{V(\mu_0)}{(d\mu/d\eta|_{\mu=\mu_0})^2}}$$

Hence, for method 1, we obtain

$$\sqrt{N} = \frac{Z_{1-\frac{\alpha}{2}}\sqrt{\left(\frac{1}{Q_1}+\frac{1}{Q_0}\right)\frac{V(\mu_0)}{(d\mu/d\eta|_{\mu=\mu_0})^2}} + Z_{1-\beta}\sqrt{\frac{1}{Q_1}\frac{V(\mu_1)}{(d\mu/d\eta|_{\mu=\mu_1})^2} + \frac{1}{Q_0}\frac{V(\mu_0)}{(d\mu/d\eta|_{\mu=\mu_0})^2}}}{g(\mu_0) - g(\mu_1)} \tag{5}$$

Zhu and Lakkis [17] find that the test characteristics are generally better if, instead, μ_1 is used for the intervention arm under the null hypothesis ('method 2'), so Σ_O equal Σ_A, and

$$\sqrt{N} = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)\sqrt{\frac{1}{Q_1}\frac{V(\mu_1)}{(d\mu/d\eta|_{\mu=\mu_1})^2} + \frac{1}{Q_0}\frac{V(\mu_0)}{(d\mu/d\eta|_{\mu=\mu_0})^2}}}{g(\mu_0) - g(\mu_1)} \tag{6}$$

Equations (5) and (6) are general, with special distributional cases being easily determined. We will use equation (6) except when referring to previous work based on method 1.

Negative binomial distribution

The negative binomial distribution is a generalization of the Poisson for count data, with an additional parameter (k) which can describe over-dispersion [28]. Small k implies a large variance and as $k \to \infty$ the distribution tends to Poisson. We derive results first for the negative binomial distribution, then for the Poisson as a limiting case. Let Y be a random variable which follows the negative binomial distribution with population mean μ and dispersion parameter k, with the variance function being $V(\mu) = \mu + (\mu^2/k)$ and density as shown in Additional file 3. Analysis by GLM usually employs a natural logarithm link function [25] for which $d\mu/d\eta = \mu$. Substituting into equation (5) gives

$$\sqrt{N} = \frac{Z_{1-\frac{\alpha}{2}}\sqrt{\left(\frac{1}{\mu_0}+\frac{1}{k_0}\right)\left(\frac{1}{Q_1}+\frac{1}{Q_0}\right)} + Z_{1-\beta}\sqrt{\frac{1}{Q_1}\left(\frac{1}{\mu_1}+\frac{1}{k_1}\right) + \frac{1}{Q_0}\left(\frac{1}{\mu_0}+\frac{1}{k_0}\right)}}{\log(\mu_0) - \log(\mu_1)} \tag{7}$$

For the special case of equal sample sizes and ($Q_0 = Q_1 = 0.5$) and k parameters ($k_0 = k_1$) this reduces to the equation by Brooker et al. [29]. Using equation (6) instead gives:

$$\sqrt{N} = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)\sqrt{\frac{1}{Q_1}\left(\frac{1}{\mu_1}+\frac{1}{k_1}\right) + \frac{1}{Q_0}\left(\frac{1}{\mu_0}+\frac{1}{k_0}\right)}}{\log(\mu_0) - \log(\mu_1)} \tag{8}$$

A normal approximation can be obtained by applying equation (6) on the identity scale, with variances equal to $\mu_i + \mu_i^2/k_i (i = 0, 1)$:

$$\sqrt{N} = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)\sqrt{\frac{1}{Q_1}\left(\mu_1 + \frac{\mu_1^2}{k_1}\right) + \frac{1}{Q_0}\left(\mu_0 + \frac{\mu_0^2}{k_0}\right)}}{\mu_0 - \mu_1} \tag{9}$$

We used simulation to estimate the actual power sample sizes obtained from equations (8) and (9), by generating repeated datasets of the calculated sizes and analysing them by GLM and Wald tests. We also used likelihood ratio tests, with similar results, unless where commented. For this we used the rnegbin and glm.nb function of the MASS package in R.

Poisson distribution

Let Y be a random variable denoting the number of events per unit time (for example, per study duration) then Y follows the Poisson distribution with mean μ. By letting k tend to infinity in equation (8), or, equivalently, from equation (6) with log link and $V(\mu) = \mu$, we obtain:

$$\sqrt{N} = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)\sqrt{\frac{1}{Q_1}\left(\frac{1}{\mu_1} + \frac{1}{k_1}\right) + \frac{1}{Q_0}\left(\frac{1}{\mu_0} + \frac{1}{k_0}\right)}}{\log(\mu_0) - \log(\mu_1)} \tag{10}$$

This is compared by simulation, for the case $Q_0 = Q_1 = 0.5$ (equal size arms), with the following normal approximation, on the scale of the identity link, obtained from equation (9) by again letting k tend to infinity:

$$\sqrt{N} = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)\sqrt{2(\mu_1 + \mu_0)}}{\mu_0 - \mu_1} \tag{11}$$

This is also used, for example, by Kirkwood & Sterne [7], except that here we include a factor of 2 inside the square root to obtain the total study size.

Binomial distribution

Let Y be a binomial random variable denoting the number of successes in d independent Bernoulli events, each with probability μ. The most common situation is to have $d = 1$, with each unit (person) having a response of 1 or 0 (e.g. positive or negative). An assumption of $d = 1$ may explain why the literature does not always show d in the variance function: we follow Fox [30] in using $V(\mu) = \mu(1\text{-}\mu)/d$. For the canonical logit link, $d\mu/d\eta = \mu(1\text{-}\mu)$, so, from equation (6), we obtain

$$\sqrt{N} = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)\sqrt{\frac{1}{Q_1}\frac{1}{\mu_1(1-\mu_1)} + \frac{1}{Q_0}\frac{1}{\mu_0(1-\mu_0)}}}{\sqrt{d}(\mathrm{logit}(\mu_0) - \mathrm{logit}(\mu_1))} \tag{12}$$

On the scale of difference in proportions (identity link), the corresponding equation is:

$$\sqrt{N} = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)\sqrt{\mu_1(1-\mu_1)\frac{1}{Q_1} + \mu_0(1-\mu_0)\frac{1}{Q_0}}}{\sqrt{d}(\mu_0 - \mu_1)} \tag{13}$$

This differs from the Lachin's equation (12), and that of Kirwood and Sterne, both of which have Z_α multiplied by a function of $\bar{\pi}(1-\bar{\pi})$, where $\bar{\pi}$ is an average of the μ_0 and μ_1. Some outcomes, in particular the occurrence of a given condition, could be quantified either as a Poisson rate (events per unit time, with rate μ) or as a binomial proportion (fraction of people experiencing the condition in a given period T). These options can be linked mathematically, with the latter probability equalling $1\text{-}e^{-\mu T}$. This relation can, in turn, be used to compare the power or sample size for quantifying a given scenario as either a rate or proportion. In this case the rate is the more powerful option [31]. This is to be expected, since the proportion loses information by considering all those with one or more events as a single category.

Gamma distribution

The gamma is a two-parameter continuous distribution family over positive values. Special cases include the exponential distribution, and the sum of identical independent exponentials. In applications it typically models right-skewed data [32]. If Y is such a random variable with shape parameter κ and scale parameter θ, then $E(Y) = \kappa\theta \equiv \mu$ and $V(\mu) = \kappa\theta^2 = \mu^2/\kappa$ [33]. Here we use the logarithmic link, although the reciprocal is canonical. Hence $d\mu/d\eta =$ and $w_i = \mu^2/(\mu^2/\kappa_i) = \kappa_i$ so equation (6) becomes

$$\sqrt{N} = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)\sqrt{\frac{1}{Q_1\kappa_1} + \frac{1}{Q_0\kappa_0}}}{\log(\mu_0) - \log(\mu_1)} \tag{14}$$

Results

Berry-Esséen bounds

For the example of a fixed sample size of 100, the Berry-Esséen bounds are shown in the Table 1, along with corresponding values based on computation of the non-Gaussian CDFs. As expected, both methods show the normal approximation to be better for larger means. The Berry-Esséen bounds are often much wider than those obtained from explicit computation. Hence we concentrate on the latter approach. Figure 1 shows the results for binomial distributions of varying sample size and proportion (μ). As expected, the discrepancy in the CDF of the normal approximation is generally larger for smaller sample sizes and values of μ further from 0.5. The differences are non-negligible for parameter values found in some research studies, in particular for small values of μ, say between 1 and 5%, which would be

Table 1 Maximum discrepancy in the approximating normal CDF, for sample size 100, in terms of Berry-Esséen bounds, and via computation

Distribution	Parameter estimates		Maximum error	
			Berry-Esséen	Exact CDF
Negative Binomial	***k***	***μ***		
	0.05	0.05	28.9%	12.5%
		0.1	27.9%	9.8%
		10	27.3%	6.1%
		50	27.3%	6.0%
	0.1	0.05	22.3%	11.9%
		0.1	20.6%	8.8%
		10	19.5%	4.4%
		50	19.5%	4.4%
	0.5	0.05	15.7%	11.3%
		0.1	12.5%	8.3%
		10	9.4%	2.1%
		50	9.4%	1.9%
Poisson		***μ***		
		0.05	13.7%	11.6%
		0.1	9.8%	8.2%
		10	4.9%	3.2%
		50	4.9%	1.5%
Binomial		***μ***		
(*d* = 1)		0.05	19.5%	11.6%
		0.1	12.8%	8.3%
		0.5	4.7%	4.0%

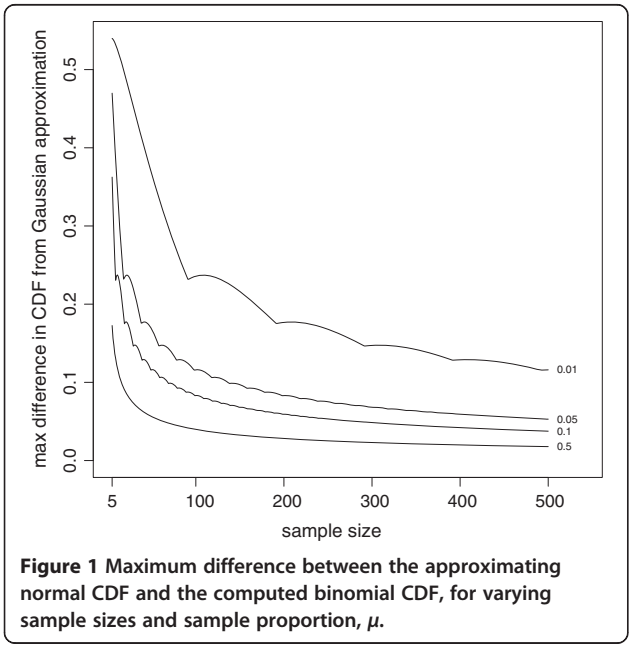

Figure 1 Maximum difference between the approximating normal CDF and the computed binomial CDF, for varying sample sizes and sample proportion, *μ*.

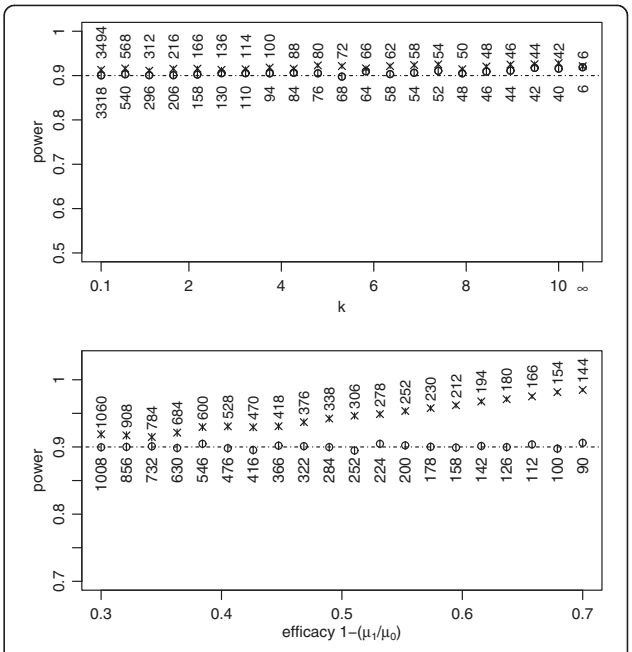

Figure 2 Negative binomial sample size. Each vertical axis shows power estimated by simulation for comparing two means and 90% nominal power. Each data point is based on 10,000 simulated datasets with two arms of equal size ($Q_0 = Q_1 = 0.5$), the ° symbol shows results for logarithmic link, equation (7), and × for identity link, equation (9). The vertical lines show the 95% confidence interval for the proportion of datasets reaching 5% two-sided significance level when analysed by GLM with logarithmic link, i.e. a measure of Monte Carlo error. The number beside each plot symbol is the sample size calculated by the relevant method. In the upper panel, $\mu_0 = 71.4$, $\mu_1 = 50$ (30% vaccine efficacy) and various values of k (assumed equal in both arms) on the horizontal axis. In the lower panel, μ_0 is as before, k is 0.33 and various values of efficacy are used (horizontal axis) with $\mu_1 = \mu_0 \times$ (1-efficacy).

expected to approximate Poisson. This tends to sustain a concern that power calculations based on normal approximations may not be accurate.

GLM approach for negative binomial distribution

We first revisit the example of Brooker et al. [29], which was motivated by the Human Hookworm Vaccine Initiative (HHVI). The degree of hookworm morbidity depends on the numbers of parasites in the intestines. Hence a quantitative endpoint is of interest for vaccine trials, and one option is the faecal egg count per Kato Katz slide. The negative binomial is often a good approximation to the distribution of such data, and the mean is a suitable summary measure [34]. For this, $\mu_1 = 50$, $\mu_0 = 71.4$ (30% vaccine efficacy), $k_0 = k_1 = 0.33$, $Q_0 = Q_1 = 0.5$, a null hypothesis of both means being equal to 71.4, 90% power and 5% significance level (two-sided). From equation (7) we again obtain a sample size of 505 per arm. From equation (8) we obtain 505 once more.

This is because the methods differ in terms of the form $1/\mu + 1/k$ and, for this example, $1/k$ dominates $1/\mu$, and k did not change. With the same parameter values, the normal approximation in equation (9) gives 531 per arm.

Two sets of simulations were done: a) k was allowed to vary from 0.1 to 10, with the Poisson as a final limiting case ($k = \infty$); b) the efficacy, i.e. 1-(μ_1/μ_0), was allowed to vary from 0.3 to 0.7. Otherwise the parameters were held constant. The results are shown in Figure 2, where each data point is based on 10,000 simulations. For 30% vaccine efficacy, using the log link maintains close to the nominal power and the identity link is only slightly conservative (upper panel). As the efficacy, and the difference between the means, increases, the log link still maintains close to the nominal power whereas the identity link over-estimates the sample size, by more than 50% for the largest values of efficacy (lower panel).

GLM approach for Poisson distribution

Equations (10), on the log scale, and (11), on the untransformed scale, were compared, with the power again set at 90%, and with three values of the mean in the control arm (μ_0): 5, 2 and 0.2. Again using 10,000 simulations for each combination, the results are shown in Figure 3. The two methods are similar, and both slightly conservative for higher efficacies; the log link slightly more so.

GLM approach for binomial distribution

Similar simulations were done for equations (12) and (13) with $d = 1$ and various values of μ_0 and efficacy (1 minus the odds ratio). As before, data for each set of values was simulated 10,000 times. For $\mu_0 = 0.5$ both methods give close to nominal power. For $\mu_0 = 0.1$ and 0.05, the pattern is similar to the smaller Poisson means, with both being slightly conservative for higher

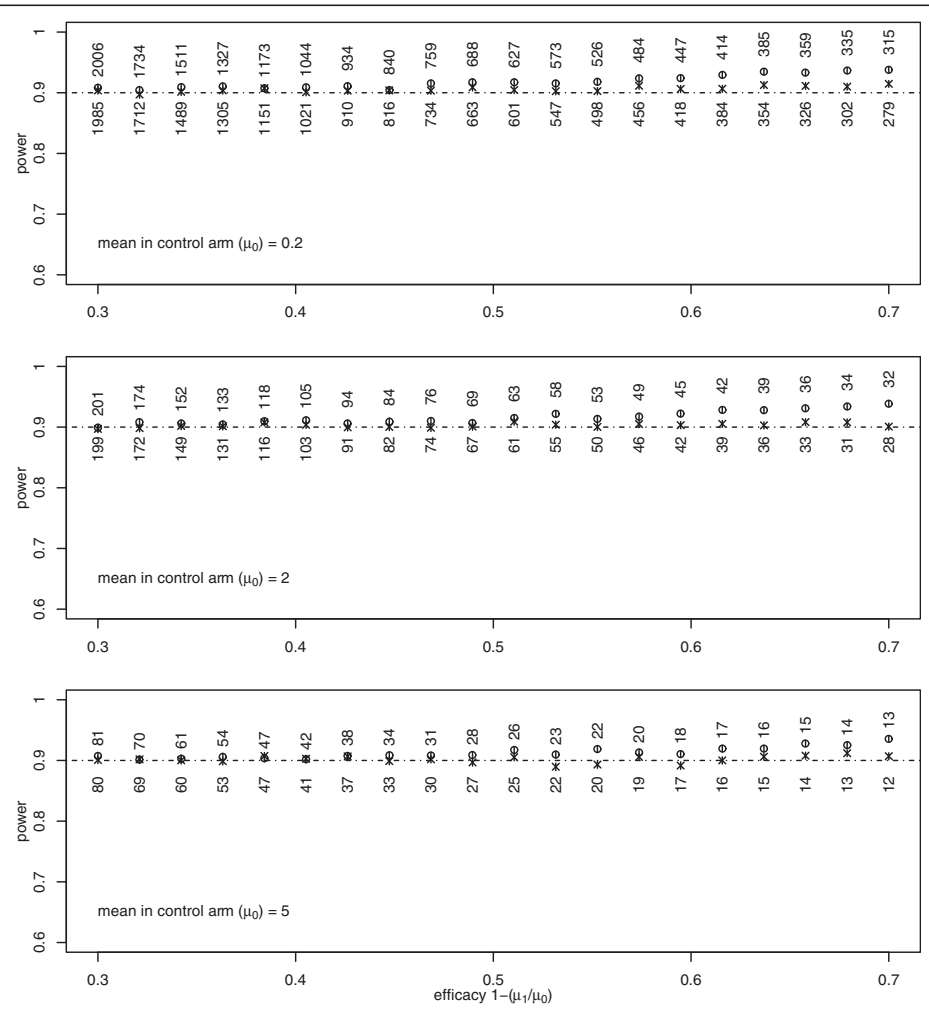

Figure 3 Poisson. This is similar to the lower panel of Figure 2, with each panel comparing two Poisson means. In each panel, the value of μ_0 is shown, and $\mu_1 = \mu_0 \times$ (1-efficacy).

efficacies; the logit link slightly more so (not shown). The simple dependency of the equations on d means that similar patterns were seen for d equal to 5 and 10 (not shown).

GLM approach for gamma distribution

Here we use an example based on concentrations of the insecticide deltamethrin on hammock nets in the Colombian Amazon [35], the mean being 8.46 mg/m^2, and κ estimated as 0.639. As before, we compare the power of sample sizes from equation (14) with those from the corresponding normal approximation on the original scale. The results are shown in Figure 4. As in Figure 2, the sample size calculated on the scale of the link function maintains close to nominal power, while the normal approximation over-estimates the necessary sample size, by 50% or more for the larger differences in means.

In this case, the likelihood ratio test resulted in higher estimated powers for both tests (not shown). Since the sample size inputs were the same for both test methods, the difference scale again had appreciably more power than the logarithmic scale.

Summary of simulation results

For the Poisson and binomial distributions, the results show little or no advantage for sample size calculations on the scale of the link function, i.e. log rates or log-odds, as opposed to the difference in rates or in proportions. By contrast, for the negative binomial and gamma distributions, which have additional parameters which can reflect skewness, sample size calculations based on differences in means can be very conservative, giving larger numbers which substantially exceed the required power. Sample size calculations on the log scale, however, retain close to the nominal power for the examples studies.

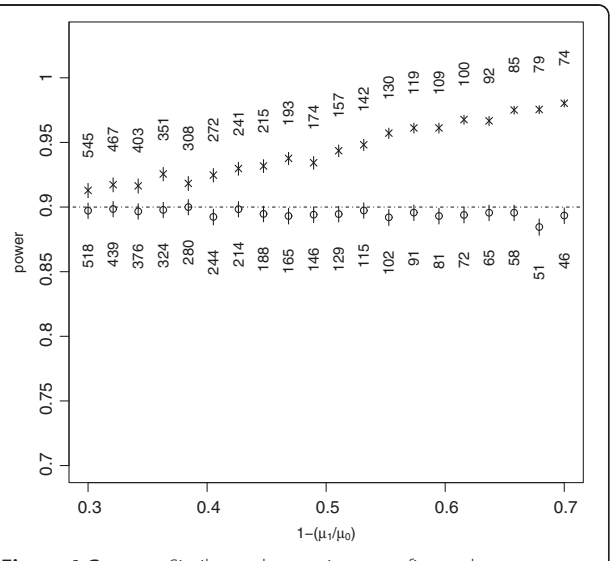

Figure 4 Gamma. Similar to the previous two figures but comparing means of two gamma distributions, with parameters based on a study of the insecticide deltamethrin on hammock nets.

Discussion

Normal approximations to distributions are often used to estimate sample sizes for discrete data, even when the data are to be analysed by generalized linear models. As well as being logically inconsistent, the magnitude of error is potentially large, judging by the discrepancies in CDF between the normal approximation and the exact distributions, whether assessed by the Berry-Esséen theorem or directly from distribution functions. This tends to sustain concerns about lack of robustness of normal approximations. Berry-Esséen and related theorems can, in principle, be used to estimate the speed of convergence of the normal approximation to that specified by the central limit theorem [20,22]. However, their bounds proved to be often markedly wider than those obtained from computing the CDF of the relevant distribution.

Considering robustness at the analysis stage, the t test performs well under certain large departures from normality [36]. Nevertheless, it is liable to break down when 'skew is severe or when population variances and sample sizes both differ' [37,38]. These are the circumstances for which we suggest the methods presented in the current paper are most suitable. The negative binomial and gamma distributions can capture severe skewness, and their variances differ between samples if the means do, due to their variance functions ($V(\mu)$). We have used examples related to parasitology and entomology, but numbers of events, such as clinic visits or epileptic fits can also yield skewed count data. On the other hand, if a particular distribution family cannot be assumed then methods are available for sample sizes for non-parametric tests [39].

Under the simulation scenarios examined, where the proposed and standard methods differ, the latter tend to be conservative. The fact that many trials do not recruit their target sample sizes [40] may suggest acquiescence in such sample size over-estimation. However, compliance with the ethical requirement to avoid unnecessary exposure to novel treatments [41] — both to reduce potential harms, and to speed the acceptance of favourable interventions — would seem to be better assured by improving both the mathematical estimation and the recruitment process, rather than anticipating a tendency for their errors to cancel.

Some previous sample size methods for GLMs concentrate on single or multiple continuous predictor variables. They tend to be complex and do not always involve an explicit expression for the sample size. Here we have obtained simple equations for the comparison

of two means, which is the most common situation for clinical trials. For the negative binomial, the method shown here corresponds to Zhu and Lakkis 'Approach 2' [17], although we allow k to differ between the arms (our k is the reciprocal of Zhu and Lakkis'). The approach was motivated by the need to plan later phase trials of vaccines against hookworm [29], a disease whose morbidity is related to infection intensity which in turn is measured by faecal egg counts. The high skewness of these counts seemed to preclude the use of normal approximations [34]. Negative binomial modelling may be appropriate for other parasite species [42] and other types of count [28], including insects [43] disease episodes [44], lesions [45], and cells [46]. For this distribution, there is a visible correspondence between the current formulae and that given by Krebs for estimating a mean with given percentage precision [47]. In fact our approach does not require specification of the complete distribution but only the link and variance functions. For the gamma, another example in the hookworm vaccine trials was the use of faecal heme as a candidate secondary endpoint. This is likely to be roughly proportional to the number of adult worms in the gut, and a gamma distribution was found to be a good fit to available data. More generally, gamma GLMs are commonly used for analysis of data on costs and length of stay in health facilities [32]. Despite the typically high skewness of cost data, analysis of arithmetic mean is statistically valid, and relevant due to it being proportional to total cost [48]. Other continuous skewed variables, for which gamma GLMs can be used, include serum concentrations of lipids, cytokines or hormones [49,50].

Conclusions

The method seems most useful for the negative binomial and the gamma distributions which, depending on their parameters, can be highly skewed, making a normal approximation less accurate for the sample mean. Motivated by two biomedical studies, we have shown that the method can be advantageous. Generalized linear models are commonly used to compare means of non-normal distributions and our method is well aligned with this, as well as being simple to use. We hope it will prove useful for situations in which the response variable is expected to be highly skewed, and for which the accuracy of normal approximations are likely to be poor.

Additional files

Additional file 1: Expression of the Berry-Esséen bound in terms of the third non-absolute central moment and a finite sum.

Additional file 2: Third central moments and probability density function for non-Gaussian distributions.

Additional file 3: 'nGLM.r'. R code to implement the methods described in the paper. It can be opened in any text editor. Instructions are at the top of the file [51,52].

Abbreviations

CDF: Cumulative distribution functions; GLM: Generalized linear model; HHVI: Human Hookworm Vaccine Initiative.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

NA conceived the approach. Both authors carried out the algebraic and computational calculations for the Berry-Esséen bounds. NA derived the GLM-based sample size equations, and carried out the simulations of their performance. BC wrote the first draft of the paper. Both authors edited the paper. Both authors read and approved the final manuscript.

Acknowledgements

We are grateful to Karim Anaya-Izquierdo for useful discussion, and Irina Shevtsova and Stephen Walters as referees. This work was funded by a) the Albert B. Sabin Vaccine Institute, which receives support from the Bill and Melinda Gates Foundation, and b) the United Kingdom Medical Research Council (MRC) and Department for International Development (DFID) (MR/K012126/1). The Sabin Vaccine Institute ratified the decision by the authors to submit the report for publication, without seeking any changes to its content.

Received: 9 September 2014 Accepted: 24 March 2015
Published online: 02 April 2015

References

1. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Control Clin Trials. 1981;2:93–113.
2. Wong KS, Chen C, Fu J, Chang HM, Suwanwela NC, Huang YN, et al. Clopidogrel plus aspirin versus aspirin alone for reducing embolisation in patients with acute symptomatic cerebral or carotid artery stenosis (CLAIR study): a randomised, open-label, blinded-endpoint trial. Lancet Neurol. 2010;9(5):489–97.
3. Watson-Jones D, Weiss HA, Rusizoka M, Changalucha J, Baisley K, Mugeye K, et al. Effect of herpes simplex suppression on incidence of HIV among women in Tanzania. N Engl J Med. 2008;358(15):1560–71.
4. Kessler D, Lewis G, Kaur S, Wiles N, King M, Weich S, et al. Therapist-delivered Internet psychotherapy for depression in primary care: a randomised controlled trial. Lancet. 2009;374(9690):628–34.
5. Holland R, Lenaghan E, Harvey I, Smith R, Shepstone L, Lipp A, et al. Does home based medication review keep older people out of hospital? The HOMER randomised controlled trial. BMJ. 2005;330(7486):293.
6. Kaul R, Kimani J, Nagelkerke NJ, Fonck K, Ngugi EN, Keli F, et al. Monthly antibiotic chemoprophylaxis and incidence of sexually transmitted infections and HIV-1 infection in Kenyan sex workers: a randomized controlled trial. JAMA. 2004;291(21):2555–62.
7. Kirkwood BR, Sterne JAC. Essentials of medical statistics. 2nd ed. Oxford: Blackwell Scientific Publications; 2003.
8. van Belle G. Statistical rules of thumb. 2nd ed. Hoboken, N.J.: Wiley-Interscience; 2008.
9. Rosner B. Fundamentals of biostatistics. 7th ed. Boston: Duxbury Press; 2010.
10. Daly L, Bourke GJ. Interpretation and uses of medical statistics. 5th ed. Oxford: Blackwell Science; 2000.
11. Whittemore AS. Sample size for logistic regression with small response probability. J Am Stat Assoc. 1981;76(323):27–32.
12. Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. Stat Med. 1998;17:1623–34.
13. Vaeth M, Skovlund E. A simple approach to power and sample size calculations in logistic regression and Cox regression models. Stat Med. 2004;23(11):1781–92.
14. Alam MK, Rao MB, Cheng F-C. Sample size determination in logistic regression. Sankhya. 2010;72-B(1):58–75.
15. Signorini DF. Sample size for poisson regression. Biometrika. 1991;78:446–50.

16. Shieh G. Sample size calculations for logistic and poisson regression models. Biometrika. 2001;88(4):1193–9.
17. Zhu H, Lakkis H. Sample size calculation for comparing two negative binomial rates. Stat Med. 2014;33(3):376–87.
18. Self SG, Mauritsen RH. Power/sample size calculations for generalized linear models. Biometrics. 1988;44:79–86.
19. Shieh G. On power and sample size calculations for likelihood ratio tests in generalized linear models. Biometrics. 2000;56(4):1192–6.
20. Feller W. An introduction to probability theory and its applications. 2nd ed. New York: Wiley & Sons; 1971.
21. Shevtsova IG. On the absolute constants in the Berry–Esseen inequality and its structural and nonuniform improvements. Informatika i Ee Primeneniya [Informatics and its Applications]. 2013;7(1):124–5.
22. Korolev VA, Shevtsova I. An improvement of the Berry-Esseen inequality with applications to poisson and mixed poisson random sums. Scand Actuar J. 2012;2012:81–105.
23. Hipp C, Mattner L. On the normal approximation to symmetric binomial distributions. Theory Probability Appl. 2008;52(3):516–23.
24. Nagaev SV, Chebotarev VI. On the bound of proximity of the binomial distribution to the normal one. Theory Probability Appl. 2012;56(2):213–39.
25. Hilbe JM. Negative binomial regression. 1st ed. Cambridge: Cambridge University Press; 2007.
26. Zelterman D. Discrete distributions: applications in the health sciences. Chichester: Wiley; 2004.
27. McCullagh P, Nelder JA. Generalized linear models. 1st ed. London: Chapman and Hall; 1983.
28. Hilbe JM. Negative binomial regression. 2nd ed. Cambridge: Cambridge University Press; 2011.
29. Brooker S, Bethony JM, Rodrigues LC, Alexander N, Geiger S, Hotez PJ. Epidemiological, immunological and practical considerations in developing and evaluating a human hookworm vaccine. Expert Rev Vaccines. 2005;4(1):35–50.
30. Fox J. Applied regression analysis and generalized linear models. 2nd ed. Thousand Oaks, California: Sage Publications, Inc; 2008.
31. Alexander N, Cundill B, Sabatelli L, Bethony JM, Diemert D, Hotez P, et al. Selection and quantification of infection endpoints for trials of vaccines against intestinal helminths. Vaccine. 2011;29(20):3686–94.
32. Manning WG, Basu A, Mullahy J. Generalized modeling approaches to risk adjustment of skewed outcomes data. J Health Econ. 2005;24(3):465–88.
33. Evans M, Hastings N, Peacock B. Statistical distributions. 3rd ed. New York: Wiley; 2000.
34. Alexander N. Analysis of parasite and other skewed counts. Trop Med Int Health. 2012;17(6):684–93.
35. Rodríguez M, Pérez L, Caicedo JC, Prieto G, Arroyo JA, Kaur H, et al. Composition and biting activity of Anopheles (Diptera: Culicidae) in the Amazon region of Colombia in relation to mosquito net policy. J Med Entomol. 2009;46(2):307–15.
36. Heeren T, d'Agostino R. Robustness of the two-independent samples *t*-test when applies to ordinal scale data. Stat Med. 1987;6:79–90.
37. Boneau CA. The effects of violations of assumptions underlying the t test. Psychol Bull. 1960;57(1):49–64.
38. Stonehouse JM, Forrester GJ. Robustness of the t and U tests under combined assumption violations. J Appl Stat. 1998;25(1):63–74.
39. Noether GE. Sample size determination for some common nonparametric tests. J Am Stat Assoc. 1987;82(398):645–7.
40. Sully BG, Julious SA, Nicholl J. A reinvestigation of recruitment to randomised, controlled, multicenter trials: a review of trials funded by two UK funding agencies. Trials. 2013;14:166.
41. Bacchetti P, Wolf LE, Segal MR, McCulloch CE. Ethics and sample size. Am J Epidemiol. 2005;161(2):105–10.
42. Anderson RM, May RM. Infectious diseases of humans: dynamics and control. 1st ed. Oxford: Oxford University Press; 1991.
43. Nedelman J. A negative binomial model for sampling mosquitoes in a malaria survey. Biometrics. 1983;39:1009–20.
44. Mwangi TW, Fegan G, Williams TN, Kinyanjui SM, Snow RW, Marsh K. Evidence for over-dispersion in the distribution of clinical malaria episodes in children. PLoS One. 2008;3(5):e2196.
45. Aban IB, Cutter GR, Mavinga N. Inferences and power analysis concerning two negative binomial distributions with an application to MRI lesion counts data. Comput Stat Data Anal. 2008;53(3):820–33.
46. Periwal SB, Spagna K, Shahabi V, Quiroz J, Shroff KE. Statistical evaluation for detection of peptide specific interferon-gamma secreting T-cells induced by HIV vaccine determined by ELISPOT assay. J Immunol Methods. 2005;305(2):128–34.
47. Krebs CJ. Ecological methodology. 2nd ed. Menlo Park: Benjamin/Cummings; 1999.
48. Barber JA, Thompson SG. Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. Stat Med. 2000;19(23):3219–36.
49. Chaves PH, Xue QL, Guralnik JM, Ferrucci L, Volpato S, Fried LP. What constitutes normal hemoglobin concentration in community-dwelling disabled older women? J Am Geriatr Soc. 2004;52(11):1811–6.
50. Garcia-Broncano P, Berenguer J, Fernandez-Rodriguez A, Pineda-Tenor D, Jimenez-Sousa MA, Garcia-Alvarez M, et al. PPARgamma2 Pro12Ala polymorphism was associated with favorable cardiometabolic risk profile in HIV/HCV coinfected patients: a cross-sectional study. J Transl Med. 2014;12:235.
51. Balakrishnan N, Nevzorov VB. A primer on statistical distributions. Hoboken, New Jersey: Wiley-Interscience; 2003.
52. Mood AM, Graybill FA, Boes DC. Introduction to the theory of statistics. 3rd ed. New York: McGraw-Hill; 1974.