Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Hanley and Moodie, J Biomet Biostat 2011, 2:5
DOI: 10.4172/2155-6180.1000124

Research Article Open Access

Sample Size, Precision and Power Calculations: A Unified Approach

James A Hanley* and Erica EM Moodie

Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Canada

Abstract

The sample size formulae given in elementary biostatistics textbooks deal only with simple situations: estimation of one, or a comparison of at most two, mean(s) or proportion(s). While many specialized textbooks give sample formulae/tables for analyses involving odds and rate ratios, few deal explicitly with statistical considerations for slopes (regression coefficients), for analyses involving confounding variables or with the fact that most analyses rely on some type of generalized linear model. Thus, the investigator is typically forced to use "black-box" computer programs or tables, or to borrow from tables in the social sciences, where the emphasis is on correlation coefficients. The concern in the – usually very separate – modules or standalone software programs is more with user friendly input and output. The emphasis on numerical exactness is particularly unfortunate, given the rough, prospective, and thus uncertain, nature of the exercise, and that different textbooks and software may give different sample sizes for the same design. In addition, some programs focus on required numbers per group, others on an overall number. We present users with a single universal (though sometimes approximate) formula that explicitly isolates the impacts of the various factors one from another, and gives some insight into the determinants for each factor. Equally important, it shows how seemingly very different types of analyses, from the elementary to the complex, can be accommodated within a common framework by viewing them as special cases of the generalized linear model.

Keywords: Generalized linear models; Multipliers; Unit variance

Introduction

In planning statistical studies, investigators cannot turn to a single textbook source for guidance on sample size, precision and statistical power. Elementary textbooks usually deal only with the estimation of one, or a comparison of at most two, mean(s) or proportion(s). Although the common structure to the formulae for these two types of data is not emphasized in textbooks, it has been exploited in the "rough and ready" multi-purpose formula given by [1,2]. However, Lehr emphasized (memorization of) one equation, with the multiplier of 16 – for a specific power (80%) and a specific two-tailed significance level (α =0.05) – because these two test characteristics seem to "occur often in biopharmaceutical research." However, this "memorize a multiplier" approach does not allow the user to determine what the multiplier would be for other (alpha, power) configurations, or whether the formula refers to the total number of subjects, or just to the number in one of the two samples to be studied.

The most commonly used analyses in epidemiologic and medical research involve ratios of odds and rates as comparative parameters. Some guidance on the statistical precision of such comparisons is found in more advanced biostatistics textbooks [3,4]; a more detailed treatment is found in specialized textbooks dealing with specific epidemiologic designs [5,6] and in dedicated articles [7,8]. The "sample size requirements" are often presented in tables, or obtained from computer software such as EpiInfo [9]. With these, the user does not always have full control over all the specific input values used, and thus may not be able to see explicitly why the numbers change the way they do. Guidelines for sample size for survival analyses (involving ratios of medians or of hazards) tend to be found in yet other separate publications and software [10].

Very few biostatistics textbooks deal with sample size considerations for a coefficient (slope) in a simple or multiple regression. For these, users are thus forced to consult a specialized and unfamiliar text on power analysis in the social sciences [11]. In this "bible," the parameters of interest are correlation coefficients (simple and partial) for responses measured on continuous scales – rather than the more familiar regression slopes and binary responses that are more commonly used in epidemiological and clinical research. This lack of tools for regression analyses has been partly remedied by the recent inclusion in the software of calculations for studies involving linear regression – albeit only for responses measured on continuous scales [12,13].

User-friendly modules such as these are most welcome, especially if the algorithms they use are published and fully documented. However, separate calculations for each design, performed with considerable – and unnecessary – exactness, out of reach of the user, rather than in a spreadsheet or calculator, do not emphasize the common structure behind the seemingly different analyses. Nor do they emphasize the unity in modern epidemiologic and biostatistical analyses that can be achieved by viewing them as special cases of the generalized linear regression model [14]. Whereas this model is one the major statistical developments of the forty years, and nowadays routinely used in data-analysis, it is seldom used in the planning of the size of an investigation.

This note presents a single universal – even if in some instances slightly "inexact" – formula that (a) presents one overall sample size as a simple product of terms, one term per factor (b) motivates, tabulates and makes explicit the impact of each factor on precision and sample size (c) allows one to obtain a closed form expression for the value of any one factor in terms of the values of the remaining ones, (d) accommodates seemingly very different types of designs and data-analyses within a single generalized framework. The approach could easily be extended to other design features not discussed here.

We proceed as follows: first, we give a generic inequality in terms of the standard error of the estimate of the parameter of interest. We first focus on the corresponding overall sample size if the comparative parameter of interest is the slope in a multiple linear regression involving a response (Y) variable measured on a continuous scale. We then disaggregate the expression to show its distinct components,

***Corresponding author:** James A Hanley, McGill University, 1020 Pine Ave, West, Montreal, Quebec, H3A 1A2, Canada, E-mail: james.hanley@mcgill.ca

Received September 21, 2011; **Accepted** November 10, 2012; **Published** January 20, 2012

Citation: Hanley JA, Moodie EEM (2011) Sample Size, Precision and Power Calculations: A Unified Approach. J Biomet Biostat 2:124. doi:10.4172/2155-6180.1000124

and to show that its a special case of a generic structure. Finally, we add to, or modify, the expression for a broad range of analyses. Numerical examples are provided in some early sections to ensure that the proposed procedure is clear, but omitted from most of the later sections.

The generic form of a sample size formula

As is illustrated in Figure 1, for a 2-sided test with a false positive rate of α to have at least 100(1-β)% power against a non-zero difference, Δ, the study configuration must satisfy the inequality

$$Z_{\alpha/2}SE_{null} + Z_{\beta}SE_{alt} < \Delta,$$

where Δ is the difference between the non-null (alternative) and null values of the parameter of interest, and SE_{null} and SE_{alt} are the standard errors of the estimate of this parameter in the null and alternative scenarios, respectively. For didactic purposes we simplify the relationship by taking an "average" standard error, so that the relation can be approximated as

$$(Z_{\alpha/2} + Z_{\beta})SE < \Delta,$$

or, more usefully, in variance terms, as

$$(Z_{\alpha/2} + Z_{\beta})^2 \times \text{Var}[\text{Parameter Estimate}] < \Delta^2. \quad (2.1)$$

Some modifications to the functions of α and β are needed if sample sizes are very small, e.g., when a *t*-(rather than a *z*-) distribution is more appropriate. Since in most situations, these modifications are minor and only detract from the 'big picture,' we will ignore them.

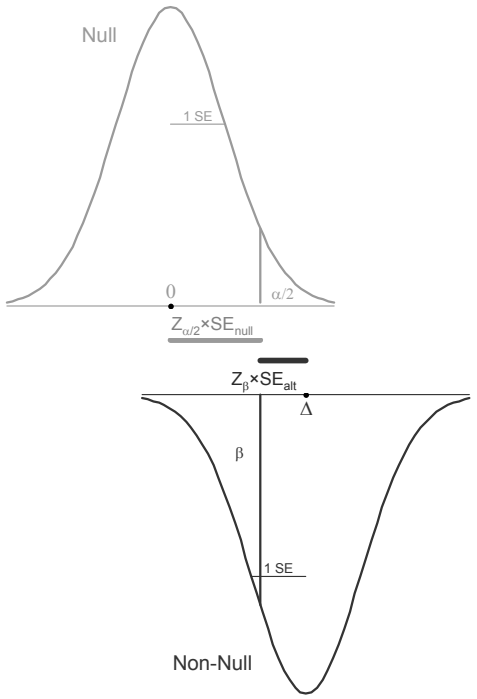

Figure 1: *Basis for sample size formulae*. Shown in a lighter colour, and with the usual orientation, is the sampling distribution of the test statistic under the null scenario. Shown in the darker colour, and upside down to distinguish its landmarks more easily, is the sampling distribution under the alternative (non-null) scenario. Although they are in this instance, the standard errors under the null and alternative will not necessarily be equal. The two distributions are separated by an amount Δ, the difference in the (comparative) parameter of interest. For a 2-sided test with a false positive rate of α to have at least 100(1-β) power against a non-zero difference Δ, the two absolute distances $Z_{\alpha/2}SE_{null}$ and $Z_{\beta}SE_{alt}$ must sum to less than Δ. In this example, α= 0:05 (so that $Z_{\alpha/2}$= 1:96) and (1-β) = 0.8 (so that Z_{β}= -0:84).

Up from multiple linear regression

Even though it may at first seem like an unusual point of departure, consider the situation where the parameter of interest is the slope in a multiple linear regression of a response (Y) variable on an X variable of interest, while "adjusting for" one of more confounding variables, denoted collectively as C. Suppose Y, X, and C are measured on numerical (but not necessarily continuous) scales. As explained in (unfortunately few) regression textbooks [15] an approximation to the sampling variation associated with the slope, estimated from n values of the response Y, measured at the n (not necessarily distinct) values

$X = \{X_1, X_2, ..., X_n\}$ is given by the expression

$$\text{Var}[(\text{all possible}) \text{ slope estimates}] \cong \frac{\text{Var}[Y \mid X]}{n \times \text{Var}[X_1, .., X_n] \times (1 - r^2_{X \text{ with } C})},$$

where $\text{Var}[Y \mid X]$ is the (presumed homogeneous over X) variance of the (infinite number of) possible Y values at each X value, and $r^2_{X \text{ with } C}$ is the square of the simple/multiple correlation of X with C. This expression illustrates how the slope is more volatile the larger the variation of the Y values from the true line, the greater the correlation of X with other influential factors, and the smaller the sample size n. It is less volatile the larger the spread of the n X values.

Substituting this specific sampling variance for the slope estimate into the generic form in the previous section, and re-arranging terms so as to isolate n, we obtain

$$n > (Z_{\alpha/2} + Z_{\beta})^2 \times \text{Var}[Y \mid X] \times \frac{1}{\text{Var}[X_1, .., X_n]} \times \frac{1}{(1 - r^2_{X \text{ with } C})} \times \frac{1}{\Delta^2}, \quad (3.1)$$

where Δ refers to the difference in the true effect of X on $E(Y \mid X)$, i.e., the difference in the slope, $\beta_{Y|X}$, under the null and alternative scenarios.

Although this expression seems to be specific to a regression approach, it also deals implicitly with several other contexts and designs. In the following sections, we will take advantage of this form – a product of several separate factors applied to the base-sample size shown in Table 1 – to emphasize the generality of our approach.

Special case: Sample size formulae for difference of means and proportions

The difference of two sample means $\bar{y}_1 - \bar{y}_0$ is equivalent to the slope of the simple regression equation fitted to the n (= $n_1 + n_0$) datapoints $\{X_1, y_1\}$, ..., $\{X_n, y_n\}$, with each of the n accompanying X's given the numerical value of 1 to indicate an observation from sub-population 1, or 0 to indicate sub-population 0. Since X is a 2-point or binary variable, the true slope β_X is simply the difference between the mean Y in the sub-population where X = 1, i.e., $\mu_{X=1}$, and its counterpart $\mu_{X=0}$, i.e., $\beta_X = (\mu_{X=1} - \mu_{X=0})/1$.

If the two sample sizes are equal, i.e., if in the simple regression, half

Significance level (α)		Power (1-β)				
Two-sided	One-sided	0.5	0.8	0.9	0.95	0.99
0.20	(0.10)	1.7	4.6	6.6	8.6	14
0.10	(0.05)	2.8	6.6	8.6	11	16
0.05	(0.025)	3.9	7.9	11	13	19
0.01	(0.005)	6.7	12	15	18	25
0.001	(0.0005)	11	18	21	25	32

Table 1: Base sample size, $(Z_{\alpha/2} + Z_{\beta})^2$, as a function of significance level (α) and power (1 − β).

of the n observations have the value $X = 1$ and half the value $X = 0$, then $\bar{X}$ =0.5, and the "unit" variance of these n values of X is

$$\text{Var}[X_1, ..., X_n]=(-0.5)^2\times(1/2) + (+0.5)^2\times(1/2) = 1/4.$$

Upon substituting the reciprocal, (1/(1/4)) = 4, as the third component into Equation (3.1), the inequality, in terms of the total sample size n, becomes

$$n_{\text{total}} > (Z_{\alpha/2} + Z_{\beta})^2 \times \text{Var}[\text{Y 's in the same sub-population}] \times 4 \times (1/\Delta)^2.$$

Many textbooks, such as [16], write it in the equivalent form for the per group sample size

$$n_{\text{per group}} > 2\times (Z_{\alpha/2} + Z_{\beta})^2 \times \text{Var}[Y\text{'s in the same sub-population}]/\Delta^2.$$

This form confuses many investigators, since they think that the '2' in the formula 'already takes care of' the 2 groups (in reality, the '4' is a special case of a much more general formula). Armitage et al., page 140 [3], using $\text{Var}[Y|X]=0.25$ and $\delta = 0.25$ calculated an n per group of 62.8. Our 'separate multipliers' approach (with the 7.9 for Table 1) suggests a total sample size of $7.9 \times 0.25 \times 4 \times \{1/(0.25^2)\} = 126.4$.

The corresponding approximate formula for a comparison of two proportions (expressed as a risk difference, RD, or prevalence difference, PD) can be obtained by replacing the unit variance Var[Y 's in the same sub-population] by the Bernoulli unit variance $\pi(1 - \pi)$, where π is the expected proportion of observations where $Y = 1$. The unit variances are given in the first row of Table 2 (the other rows will be discussed later). Then, with Δ the difference in risk or prevalence under the null and alternative scenarios,

$$n_{total} > (Z_{\alpha/2} + Z_{\beta})^2 \times [\pi(1 - \pi)] \times 4 \times (1/\Delta)^2.$$

The reason this is an approximation goes back to our simplification, stated at the outset, in which we forced the standard error to be the same under the null and the alternative scenarios. Save for a few exceptional cases, the null and alternative standard errors differ when we are dealing with binary Y's, since the unit variance $\pi(1 - \pi)$ is a function of π itself. Under the null, the standard error of an estimate of a difference in proportions is a function just of $\pi_0(1 - \pi_0)$, but under the alternative it involves a combination of both $\pi_0(1-\pi_0)$ and $\pi_{\text{alt}}(1-\pi_{\text{alt}})$. In practice, to simplify matters, we – just as many authors do – use a common unit variance, evaluated at some intermediate, or "average," value of π. For non-extreme values of π, and unless $Z_{\alpha/2}$ and Z_{β} are very different in magnitude, this simplification introduces only a slight approximation. It is however just one of the many sources of differences between the sample sizes obtained with different formulae or software programs.

Using π_1 =0.25, π_2 =0.35, α =0.05, and β =0.1, and no continuity correction, [3], page 142, calculated an *n per group* of 439. The product of 11 from Table 1; a 'unit variance', calculated at π =0.3, of $\pi(1 - \pi)$=0.21; the reciprocal of the 'unit variance', of the group indicators (the x's) of 4; and $(1/\delta)^2 = 100$, yields 11×0.21×4×100 = 924. [3] repeated their calculation using the continuity-correction used in the tabulations of Fleiss, and in Table A8 of their own textbook, and obtained a total of 918. Not surprisingly, given that Fisher's exact text and chi-squred tests yield similar results when the numbers of events are sizable, the table of [17] also yielded a total of 918 in this application, whereas it would not do so in other situations.

An insight from use of a regression approach: dealing with unequal sample sizes

Although such designs are commonly used, few textbooks give – or if they do, adequately motivate – the corrections needed for comparisons involving *unequal* sample sizes. However, the correction can easily be derived and understood by examining the term

$$\frac{1}{Var[X_1,...,X_n]}$$

in Equation (3.1) for the variance of a slope. For, if a proportion $P_{X=1}$ of the observations are from group $X = 1$, and the remainder $1 - P_{X=1}$ are from group $X = 0$, then the reciprocal of the unit variance of the group 1 indicator values can be written as

$$\frac{1}{Var[X_1,..,X_n]} = \frac{1}{P_{X=1}\times(1-P_{X=1})} = \frac{1}{P_{X=1}} + \frac{1}{P_{X=0}}.$$

This multiplier takes on a minimum value of 4 when $P_{X=1}$ =1/2 and

Scale	Unit variance	Proportion (π) 0.05	0.1	0.15	0.2	0.3	0.4	0.5
Proportion, π (Risk difference)	$\pi(1-\pi)$	0.048	0.090	0.13	0.16	0.21	0.24	0.25
Logit: $\log(\pi/(1-\pi))$ (Odds ratio)	$[\pi(1-\pi)]^{-1}$	22	12	7.9	6.3	4.8	4.2	4.0
Log: $\log(\pi)$ (Risk or rate ratio)	$(1-\pi)/\pi$	19	9.0	5.7	4.0	2.4	1.5	1.0

Table 2: Sample size multiplier (unit variance) when studying a difference or ratio of two proportions, or the ratio of their odds. Multipliers are shown as a function of the scale involved and of the value of response proportion (π), rounded up, and to two significant digits. As in generalized linear models, the multiplier for the logit and log of the proportion were derived by multiplying the unit Bernoulli variance under the identity link function by the Jacobian associated with the change of scale.

	Relative sample sizes 50:50 (1:1)	60:40 (1.5:1)	67:33 (2:1)	75:25 (3:1)	80:30 (4:1)	83:17 (5:1)	91:9 (10:1)
$[P_{X=1} \times (1 - P_{X=1})]^{-1}$	4.0	4.2	4.5	5.4	6.3	7.2	13

Table 3: Sample size multiplier, $1/\text{Var}[X_1, ..., X_n]=[P_{X=1} \times (1 - P_{X=1})]^{-1}$, when X is binary, as a function of relative sizes of samples in which $X = 1$ and $X = 0$. Multipliers rounded up, and to two significant digits.

$P_{X=0}$=1/2, and increasingly higher values the more the sub-sample sizes differ from each other. As can be seen in Table 3, the correction for this "inefficiency" is slight if the split is no worse than 60:40 (1/0.24 = 4.2, only 5% larger than the minimum of 4) but increases rapidly as it becomes more extreme: e.g., 1/0.21 = 4.7, or 19%, if 70:30; but 1/0.16 = 6.3, or 56% if 80:20; and 1/0.09 = 11, or 178%, if 90:10. Some readers may have seen this 'law' expressed in a formula for the *smaller* of the two sample sizes,

$$n_{\text{smaller}} = \frac{k+1}{k} \times (Z_{\alpha/2} + Z_{\beta})^2 \times (\sigma / \Delta)^2,$$

where k is the the ratio of the larger to the smaller sample size. Or, they might have calculated the total sample size, as [3] do, using the formula

$$n_{\text{total if unequal n's}} = \frac{(k+1)^2}{4k} \times n_{\text{total if equal n's}} \cdot$$

In our approach, one obtains the required total directly by using the reciprocal of the variance of the x's as the multiplier; this approach also encourages end-users to think of even the simple comparison of two means as just a special case of a general regression model – one with a binary group-indicator variable as the x. Ury HK [18] was one of the first to consider the statistical efficiency for "case-control" comparisons of both means and proportions; his $2k/(k + 1)$ efficiency ratio, for a study with k controls per case, relative to 1 control per case is widely cited – it ranges from 1 at $k = 1$ to an asymptote of 2 at $k = \infty$.

Again, when the outcome is binary – i.e., when comparing proportions – unless one simplifies matters by taking a common standard error that is intermediate between that under the null and alternative scenarios, it is not possible to cleanly separate the effect of the ratio k from that of π_0 and π_{alt} .In practice, as can be seen by studying [19], the modifications to deal with this seem hardly worth the additional complexity, particularly in view of the tentative and approximate nature of sample sizes to begin with.

Sample size for testing a simple correlation

By definition, the correlation, ρ, between two variables X and Y is independent of the units in which they are measured, and so the sample size depends only on the correlation itself. On the scale $T[r] = \frac{1}{2}\log\frac{1+r}{1-r}$ introduced by Fisher, $\text{Var}(T[r]) = \frac{1}{n-3}$. Thus, equation (2.1) can be re-arranged as $n = (Z_{\alpha/2} + Z_{\beta})^2 \times (1/\Delta)^2$,

where

$$\Delta = \frac{1}{2} \times \log\left[\frac{(1+\rho_{alt})/(1-\rho_{alt})}{(1+\rho_{null})/(1-\rho_{null})}\right].$$

Even though Fisher's transformation is not found in the generalized linear model packages, its 'canonical' form does map r into the full $-\infty$, ∞ scale. Table 4 gives the $(1/\Delta)^2$ multipliers based on this transformation. Use of this transform should discourage end-users from using the *null* variance in confidence interval and sample size calculations involving r.

The sample size cost of having to adjust for confounding

A multiple regression, or an analysis of covariance, is often used to remove the bias that would be generated by omitting an important covariate (C) that is correlated with the determinant (X) of interest. In the simplest case, where X is binary, the fitted regression coefficient $b_{X|C}$ in the multiple linear regression is algebraicly equivalent to subtracting a correction factor from the fitted 'crude' regression coefficient

$$b_X = \bar{y}_{X=1} - \bar{y}_{X=0}\text{ , i.e., [20]}$$

$$b_{X/C} = \left(\bar{y}_{X=1} - \bar{y}_{X=0}\right) - b_{X/C} \times \left(\bar{C}_{X=1} - \bar{C}_{X=0}\right).$$

The adjustment in sample size to achieve the same power as when there is no C is also in the form of a multiplier, $1/\left(1-r^2_{X\text{ with }C}\right)$, which some will recognize from courses in multiple regression as the variance inflation factor or VIF. The multipliers are given in Table 5.

We could equally call them the "sample size inflation factors" since, statistically speaking, we counteract variance by sample size, i.e. they are inverses of each other. To help understand how the control of confounding comes at a sample size cost, and to make a link with sample size formulae for simple linear regression, it may help to take a simplified example, and to borrow again from the regression framework underlying, and unifying, all of the formulae presented here. Consider a non-experimental study (cross-sectional) that investigates, in a multiple linear regression, how Y , the amount of hearing loss, is influenced by X, the number of years employed in a noisy workplace, while having to take account of C, the worker's age. Conceptually, using C in the multiple regression is similar to combining the slopes from the C-specific simple linear regressions of Y on X. Unless the sampling of subjects is designed to ensure the same (full) range of X's in each age-band, the naturally high correlation between X and C will produce a range of X at each value of C which is narrower than the full variance Var[X]. In fact, the C-specific variance of X is only $\text{Var}[X]\times\left(1-r^2_{X\text{ with }C}\right)$. Also, in a simple linear regression, the variance of the slope is governed by the reciprocal of the variance of the X's at which the Y observations are made. Thus the decreased precision (or the increased sample size required for the same precision) for the regression coefficient for X in the multiple regression can be seen as a case of a decreased effective range of X in a series of C-specific simple regressions of Y on X.

	ρ_{alt}								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
ρ_{null}									
0	99	24	10	5.6	3.3	2.1	1.3	0.8	0.5
0.1		95	23	9.6	5	2.8	1.7	1	0.5
0.2			88	20	8.3	4.2	2.3	1.2	0.6
0.3				77	17	6.8	3.2	1.6	0.7
0.4					63	14	5.1	2.2	0.9
0.5						48	9.9	3.3	1.2
0.6							33	6.1	1.6
0.7								19	2.7
0.8									7.2
	entries are $1/(T[\rho_{alt}] - T[\rho_{null}])^2$ where $T[\rho] = (1/2)\log[(1+\rho)/(1-\rho)]$.								

Table 4: Sample size multipliers, $(1/\Delta)^2$, for testing a correlation; rounded up, to two significant digits.

	(Multiple) Correlation, $r_{X\text{ with }C}$, between X and covariate(s) C							
	0.0	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$[1 - r^2_{X\text{ with }C}]^{-1}$	1.0	1.1	1.2	1.4	1.6	2.0	2.8	5.3

Table 5: Sample size multiplier to cover the cost of adjusting for one or more confounding variables.The multiplier, $[1-r^2_{X\text{ with }C}]^{-1}$ is given as a function of the (multiple) correlation, $r_{X\text{ with }C}$, between X and the set of confounding variables C. It is rounded up, and given to two significant digits.

Technically speaking, when covariates – whether or not they are confounders – are included in a multiple regression, the sample size formulae involves the smaller Var[Y |X,C], rather than Var[Y |X]. Given the scanty reporting of X-specific or X-and-C-specific summary statistics in many research reports, it is difficult to find either of these in the literature. Thus, in practice, when planning a new study, one often must make do with estimates of the – even larger – unconditional Var[Y]. In randomized trials, where X and C are relatively uncorrelated, one may still wish to reduce the within-group variation Var[Y |X] by using C, in addition to X, in a two-way analysis of variance, or – if C is numerical – in an regression-based analysis of covariance. If the analysis will involve the identity link and Gaussian variation, then using Var[Y] or Var[Y |X] rather than Var[Y |X, C] to plan sample sizes will likely underestimate the actual power [21]. In non-experimental studies, if there is considerable confounding, i.e., if $r^2_{X \text{ with } C}$ is much greater than zero, the gains in precision from including C (and X!) in the model may be partly offset by the variance inflation factor. The net result can often only be guessed at, unless one can provide estimates of the various components from analyses of one's own earlier data from similar situations.

The sample size cost of using mis-measured responses (Y's)

Fleiss [22] emphasizes the cost of unreliable responses by multiplying the sample sizes needed in the case of error-free (Y) measures by the reciprocal of the reliability coefficient R_Y. This correction stems from the definition of the reliability coefficient as the ratio of the variance of the error-free Y to the variance of the error-containing $Y^* = Y + \varepsilon$, i.e.,

$$R_{Y*} = ICC_{Y*} = \frac{Var[Y]}{Var[Y^*]} = \frac{Var[Y]}{Var[Y] + Var[\varepsilon]}.$$

Thus, since the power of the study will be influenced by the variation in the Y^*'s rather than in the Y's, the n_{total} calculated based on an error-free (Y) needs to be inflated by a factor of $\frac{1}{R_{Y*}}$. Thus, if the reliability is expected to be 0.8, 0.6, 0.4, or 0.2, the planned sample size needs to be inflated by a factor of 1.25, 1.67, 2.5 and 5.00 respectively.

The sample size cost of using mis-measured X's

Imprecision in the measured X's has a more insidious effect of attenuating the fitted regression coefficients [23]. Suppose, for example, that the true relation, involving error-free X's, is

$$E[Y|X+1] - E[Y|X] = \beta_{Y|X}.$$

The use of error-containing, X^*'s instead of error-free X's, will, on average, lead to an observed slope that is closer to the null, i.e,

$$E[Y|X^*+1] - E[Y|X^*] = \beta_{Y|X} \times ICC_{X^*},$$

where ICC_{X^*} is the reliability of the X^* measurements. In this situation, the required sample size from equation (1) – or the detectable effect – should be multiplied by the sample-size-inflation-factor, $1/ICC_{X^*}$.

When the parameter of interest is the correlation coefficient, $\rho_{Y,X}$, the attenuation involves the reliability of both the Y's and the X's:

$$\rho[Y^*, X^*] = \rho[Y, X] \times \sqrt{ICC_{Y*}} \times \sqrt{ICC_{X^*}}.$$

Thus, the required sample size should be multiplied by the two sample-size-inflation factors

$$\frac{1}{\sqrt{ICC_{Y*}}} \text{ and } \frac{1}{\sqrt{ICC_{X^*}}}$$

Prevalence-, risk-and odds-ratios

The log of an odds ratio (OR) is the difference between the logs of the two compared odds, and can be represented as the slope of the regression of the *logits* on the values (0/1 or continuous) of the X (exposure) variable. The use of logits transforms the traditional response parameter scale of $0 < \pi = P(Y = 1) < 1$ to the larger scale $-\infty < \log[\pi/(1-\pi)] < \infty$. This enlargement of the response scale also enlarges "unit" variances associated with Y. Thus, as is highlighted in the later rows of Table 2, the unit (Bernoulli) variance of $\pi(1-\pi)$ in the general sample size expression must be replaced by the unit variance in the logit(π) scale, namely $1/[\pi(1-\pi)]$, to give

$$n_{\text{total}} > (Z_{\alpha/2} + Z_{\beta})^2 \times \frac{1}{\pi(1-\pi)} \times \frac{1}{Var[X_1, .., X_n]} \times \frac{1}{\Delta^2}.$$

and – because of the change of scale – the Δ in Table 3 refers to the difference in the true log odds (i.e., the log of the odds ratio) under the null and alternative situations. There is a large literature on sample size calculations for logistic regression with continuous covariates: see [24] for a recent example and for links to earlier work. If, instead, the focus is on a simple ratio of two proportions, inference is usually carried out in the log scale. Thus, Table 5 also shows the unit variance in this scale, and the Δ in Table 6 refers to the difference in the true log prevalence or risk ratio under the null and alternative situations.

Outcome-based sampling: Not all studies have an '$X \rightarrow Y$' design. When either $Y = 1$ or $Y = 0$ is uncommon, it may be more efficient to use 'outcome-based' sampling, by merely assembling sufficient numbers of instances of each of $Y = 1$ and $Y = 0$ and then measuring X in each instance. In epidemiology, this study design is often called a case-control study [6, 5] whereas in economics and marketing, it might be called choice-based sampling [25].

In the planning of – but not the analysis of the data from – such studies, it helps to reverse the usual roles of X ("exposure") and Y ("outcome"), so that $Y' = X$ and $X' = Y$. Although, conceptually, we wish to compare the event-rate in the exposed with that the unexposed, most authors calculate the sample size for case-control studies by 'comparing the cases with the controls' with respect to the proportion exposed. Thus, in their example 4.16, [3] use the expected 1:4 distribution of exposed:nonexposed in the general population, and the postulated rate ratio of 2, to calculate that the distribution in the cases would be 1:2. Thus, they converted the calculation of the required number of cases, and an equal number of controls (a 1:1 ratio), into a comparison of two exposure proportions, with values 0.20 and 0.20 under the null, and 0.20 and 0.33 under the alternative. With α =0.05, and β =0.2, they calculated that the required numbers are 187:187 (374 in all).

Although our approach anticipates that in the actual data-analysis,

	Risk difference (Δ)								
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
Δ^{-2} :	400	100	45	25	16	11	8.2	6.3	4.9
	Odds, risk, or event-rate ratio (e^{Δ})								
	1.1	1.25	1.5	1.75	2.0	2.5	3.0	4.0	5.0
Δ^{-2} :	110	20	6.1	3.2	2.1	1.2	0.83	0.52	0.39

Table 6: Sample size multiplier, Δ^{-2}, when studying a difference or ratio of two proportions, or the ratio of their odds, or the ratio of two event-rates. Multiplier is given as a function of the differerence, Δ, in the scale in question, between the alternative and null values of the comparative parameter.

statistical tests and interval estimates will be based on differences in logits, we too take advantage of the fact that planning for a certain precision using a case:control (Y =1 : Y =0) ratio in an outcome-based-sampling study is analogous to planning the distribution of X in an 'X' → Y'' design. Thus, in the just-cited example, a ratio of 1 control (Y = 0) per case (Y = 1) would be represented as $P_{X'=1}$ =1/2 and $P_{X'=0}$ =1/2, yielding an X'-associated multiplier of 1/(1/2) + 1/(1/2) = 4. Likewise, an (average) exposure (X) prevalence of 25% or π =1/4, say, would be represented as a Y'-associated multiplier of

$$1/[\pi(1-\pi)] = 1/(1/4) + 1/(3/4) = 5\frac{1}{3}.$$

(Of note is that, expanded, the product of these two multipliers, is 1/(1/8) + 1/(1/8) + 1/(3/8) + 1/(3/8): this sum, divided through by the overall sample size n, has the same form as Woolf's formula for the variance of a log odds ratio. The fact that the product of the two multipliers is the same no matter which item (exposure or case/control status) is assigned the role of X and which the role of Y reflects the invariance of the odds ratio with respect to the sampling strategy.) Thus, the total number of subjects is the product of 7.9 from Table 1, the 4 and $5\frac{1}{3}$ reflecting the outcome and exposure distributions, and the 2.1, representing the `signal', from the last row of Table 6. It comes to 354. If unsure whether to calculate the unit variance for Y′ at an 'average' exposure of 0.20 say, or at 0.265 = (0.2 + 0.33)/2, or at 0.33 it is safer to take the most conservative one: an average of exposure prevalence of say 20% yields a unit variance (multiplier) for Y′ of 1/(1/5) + 1/(4/5) = $6\frac{1}{4}$ rather than $5\frac{1}{3}$, giving a result of 414, slight larger than the 374 derived by [3].

Although our approach may seem somewhat complex for what appears to be a simple comparison of two proportions, the fact is that odds ratios from case-control studies are seldom derived from simple 2 × 2 tables taught in introductory epidemiology courses; nowadays, they are typically derived from multiple logistic regression. And while the above logit-based sample-size calculation does not include any adjustment variables, it does come closer to actual analysis practice. A second advantage of our general approach is that its covers all possible design configurations, not just those are typically tabulated. Because controls are more readily available, and increase statistical power, investigators typically assemble more controls than cases. Yet, tables in textbooks limit themselves to control:case ratios of 1 or 2 or 4 (Table A9 of [3] only covers the 1:1 design, α =0.05 and β =0.1/0.2.) In our approach, ratios of k = 3 – or 5, or 20 – controls (Y = 0) per case (Y = 1) are easily handed by using a single X'-associated multiplier of $1/[1/(k+1)] + 1/[k/(k+1)] = (k+1)^2/k$.

The effect of confounding variables on sample size: The above expressions are for crude comparisons, and thus omit the VIF. An adjusted odds ratio can be obtained by a Mantel-Haenszel summary measure or by exponentiating the Woolf-weighted average of the stratum-specific log-odds ratios or by including the confounding variable(s) in a logistic regression regression model. The way in which the variance of the parameter estimate in a logistic regression model depends on the distribution of the Y, X, and C variables is quite complicated, and – aside from the special situation where the prevalence of Y = 1 instances is low [26] – not easily simplified to a usable closed form expression. Unlike the case of an identity link and Gaussian variation, it is even difficult to predict whether adding C to a logistic regression of Y on X will increase or decrease the variance of the estimated regression coefficient associated with X [27]. However, this does not mean that if C is uncorrelated with X (as it would be expected to be if, in an RCT, treatment, X, was allocated independently of C), one should omit it from the logistic regression: omitting it tends to produce an underestimate of the parameter of interest [28]. Given these complexities, the approach of [29], who (effectively) suggest using the same VIF used in Equation (3.1) above, along with the crude variance in the logit(π) scale, seems to be a sensible one, unless the effects of both X and C are very strong.

Risk/Prevalence ratios: The log of a relative risk (RR), or of the relative prevalence, is the difference between the logs of the two compared proportions. The entries in the last row of Table 2 reflect this change in scale, whereby the Bernoulli unit variance is replaced by the unit variance in the log(π) scale, namely $(1-\pi)/\pi$, to give

$$n_{\text{total}} > (Z_{\alpha/2} + Z_{\beta})^2 \times \frac{1-\pi}{\pi} \times \frac{1}{Var[X_1,..,X_n]} \times \frac{1}{\Delta^2}.$$

Of course, Δ now refers to the differerence between the non-null and null values of log(RR).

Rate ratios

For statistical purposes, one can treat a Poisson-based analysis of rate-and hazard ratios by re-expressing the overall person time as a large number (total: n) of person-moments, each with a small probability (π), indexed by X, of producing an event. The resulting unit variance in the log(π) scale then simplifies to $1/\pi$, to give

$$n_{\text{total person-moments}} > (Z_{\alpha/2} + Z_{\beta})^2 \times \frac{1}{\pi} \times \frac{1}{\text{Var}[X_1,..,X_n]} \times \frac{1}{\Delta^2}.$$

Multiplying through by π allows the inequality to be written in terms of overall number ($n\pi$) of events

$$n_{\text{total events}} > (Z_{\alpha/2} + Z_{\beta})^2 \times \frac{1}{\text{Var}[X_1,..,X_n]} \times \frac{1}{\Delta^2}.$$

where Δ is the difference between the null and alternative values of the log of the rate ratio. This is similar to the formulae given by [6,7]. However, they, as well as [30], give expressions for the numbers of events or person time in one group. Again, in order to simplify the components in the expression, we treat π as constant (and the associated variance) over the various (X,C) configurations but the simplification still gives a reasonable approximation.

As an example, consider the total number of events required for a 'rate' contrast within the same cohort study, a topic addressed in some detail in [6]. In the (unusual) situation of equal sized sub-cohorts, their quite simple formula, involving just Z_{α}, Z_{β} and the Rate Ratio, RR, yields, in the case of RR = 0.5, a total of 69 events, whereas ours yields 65. For a smaller signal, RR = 0.9, the corresponding numbers are 633 and 630. In practice, the contrasted amounts of experience are likely to be unequal, typically with less (say 0.3) in the index category and more (0.7) in the reference category. In this situation, the more unwieldy formula from [6] yields 98 whereas ours (with Var[X]= 0.21) yields 78. For a smaller signal, RR = 0.9, the corresponding numbers are 801 and 750. At these higher numbers of events, differences in the links used have less impact on the Gaussian approximation to the distribution of the test statistic: we use the log link and thus the log(Poisson count) and the log(RR) scales, whereas [6] use the fact that, conditional on the total number of events, the split into 'index' vs 'reference' category events is binomial, with $\pi = 0.3RR/(0.3RR + 0.7)$.

Correlated responses

In some studies, two or more observations come from each of several clusters, for example from paired organs or limbs or anatomical regions in the same individual, or from twin pairs, or members of the same family, school, professional practice, etc. [31-33]. Even after adjustment – through say the use of fixed-effects terms in a regression analysis – for characteristics of the individual (e.g., sex, age) and the cluster (e.g., family income, age, sex, and training of the practice professional), the residuals of the responses of individuals in the same cluster may still tend to be of the same sign. This correlation exists because not all of the shared factors that influence responses in the same cluster are measured, or measurable, or even known.

We use a simple 'one-sample' (intercept-only, constant X) example to help explain the impact of the correlation. Suppose that, in order to estimate the average level of a variable Y in a population of children, observations are made on a total of $n = 100$ children, 2 from each of 50 households. Depending on how closely correlated the levels from children in the same household, the effective sample size is somewhere between 50 – if levels from the 2 children in the same household are identical – and 100 – if they are no more alike than levels from 2 children in different households. The population mean is estimated from all n observations by giving each one a weight between 0.5 (if perfect correlation) and 1 (no correlation), and using them to take a weighted average of the n response values, Y. In such situations, the similarity is usually measured by the intra-class correlation (ICC_Y) and the weight associated with the observation on each pair-member becomes $1/(1 + ICC_Y)$. More generally, if the number of sampled children varies from household to household, the weights become 1 each for the statistically independent singletons, $1/(1+ICC_Y)$ for children in 2-children clusters, $1/(1 + 2 \times ICC_Y)$ for those from 3-children clusters, and so on [34]. Thus, if the average cluster size is k, and one wishes to achieve the same statistical precision as one would have with n independent observations, the overall size of the sample needs to be approximately $(1 + (k - 1) \times ICC_Y)$ times larger than n. One can incorporate this requirement into sample size formulae (1) by multiplying the unit variance $\mathrm{Var}[Y|X]$ for independent units by this variance inflation factor or effect size of $(1 + (k - 1) \times ICC_Y)$, i.e.,

'unit variance' if $Y_1, Y_2, ..., Y_k$ from same cluster = $\mathrm{Var}[Y|X] \times (1+(k-1) \times ICC_Y)$.

The same variance inflation factor applies in a two-sample comparison of levels in n_0 children from several families in group $X = 0$ with n_1 children, from several other families, in group $X = 1$. More generally, it applies in any multiple regression situation where observations from the same cluster are not split up across the different levels of X, the contrast-variable of interest.

The opposite of variance inflation occurs when observations from the same cluster are split up across the different levels of X. Consider as an example a comparison of Y levels across two conditions ($X = 0$ and $X = 1$), but with both conditions studied in the same cluster, *allowing for a within-cluster comparison*. The cluster might be an individual (as in a crossover design) or a twin (or otherwise-matched) pair. In such instances, the variance associated with a difference $Y_1 - Y_0$ measured within same or matched individual is *reduced* by the within-pair correlation, so that the 'unit variance' of within-cluster contrast = usual unit variance $\times (1 - r)$, where r is the correlation of the (Y_0, Y_1) pairs. The multiplier in Equation (3.1) associated with the 50:50 distribution of X remains as 4 and the sample size continues to be in terms of the *overall* or total number of *observations* (Y values); however, with *self* -pairing, the number of *subjects* is half this number. The contrast between variance inflation and variance reduction is nicely illustrated by [35].

Given that correlation can lead to a loss of information, it may seem surprising that repeated measures designs are used so commonly. However, when interest centres on a change in response under different conditions or over time, the longitudinal correlation between repeated observations means that within-person changes can be highly informative because they minimize the noise arising from between-person variability. Thus, if one wished to test a new drug purporting to increase height in middle-aged adults, the fact that height is essentially constant in this age group means that the change in height within subjects (before drug versus after) will provide a powerful test of efficacy. In such circumstances, ignoring the correlation structure can waste important information and can make standard errors too large, as when an unpaired t-test is used on paired data with a positive intraclass correlation.

Special cases

Equation (1) can easily be modified easily to accommodate other situations.

Interval estimation rather than hypothesis testing: In this context, the $(Z_{\alpha/2} + Z_\beta)^2$ term in Equation (3.1) becomes simply $Z^2_{\alpha/2}$ and Δ refers to the acceptable margin of error in estimating the parameter of interest, i.e., half the width of the confidence interval.

One sample tests/confidence intervals: Since there is no variation in, and thus no comparison across levels of, X, the multiplier involving the variance of X in Equation (3.1) simply drops out.

Ordinal covariates, *X*, and tests of trend: Although the emphasis here has been on an all-or-none covariates X, the formulae can also be used with an ordinal or interval X. One specifies, or anticipates, via its variance the approximate distribution of the X values.

Discussion

Many clinical and epidemiological studies involve sample size considerations that are much more complex that those covered in mainstream textbooks, so much so that some users resort to specialized commercial (and often expensive!) sample size software to do the calculations. However, the approach described here shows that there is considerably more unity, and thus a greater transparency, to these formulae than can be seen with the prevailing tools. In particular, by adopting a multiple regression approach, different designs involving unequal sample sizes, differences in means, regression slopes, differences of (and linear trends in) proportions, and allowance for the cost of controlling for confounding, can all be accommodated within the same framework. Generalized linear regressions with the logit or log link, and Binomial or Poisson Variation are already widely used models in data-analysis. By formulating contrasts within such a generalized linear regression framework, [36], and recognizing the change of scale – and accompanying change in variance – that these links induce, the sample size formulae can easily be extended to deal with effect measures expressed as ratios.

To be able to show the (entire) sample size formula as a product of entirely separate elements, one simplifying approximation was

required. Even when the variance is itself a function of the mean (as in Binomial and Poisson variation), the approach described here does not evaluate this variance separately under the null and under the alternative scenarios. Instead, these variances are evaluated at a common intermediate value. Quite apart from the mainly didactic purpose of this presentation, this simplification, and the other approximations involved when dealing with models other than the identity-Gaussian variation model, can be justified on several grounds. Some of these are purely statistical; some have to do with the large gap between the before-and after-the-data-collection realities and the amount of realism that can be incorporated in the formulae. Even from a purely technical statistical viewpoint, despite appearances to the contrary, all of the seemingly exact formulae are themselves approximations at best. For example, in practice, even when Y is recorded on an interval scale (as for example with birthweights), it is biologically implausible than an effective intervention will simply shift the null distribution, but not affect its spread. With binary *Y* 's, or counts, the groups studied may contain hidden mixtures that create extra-binomial or extra-Poisson variation than may be difficult to quantify at the time of planning. Moreover, it is quite difficult to anticipate what the magnitudes of the response means (or proportions), their variances – crude and net – and the multiple correlation coefficient of *X* with *C*, will actually be. Nor is it realistic to assume that one can pre-specify exactly the set of variables that will go to make up *C*, or how these variables will be represented in the final models, or how much one should adjust for the way degrees of freedom are 'spent' in arriving at final models [37].

Another reason not to overemphasize exactness in sample size formulae has to do with the fact that Δ is not what it seems. Despite it being largely a matter of judgment as to what difference would make a difference, and the costs associated with this benefit, the value used is often based on small empirical pilot studies rather than on the judgment of experts. Moreover, the before-the-study value of Δ, and the focus on detecting this difference, are – curiously – seldom considered in the interpretation of the results.

Some readers may have expected us to provide extensive comparisons of the formulae here with those in textbooks and software. The main reason we did not do so is that there is no one perfectly correct formula for any given problem. For example we regularly notice that investigators will – at the planning stage – use a sample size formula based on a specific method of data-analysis (e.g. a comparison of proportions based on the arcsine transform) and then – at the actual analysis stage – use quite a different method (e.g. a chi-square or an exact test, or focus on one coefficient in a multiple logistic regression).

The increasing use of meta-analyses correctly emphasizes the accumulation of evidence, over the single-study "yes/no" statistical decision framework that underlies most sample size formulae. In this spirit, when planning the size of an investigation, perhaps it is best to focus less on the α, β, and Δ, and more on other factors in the formulae. These are the factors that determine how much precision the investigation will 'buy' with the outlay being considered/requested, and how much – and with what efficiency – the investigation will contribute to the ultimate meta-analyses. One of our colleagues likens the issue to how much to give when the collection plate is passed around in a place of religious worship: people give according to their means: it is the aggregate of the individual contributions that ultimately matters. The scientific aggregation ultimately takes place at the time of a meta-analysis. Thus, just as there is no right individual amount to put in the collection plate, there is no right sample-size: more is usually better, unless resource constraints force sponsors to fund only the subset of studies that collectively yield the most precise overall parameter estimate for the total budget available.

Postscript

At the beginning of section 3, we noted that textbooks miss an important opportunity to re-write the term $\Sigma(X_i - \bar{X})^2$ in equation (1) in the more instructive form $n \times \text{Var}\,[X_1, ..., X_n]$, and to show the familiar $\sqrt{n}$ explicitly in its usual location in the denominator of the standard error. We appreciate that some authors prefer to define all variances in terms of a divisor of $(n-1)$. However, regression analyses treat the *n X* values as "fixed" – they may even have been designated by the investigator – it is legitimate to think of $\Sigma(X_i - \bar{X})^2$ as *n*, rather than as $(n-1)$ times the variance of these particular n values of *X*. In any case, n is typically large enough, and as most power calculations are projections based on the estimates of the magnitude of Var$[Y|X]$ – and if the *X* values are not designated ahead of time, on estimates of $\Sigma(X_i - \bar{X})^2$. Thus, the appropriateness of "approximating $(n\text{-}1)$ by *n* is moot. As statistician Karl Pearson wrote to Guinness experimentalist Willam Gosset in 1912, "only naughty brewers take *n* so small that the difference is not of the order of the probable error!".

Acknowledgements

Funding was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) and le fonds québécois de la recherche sur la nature et les technologies. We thank Piotr Biernot for research assistance.

References

1. Lehr R (1992) Sixteen s-squared over d-squared: a relation for crude sample size estimates. Statistics in Medicine 11: 1099–1102.
2. Van Belle G, Fisher LD, Heagerty PJ, Lumley TS (2004) Biostatistics: A Methodology for the Health Sciences. Wiley.
3. Armitage P, Berry G, Matthews JNS (2002) Statistical methods in medical research. Blackwell Science.
4. Fleiss JL, Levin B, Paik MC (2003) Statistical methods for rates and proportions. Wiley.
5. Schlesselman JJ (1982) Case-Control Studies: Design, Conduct, Analysis. Oxford University Press.
6. Breslow NE, Day NE (1980) Statistical methods in cancer research: The design and analysis of cohort studies. International Agency for Research on Cancer.
7. Brown CC, Green SB (1982) Additional power computations for designing comparative Poisson trials. Am J Epidemiol 115: 752–758.
8. Gail MH, Mark SD, Carroll RJ, Green SB, Pee DA (1996) On design considerations and randomization-based inference for community intervention trials. Statistics in Medicine 15: 1069–92.
9. CDC (2004) Epi info. Technical report, Atlanta.
10. Schoenfeld DA, Richter JR (1982) Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. Biometrics 38: 163–170.
11. Cohen J (1988) Statistical power analysis for the behavioral sciences. L Erlbaum Associates.
12. Dupont WD, Plummer WD (1990) Power and Sample Size Calculations: A Review and Computer Program. Controlled Clinical Trials 11: 116–128.
13. Dupont WD, Plummer WD (1998) Power and Sample Size Calculations for Studies Involving Linear Regression. Controlled Clinical Trials 19: 589–601.
14. Wacholder S (1986) Binomial regression in GLIM: estimating risk ratios and risk differences. Am J Epidemiol 123: 174–184.
15. Glantz SA, Slinker BK (2000) Primer of Applied Regression Analysis & Analysis of Variance. McGraw Hill.

16. Colton T (1974) Statistics in medicine. Little, Brown.

17. Casagrande JT, Pike MC, Smith PG (1978) The power function of the 'exact' test for comparing two binomial distributions. Applied Statistics 27: 176–180.

18. Ury HK (1975) Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. Biometrics 31: 643–649.

19. Ury HK, Fleiss JL (1980) On Approximate Sample Sizes for Comparing Two Independent Proportions with the Use of Yates' Correction. Biometrics 36: 347– 351.

20. Anderson S, Auquier A, Hauck WH, Oakes D, Vandaele W, et al. (1980) Statistical Methods for Comparative Studies: Techniques for Bias Reduction. Wiley.

21. Aickin M, Ritenbaugh C (1991) A criterion for the adequacy of a simple design when a complex model will be used for analysis. Controlled Clinical Trials, 12: 560 – 565.

22. Fleiss JL (1986) The design and analysis of clinical experiments. Wiley.

23. Hutcheon JA, Chiolero A, Hanley JA (2010) Random measurement error and regression dilution bias. BMJ 340: 1402–1406.

24. Novikov I, Fund N, Freedman LS (2010) A modified approach to estimating sample size for simple logistic regression with one continuous covariate. Statistics in Medicine 29: 97–107.

25. McFadden D(1980) Econometric models for probabilistic choice among products. The Journal of Business 53: 13–29.

26. Whittemore A (1981) Sample Size for logistic regression with small response probability. J Amer Statist Assoc 76: 27–32.

27. Robinson LD, Jewell NP (1991) Some Surprising Results About Covariate Adjustment in Logistic Regression Models. International Statistical Review 58: 227– 240.

28. Gail MH, Wieand S, Piantadosi S (1984) Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika 71: 431-444.

29. Smith PG, Day NE (1984) The design of case-control studies: the influence of confounding and interaction effects. Int J Epidemiol 13: 356–365.

30. Methods for field trials of interventions against tropical diseases: a toolbox (1991) Oxford University Press.

31. Rosner B, Milton RC (1988) Significance testing for correlated binary outcome data. Biometrics 44: 505–512.

32. Donner A, Klar N (1994) Methods for Comparing Event Rates in Intervention Studies When the Unit of Allocation is a Cluster. Am J Epidemiol 140: 279– 289.

33. Donner A, Klar N (1996) Statistical considerations in the design and analysis of community intervention trials. J Clin Epidemiol 49: 435–439.

34. Hanley JA, Negassa A, Edwardes MD deB, Forrester JE(2003) Statistical Analysis of Correlated Data Using Generalized Estimating Equations: An Orientation. Am J Epidemiol 157: 364–375

35. Burton P, Gurrin L, Sly P (1998) Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. Statistics in Medicine 17: 1261-1291.

36. Tosteson TD, Buzas JS, Demidenko E, Karagas M (2003) Power and sample size calculations for generalized regression models with covariate measurement error. Statistics in Medicine 22: 1069–1082.

37. Harrell FE (2001) Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer.