rbn1	Generalized power calculations

Author: Roger Newson, Imperial College London, UK. Email: r.newson@imperial.ac.uk Date: 26 March 2012.

Abstract

The program `powercal` carries out generalized power and sample size calculations and outputs the results to variables in a Stata data set. These variables can then be plotted and listed. This usually communicates the message of power and sample size calculations, and the tradeoffs involved, better than a point result as provided by `sampsi`.

Key phrases

Power; sample size; significance level; detectable difference; standard error; influence function.

Syntax

`powercal` *newvarname* [`if` *exp*][`in` *range*], [`nunit(`*expression_1*`)` `power(`*expression_2*`)` `alpha(`*expression_3*`)` `delta(`*expression_4*`)` `sdinf(`*expression_5*`)` `tdf(`*expression_6*`)` `noceiling` `float`]

Description

`powercal` performs generalized power calculations, storing the result in a new variable with a name specified by *newvarname*. All except one of the expression options `nunit()`, `power()`, `alpha()`, `delta()` and `sdinf()` must be specified. The single unspecified option in this list specifies whether the output variable is the number of sampling units, power, alpha (significance level), delta (difference in parameter value to be detected), or the standard deviation (SD) of the influence function. Any of these 5 quantities can be calculated from the other 4. `powercal` can be used to calculate any of these quantities, assuming that we are testing a hypothesis that a parameter is zero, and that the true value is given by `delta()`, and that the sample statistic is distributed around the population parameter in such a way that the pivotal quantity

$$\texttt{PQ} = \texttt{sqrt}(\texttt{nunits}) * (\texttt{delta}/\texttt{sdinf})$$

has a standard Normal distribution (if `tdf()` is not specified) or a t-distribution with `tdf()` degrees of freedom (if `tdf()` is specified). The formulas used by `powercal` define power as the probability of detecting a difference in the right direction, using a two-tailed test.

Options

`nunit(`*expression_1*`)` gives an expression whose value is the number of independent sampling units. Sampling units are defined very generally. For instance, in an experiment involving equal-sized samples of individuals from Population A and Population B, a sampling unit might be a pair of sampled individuals, one from each population. Similarly, in a case-control study with 4 controls per case, a sampling unit might be a case together with 4 controls.

`power(`*expression_2*`)` gives an expression whose value is the power to detect a difference specified by the `delta()` option (see below). The power is defined as the probability that the sample difference is in the correct direction, and also large enough to be significant, using a 2-tailed test, at the level specified by the `alpha()` option (see below).

`alpha(`*expression_3*`)` gives an expression whose value is the size, or significance level, of the statistical test (in units of probability, not percentage).

`delta(`*expression_4*`)` gives an expression whose value is the true population difference to be detected. This difference is assumed to be positive. Therefore, if the user wishes to detect a negative difference, then s/he should specify an expression equal to minus that difference. The difference may be the log of a ratio parameter, such as an odds ratio, rate ratio, risk ratio or ratio of geometric means.

`sdinf(`*expression_5*`)` gives an expression whose value is the standard deviation of the influence function. That is to say, it is an expression equal to the expected standard error of the sample difference multiplied by the square root of the number of sampling units, where sampling units are defined generally, as specified in the option `nunit()`. In the simple case of a paired t-test, `sdinf()` is the standard deviation of the paired differences. More generally, `sdinf()` can be defined by calculating a standard error for a particular number of units, from a pilot study, from a simulation or from a formula, and multiplying this standard error by the square root of the number of units in the pilot study, simulation or formula.

`tdf(`*expression_6*`)` gives an expression whose value is the degrees of freedom of the t-distribution to be assumed for the pivotal quantity `PQ` specified above. The degrees of freedom expression is not necessarily integer-valued. If `tdf()` is absent, then `PQ` is assumed to follow a standard Normal distribution.

`noceiling` specifies that, if the output variable specified by *newvarname* is a number of units, then it will not be rounded up to the lowest integer no less than itself (as calculated by the Stata `ceil()` function). This option can be useful if the output variable is intended to specify an amount of exposure, such as a number of person-years, and the input `sdinf()` expression specifies a standard deviation of the influence function per unit exposure. If `noceiling` is not specified, and `power()`, `alpha()`, `delta()` and `sdinf()` are specified, then `powercal` rounds up the output variable, so that it contains a whole number of units

`float` specifies that the output variable will have a storage type no higher than `float`. If `float` is not specified, then `powercal` creates the output variable with storage type `double`. Whether or not `float` is specified, `powercal` compresses the output variable as much as possible without loss of precision. (See help for `compress`.)

Remarks

`powercal` carries out sample size calculations for a more general range of possible experimental designs than `sampsi`, and stores the result in a new variable, instead of reporting the result in the log. The new variable may be input to further calculations and/or plotted and/or listed. `powercal` is intended as a low-level programming tool for users intending to carry out sample size calculations for a given experimental design. It is the responsibility of the user to ensure that the expressions are correct, and to choose a parameter scale on which the parameter is expected to be Normally distributed (or t-distributed), with a variance that does not vary excessively with the size of the measured difference.

Methods and Formulas

Generalized power and sample size calculation formulas are based on the Central Limit Theorem applied to influence functions. Suppose that a sequence of (scalar or vector) random variables $\{X_i\}$ is sampled independently from a common (univariate or multivariate) population distribution, and suppose that $\theta(F)$ is a (scalar or vector) parameter, defined from the set of candidate (univariate or multivariate) cumulative distribution functions F that might apply to the X_i. Denote by $\hat{F}_n$ the sample cumulative distribution function, based on the first n of the X_i, and define $\hat{\theta}_n = \theta(\hat{F}_n)$ to be a sample estimator of $\theta(F_0)$, where F_0 is the true population cumulative distribution function of the X_i. An influence function $\Upsilon(X;\theta,F)$ is a function, defined for each possible X-value, parameter value and cumulative distribution function, and having the properties that

$$E\left[\Upsilon\left(X_i;\theta(F_0),F_0\right)\right] = 0 \tag{1}$$

and

$$\hat{\theta}_n = \theta(F_0) + n^{-1}\sum_{i=1}^{n}\Upsilon\left(X_i;\theta(F_0),F_0\right) + o_p(n^{-1/2}), \tag{2}$$

where $E[\cdot]$ denotes expectation, and $o_p(n^{-1/2})$ is a term having the property that $o_p(n^{-1/2})/n^{-1/2}$ converges in probability to zero. Therefore, in words, the sample statistic is equal to the population parameter, plus the sample mean of the population influence function, plus a third term, which is negligible if the sample size is sufficiently large. (In the simplest case, where the X_i are scalar random variables, θ is their population mean and $\hat{\theta}_n$ is the sample mean for the first n of the X_i, the influence function is $\Upsilon(X_i;\theta,F) = X_i - \theta$.)

Influence functions with properties (1) and (2) exist for a wide range of parameters, including those estimated by maximum likelihood (whether or not the likelihood function is correctly specified). They are the reason why the Central Limit Theorem can be generalized from sample means to more general sample statistics. More details about the theory, and more rigorous definitions of influence functions, can be found in Hampel (1974), Hampel *et al.* (1986) and Huber (1981). However, for power calculation purposes, the main consequences of properties (1) and (2) are that, for a wide range of parameter estimates $\hat{\theta}_n$, the quantity

$$Z_n = \frac{n^{1/2}}{\sigma}\left[\hat{\theta}_n - \theta(F_0)\right] = \left[\hat{\theta}_n - \theta(F_0)\right]/\mathrm{SE}(\hat{\theta}_n) \tag{3}$$

has an asymptotic standard Normal distribution, where

$$\sigma = E\left[\Upsilon\left(X_i;\theta(F_0),F_0\right)^2\right]^{1/2} \tag{4}$$

is the population standard deviation of the population influence function, and

$$\mathrm{SE}(\hat{\theta}_n) = \sigma/\sqrt{n} \tag{5}$$

is known as the asymptotic standard error. In the simplest case of estimating the population mean of scalar X_i by the sample mean, σ is simply the population standard deviation of the X_i. However, in the more general case, if we have a formula or estimate for $\mathrm{SE}(\hat{\theta}_n)$ for known n, then we can multiply that formula or estimate by $\sqrt{n}$ to derive a formula or estimate for σ.

In practice, when calculating confidence intervals and P-values, we usually estimate σ with a consistent estimator $\hat{\sigma}_n$, calculated from the first n of the X_i, and calculate an estimated standard error $\widehat{\mathrm{SE}}(\hat{\theta}_n) = \hat{\sigma}_n/\sqrt{n}$, and then the quantity

$$\hat{Z}_n \;=\; \frac{n^{1/2}}{\hat{\sigma}_n}\left[\hat{\theta}_n - \theta(F_0)\right] \;=\; \left[\hat{\theta}_n - \theta(F_0)\right] / \widehat{\mathrm{SE}}(\hat{\theta}_n) \tag{6}$$

is a consistent estimator of Z_n and has an asymptotic standard Normal distribution. Sometimes, the distribution of $\hat{Z}_n$ for finite n can be approximated better by a t-distribution with finite degrees of freedom, which may or may not be integer.

Most power and sample size calculations aim to calculate power and sample size to detect a non-zero value for a population difference parameter δ, estimated by a sample difference statistic $\hat{\delta}$, by showing that the confidence limits for the population δ exclude zero. (Note that a difference may be a log ratio or other difference between parameter values transformed by a Normalizing and/or variance-stabilizing transformation.) In the following formulas, we will assume that a significance threshold α is used to define $100(1-\alpha)\%$ confidence intervals, or to reject the null hypothesis $\delta = 0$ with $P \leq \alpha$. If the number of sampling units is n and the standard deviation of the influence function is σ, then the standard error of $\hat{\delta}$ is $\mathrm{SE}(\hat{\delta}) = \sigma/\sqrt{n}$, and the pivotal quantity

$$Z = (\hat{\delta} - \delta)/\mathrm{SE}(\hat{\delta}) \;=\; n^{1/2}(\hat{\delta} - \delta)/\sigma \tag{7}$$

is assumed to be distributed with a cumulative density function $G(\cdot)$ such that, for any z,

$$G(z) \;=\; \Pr(Z \leq z) \;=\; \Pr(Z < z) \;=\; 1 - G(-z). \tag{8}$$

The first equality is a definition, the second equality specifies a continuous distribution, and the third equality specifies that the distribution is symmetrical around zero. These conditions hold whether $G(\cdot)$ specifies a standard Normal distribution or a central t-distribution. If $G^{-1}(\cdot)$ is the inverse of $G(\cdot)$, then a $100(1-\alpha)\%$ confidence interval is defined (approximately) by $\hat{\theta} \pm G^{-1}(1-\alpha/2) \times \mathrm{SE}(\hat{\theta})$, and the null hypothesis $\delta = 0$ is rejected in a positive direction by a two-tailed test at $P \leq \alpha$ if and only if $Z \geq G^{-1}(1-\alpha/2)$. (We are assuming that, if the standard error is estimated, then it is estimated well, so that the $\hat{Z}_n$ of (6) is a good approximation to the Z_n of (3).) If the power to detect a positive difference δ is no less than a required level γ, then it follows that

$$\begin{aligned}
\gamma \;&\leq\; \Pr\left[\hat{\delta}/\mathrm{SE}(\hat{\delta}) \;\geq\; G^{-1}(1-\alpha/2)\right] \\
&=\; \Pr\left[\hat{\delta}/\mathrm{SE}(\hat{\delta}) - \delta/\mathrm{SE}(\hat{\delta}) \;\geq\; G^{-1}(1-\alpha/2) - \delta/\mathrm{SE}(\hat{\delta})\right] \\
&=\; 1 - G\left[G^{-1}(1-\alpha/2) \;-\; \delta/\mathrm{SE}(\hat{\delta})\right] \\
&=\; G\left[\delta/\mathrm{SE}(\hat{\delta}) \;-\; G^{-1}(1-\alpha/2)\right].
\end{aligned} \tag{9}$$

The first inequality is a requirement, the first equality follows trivially, the second equality follows from the fact that $G(\cdot)$ specifies a continuous distribution for Z, and the third equality follows from the symmetry of that distribution around zero. Applying $G^{-1}(\cdot)$ to both sides of the inequality (9), we have

$$G^{-1}(\gamma) \;\leq\; \delta/\mathrm{SE}(\hat{\delta}) \;- G^{-1}(1-\alpha/2), \tag{10}$$

or, equivalently,

$$\frac{\delta\sqrt{n}}{\sigma} \;\geq\; G^{-1}(\gamma) \;+\; G^{-1}(1-\alpha/2). \tag{11}$$

The inequality (11) expresses the power requirements elegantly and briefly, as the left hand side is increasing in δ and n and decreasing in σ, and the right hand side is the sum of two terms, the first increasing in γ and the second decreasing in α. We can therefore rearrange (11) to derive a minimum or maximum value for each of the 5 parameters γ, α, δ, σ and n, compatible with the power requirements (9) and with given values of the other 4

parameters. These minima or maximas may or may not exist for γ and α in the interval $(0, 1)$ and positive δ, σ and n, because the inequality (11) may be satisfied nowhere or everywhere in the open interval parameter range. If we define the quantities

$$R = G^{-1}(\gamma) + G^{-1}(1-\alpha/2) \text{ and } S = \delta\sqrt{n}/\sigma - G^{-1}(\gamma), \tag{12}$$

then the minima and maxima are defined as follows:

$$\begin{aligned}
\gamma_{\max} &= G\left[\,\delta\sqrt{n}/\sigma - G^{-1}(1-\alpha/2)\,\right], \\
\alpha_{\min} &= 2G(-S), \text{ if } S > 0, \\
\delta_{\min} &= \frac{\sigma}{\sqrt{n}}R, \text{ if } R > 0, \\
\sigma_{\max} &= \delta\sqrt{n}/R, \text{ if } R > 0, \\
n_{\min} &= \left\lceil \left(\frac{\sigma}{\delta}R\right)^2 \right\rceil, \text{ if } R > 0.
\end{aligned} \tag{13}$$

The operator $\lceil x \rceil$ represents the minimum integer no less than x, as calculated by the `ceil()` function in Stata. This operator is not applied if the user specifies the `noceiling` option. The inequality (11) is not satisfied by any $\alpha \in (0, 1)$ if $S \leq 0$, and is satisfied by all positive δ, σ and n if $R \leq 0$. Note that $R \leq 0$ can only be true if $\gamma \leq 1/2$, and that $S \leq 0$ can only be true if δ represents fewer standard errors than $G^{-1}(\gamma)$. In practice, we usually aim for more than 50% power to detect an interesting positive population difference, and we usually choose a sample size large enough to make the standard error small enough to prevent the *sample* difference from being negative even when the *population* difference is positive enough to be interesting.

The parameter σ must usually be provided by the user. It may be estimated by multiplying a standard error from a pilot study, a simulation or a formula by the square root of the number of units involved in calculating that standard error. In the absence of a pilot study or a simulation, a formula is usually known only for the simplest cases. For instance, in the case of a paired t-test, or a sign test, the standard deviation of the influence function is simply the standard deviation of the pairwise differences, or of the signs of these differences, respectively.

However, many experimental designs involve sampling in parallel and independently from K subpopulations of primary sampling units (PSUs), estimating a population parameter η_j for the jth subpopulation by means of a sample estimate $\hat{\eta}_j$, and thereby estimating a contrast of interest $\delta = \sum_{j=1}^{K} a_j \eta_j$, assumed to be zero under a null hypothesis to be tested. (Usually, but not always, the η_j are link functions of subpopulation means in a generalized linear model, as defined by McCullagh and Nelder, 1989. Examples include arithmetic subpopulation means, log geometric subpopulation means, log subpopulation incidence rates, or log case and control odds of exposure in an unmatched case-control study.) A sample for such a design may contain a number n of compound sampling units (CSUs), where each CSU consists of m_j PSUs sampled independently from each jth subpopulation. (For instance, an unmatched case-control study may have a fixed number of controls per case, and then $K = 2$, $a_1 = 1$, $a_2 = -1$, $m_1 = 1$, and m_2 is the number of controls per case.) Sample size calculations for such designs usually output or input numbers of CSUs, rather than numbers of PSUs. The estimate for δ is $\hat{\delta} = \sum_{j=1}^{K} a_j \hat{\eta}_j$. The standard error of $\hat{\eta}_j$ is

$$\mathrm{SE}\left(\hat{\eta}_j\right) = \sigma_j / \sqrt{n m_j}, \tag{14}$$

where σ_j is the standard deviation of the influence function (per PSU) of η_j. If η_j is a link function in a generalized linear model, and there is one observation per PSU, then the standard deviation of the per-PSU influence function is equal to

$$\sigma_j = \frac{d\eta_j}{d\mu_j}\sqrt{\phi V(\mu_j)}, \tag{15}$$

where, in the notation of McCullagh and Nelder (1989), μ_j is the subpopulation mean corresponding to η_j, $V(\mu_j)$ is the variance function, and ϕ is the dispersion parameter. The standard error of $\hat{\delta}$ is

$$\mathrm{SE}\left(\hat{\delta}\right) = \sqrt{\sum_{j=1}^{K} a_j^2 \left[\mathrm{SE}\left(\hat{\eta}_j\right)\right]^2}. \tag{16}$$

It follows that the standard deviation of the per-CSU influence function of δ is derived from the standard errors of the per-PSU influence functions of the η_j by the formula

$$\sigma = \sqrt{n} \times \mathrm{SE}\left(\hat{\delta}\right) = \sqrt{\sum_{j=1}^{K} \frac{a_j^2}{m_j}\sigma_j^2}. \tag{17}$$

Note that influence functions are "additive" because the PSUs of different subpopulations are independent.

Example 1. Geometric mean ratios

The geometric mean (defined as the antilogarithm of the arithmetic mean logarithm) is frequently used as an approximation to the median if a variable is positive-valued and positively skewed. Power calculations for ratios between geometric means usually assume that the outcome variable has a lognormal distribution, so that the log of the outcome variable has a Normal distribution. Under this assumption, the geometric mean is the median, its log is the mean log, and the standard deviation of the logs is the other parameter of the distribution, measuring dispersion. Alternative measures of dispersion for positive-valued variables, more familiar to non-mathematicians, are the coefficient of variation and the qth tail ratio, defined as the ratio of the $100(1-q)$th percentile to the $100q$th percentile if $0 < q < 1/2$. If the lognormal assumption is true, then the standard deviation of the natural logs can be calculated from the coefficient of variation or the qth tail ratio by the formulas

$$\mathrm{SD}_{\log} = \sqrt{\ln\left(\mathrm{CV}^2 + 1\right)} = -\ln(r_q) / \left[2\Phi^{-1}(q)\right], \tag{18}$$

where $\mathrm{SD}_{\log}$ is the standard deviation of the natural logs, CV is the coefficient of variation of the unlogged variable, r_q is the qth tail ratio of the unlogged variable, and $\Phi^{-1}(\cdot)$ is the inverse standard normal cumulative distribution function. (See Aitchison and Brown, 1963, or Stanislav Kolenikov's website at *http://www.komkon.org/~tacik/*, which contains some formulas from that source for quick reference.)

When we perform lognormal power calculations, the difference δ that we aim to detect is usually a linear contrast between logs of geometric means. In the notation of (14), (15), (16) and (17), the η_j are logs of geometric means, the σ_j are standard deviations of the logs, and we wish to know the standard deviation σ of the influence function of the contrast δ, so that we can apply the formulas (13). In the simplest case, we may plan to measure the ratio between geometric means in 2 treatment groups. In this case, we have $K = 2$, η_1 and η_2 are the log geometric means in treatment groups 1 and 2 respectively, $a_1 = 1$, $a_2 = -1$, and the difference to detect is the log geometric mean ratio $\delta = \eta_1 - \eta_2$. The PSUs are treated units. If we decide to apply the treatments to 2 unmatched samples of equal size, then each CSU might be a pair of PSUs, one allocated to each treatment group, and therefore we have $m_1 = m_2 = 1$. If we assume that the two treatment groups have a common coefficient of variation (and therefore common tail ratios), then we also have $\sigma_1 = \sigma_2 = \mathrm{SD}_{\log}$, where $\mathrm{SD}_{\log}$ is derived from the assumed coefficient of variation or tail ratio by (18). By (17), the standard deviation of the per-CSU influence function of δ is then given by

$$\sigma = \mathrm{SD}_{\log} \times \sqrt{2}. \tag{19}$$

Note that (19) is derived from a special case of (15), where the logs are distributed according to a generalized linear model with an identity link function and a Normal variance function. In this case, $\eta_j = \mu_j$ is the arithmetic mean log (or log geometric mean) of the jth treatment group, $d\eta_j/d\mu_j = V(\mu_j) = 1$, and $\phi = {\mathrm{SD}_{\log}}^2$ is the common variance of the logs in both treatment groups.

The following example assumes a coefficient of variation of 0.5 within each of 2 treatment groups. This implies a 20% tail ratio of 2.2147318, meaning that, within each treatment group, the bottom of the top quintile is 2.2147318 times the top of the bottom quintile. The variable `logratio` is created, containing a range of log geometric mean ratios, and the variable `gmratio` is created, containing the corresponding ratios themselves, which range from 1 to 2. We then use `powercal` to calculate, in a new variable `power`, the power to detect each geometric mean ratio with $P \leq 0.01$, using 50 units in each group (and therefore 50 CSUs) and carrying out a two-sample t-test on the logs. The power is plotted against the geometric mean ratio in Figure 1, with vertical-axis reference lines for 80% and 90% power. We see that a geometric mean ratio of 1.39 can be detected with 80% power, whereas a geometric mean ratio as high as 1.45 can be detected with 90% power.

```
. clear;
. scal cv=0.5;
. scal sdlog=sqrt(log(cv*cv + 1));
. scal r20=exp(-2*sdlog*invnorm(0.2));
. disp _n as text "Coefficient of variation: " as result cv
>   _n as text "SD of logs: " as result sdlog
>   _n as text "20% tail ratio: " as result r20;
Coefficient of variation: .5
SD of logs: .47238073
20% tail ratio: 2.2147318
. set obs 100;
obs was 0, now 100
. gene npergp=_n;
. lab var npergp "Number per group";
. powercal logratio, power(0.9) alpha(0.01) sdinf(sdlog*sqrt(2)) nunit(npergp)
```

```
>   tdf(2*(npergp-1));
Result to be calculated is delta in variable: logratio
. gene hiratio=exp(logratio);
(1 missing value generated)
. gene loratio=exp(-logratio);
(1 missing value generated)
. lab var hiratio "Detectable GM ratio >1";
. lab var loratio "Detectable GM ratio <1";
. line hiratio loratio npergp if _n>=5, xlab(0(10)100);
```

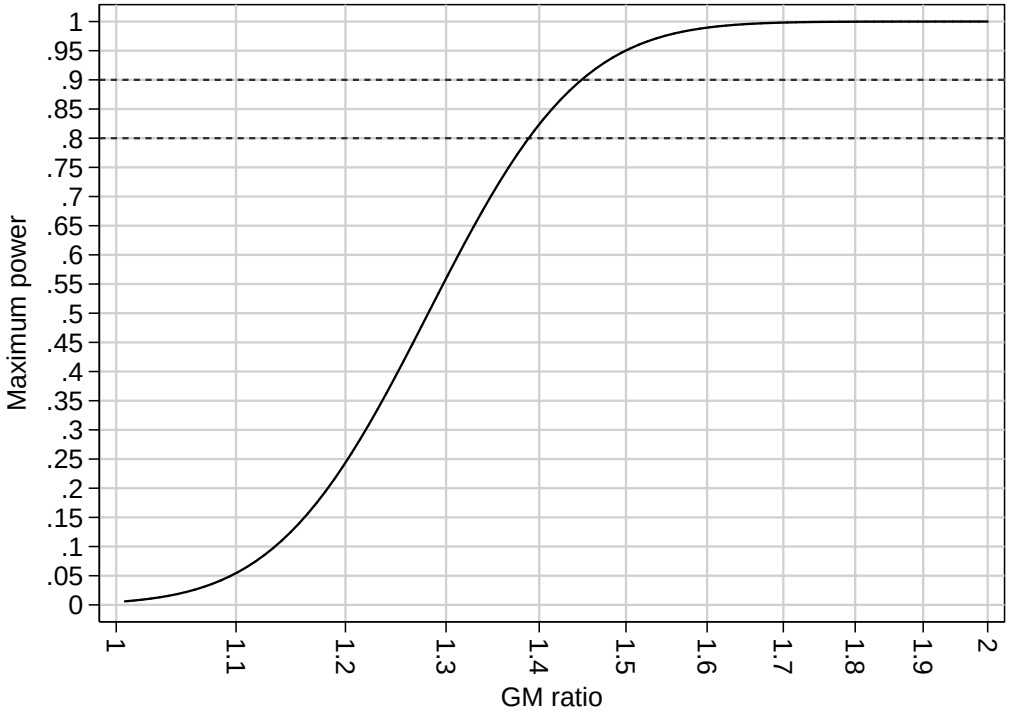

Figure 1. Power to detect geometric mean ratios

Alternatively, we might wish to calculate the detectable geometric mean ratios closest to unity as a function of sample number. The following example does this by creating a variable `npergp`, containing possible numbers per group from 1 to 100, and then using `powercal` to calculate the detectable positive log geometric mean ratio as a function of `npergp`, assuming that we require 90% power to detect a difference with $P \leq 0.01$ by t-testing the logs, and that the coefficient of variation within each treatment group is 0.5 as before. We then calculate the detectable geometric mean ratios greater than 1 and less than 1 as `hiratio` and `loratio`, respectively, and plot these against the number per treatment group, with a vertical-axis reference line indicating a ratio of 1. This plot is Figure 2. Note that we have suppressed the spectacular ratios detectable with 4 or fewer subjects per group. A plot such as Figure 2 has the advantage that it communicates to colleagues the inverse square law, whereby, to halve the detectable difference, you must approximately *quadruple* (not double) the number of subjects. Non-statisticians frequently do not appreciate this law, although they usually are vaguely aware that larger sample sizes increase power.

```
. clear;
. scal cv=0.5;
. scal sdlog=sqrt(log(cv*cv + 1));
. scal r20=exp(-2*sdlog*invnorm(0.2));
. disp _n as text "Coefficient of variation: " as result cv
>   _n as text "SD of logs: " as result sdlog
>   _n as text "20% tail ratio: " as result r20;
Coefficient of variation: .5
SD of logs: .47238073
20% tail ratio: 2.2147318
. set obs 100;
obs was 0, now 100
. gene npergp=_n;
. lab var npergp "Number per group";
. powercal logratio, power(0.9) alpha(0.01) sdinf(sdlog*sqrt(2)) nunit(npergp)
```

```
>   tdf(2*(npergp-1));
Result to be calculated is delta in variable: logratio
. gene hiratio=exp(logratio);
(1 missing value generated)
. gene loratio=exp(-logratio);
(1 missing value generated)
. lab var hiratio "Detectable GM ratio >1";
. lab var loratio "Detectable GM ratio <1";
. line hiratio loratio npergp if _n>=5, xlab(0(10)100);
```

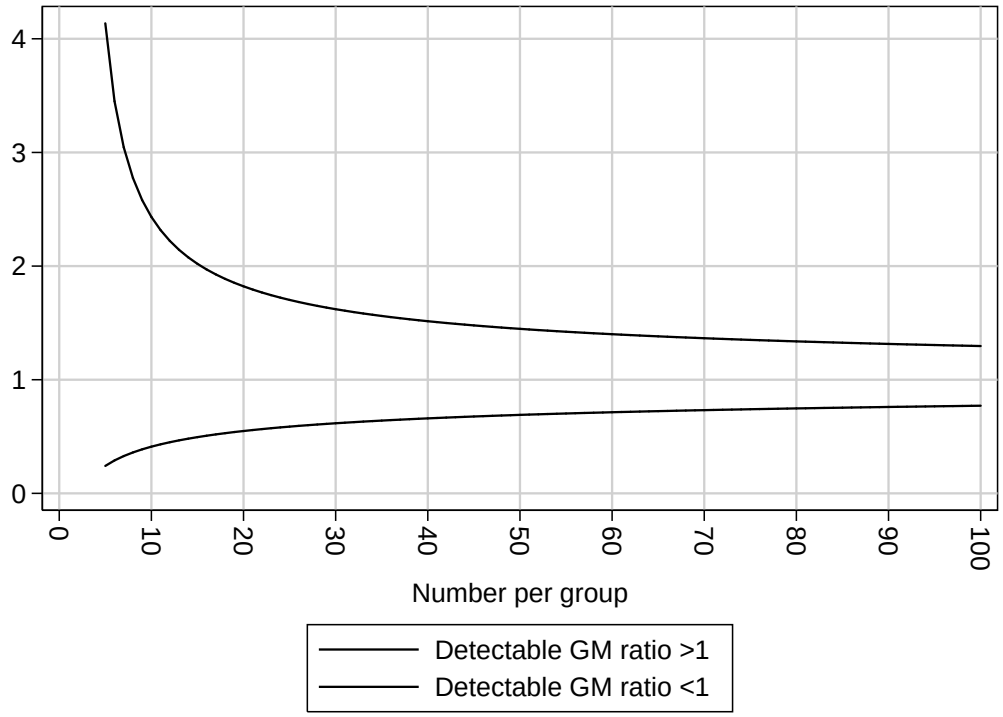

Figure 2. Detectable geometric mean ratios as a function of number per treatment group

Example 2. Odds ratios from case-control studies

Case-control studies are commonly recommended as the design of choice in genomic epidemiology for measuring an association between a gene and a disease (see Clayton and McKeigue, 2001). If we are designing an unmatched case-control study, then we typically plan to sample a given number of subjects with each possible disease status (eg "with the disease" and "without the disease"), and then measure, in each subject, the exposure, which might be the presence of a gene in a patient. The difference δ that we wish to detect is the log ratio of the odds of the exposure between cases and controls, or possibly some other linear contrast of the log odds of exposure for different disease status values, if there are more than 2 possible disease status values. If there are K possible values of disease status, and E_j is the prevalence of exposure in subjects with the jth disease status, then the odds of exposure in the jth disease category is defined as $E_j/(1-E_j)$, and its logarithm is typically used as a normalizing and variance-stabilizing transformation.

In the generalized linear model notation of (14), (15), (16) and (17), the PSUs are subjects (cases or controls), the subpopulations correspond to the possible disease status values, and the "outcome" variable is a binary exposure variable, whose distribution in each subpopulation is governed by a generalized linear model with a logit link function and a Bernoulli variance function. The mean "outcome" in the jth subpopulation is therefore $\mu_j = E_j$, the parameter estimated for the jth subpopulation is $\eta_j = \ln[\mu_j/(1-\mu_j)]$, its derivative is $d\eta_j/d\mu_j = 1/\mu_j + 1/(1-\mu_j)$, the variance function is $V(\mu_j) = \mu_j(1-\mu_j)$, and the dispersion parameter is $\phi = 1$. It follows from (15) that the standard deviation of the per-PSU influence function of the log odds η_j is

$$\sigma_j \;=\; \sqrt{1/E_j + 1/(1-E_j)}. \tag{20}$$

A CSU in this case is composed of m_j subjects sampled independently from the subpopulation with each jth disease status. This is because, although the case-control study is unmatched, we may plan to sample subjects of different

disease status according to a particular ratio, such as two controls per case. The standard deviation of the per-CSU influence function is then given by the formula (17). Note that, in the sample size calculations, the generalized linear model is defined with the disease status as the "predictor" and the exposure status as the "outcome". This is in contrast to the statistical analysis, where the disease status is usually the "outcome" and the exposure status is usually the "predictor".

In the simplest case, there are $K = 2$ possible values for disease status, namely "diseased" and "undiseased", and a CSU is a single case together with m_2 unmatched controls, so that $m_1 = 1$. We are interested in measuring a log odds ratio $\delta = \eta_1 - \eta_2$, so, in the notation of (17), we have $a_1 = 1$ and $a_2 = -1$. In this case, the standard deviation of the per-CSU influence function is given, according to (17), by

$$\sigma = \sqrt{\frac{1}{E_1} + \frac{1}{1-E_1} + \frac{1}{m_2}\left[\frac{1}{E_2} + \frac{1}{1-E_2}\right]}. \tag{21}$$

When designing a case-control study, we typically have a good prior estimate of the control exposure prevalence E_2, because the control exposure prevalence is intended to be an estimate for the total population exposure prevalence. Therefore, if we hypothesize a particular value for the odds ratio, we can multiply this odds ratio by the "known" control odds of exposure to arrive at the corresponding hypothesized case odds of exposure by the formula

$$E_1/(1-E_1) \;=\; \mathrm{OR} \times E_2/(1-E_2), \tag{22}$$

and then calculate the case exposure prevalence E_1 from the case exposure odds $E_1/(1-E_1)$. Given an estimate for the control exposure E_2, the standard deviation of the per-case influence function of the odds ratio, given by (21), is dependent on the odds ratio itself, and therefore should not be entered into the formulas (13) independently of the odds ratio itself. This is in contrast to the case with lognormal geometric mean ratios.

The following example assumes that we are planning a case-control study to measure the association of a rare disease with a binary exposure, whose control prevalence is expected to be 0.25, or 25%. We decide to recruit $m_2 = 2$ unmatched controls per case. We create a data set with 1 observation for each of a range of odds ratios from 1.25 to 5, which will correspond to relative risks of the same size, if the rare disease assumption is true. The log odds ratios are stored in the variable `logor`, the odds ratios are stored in `or`, the case exposure odds are stored in `caseodds`, the case exposure prevalences are stored in `caseprev`, and the control exposure prevalence and odds are stored in scalars. We use the formulas (20) and (21) to calculate the standard deviation of the influence function of the log odds ratio in `sdinflor`. We then use `powercal` to calculate the minimum number of cases to detect each odds ratio at significance level $P \leq 0.01$ with 90% power, and plot the odds ratio against that minimum number of cases, suppressing odds ratios requiring over 2000 cases to be detectable. The resulting graph is Figure 3. Note that uninteresting low unadjusted odds ratios are very expensive to detect, as well as being more credibly attributed to confounding than spectacular high odds ratios.

```
. clear;
. scal conprev=0.25;
. scal conodds=conprev/(1-conprev);
. disp _n as text "Expected control prevalence: " as result conprev
>   _n as text "Expected control odds: " as result conodds;
Expected control prevalence: .25
Expected control odds: .33333333
. set obs 101;
obs was 0, now 101
. gene logor=log(1.25)+(log(5)-log(1.25))*(_n-1)/(_N-1);
. gene or=exp(logor);
. gene caseodds=conodds*or;
. gene caseprev=caseodds/(1+caseodds);
. gene sdinflor=sqrt(
>   1/caseprev + 1/(1-caseprev) + (1/2)*( 1/conprev + 1/(1-conprev) )
>   );
. lab var logor "Log odds ratio";
. lab var or "Odds ratio";
. lab var caseodds "Case exposure odds";
. lab var caseprev "Case exposure prevalence";
. lab var sdinflor "SD of influence for log OR";
. desc;
Contains data
  obs:           101
 vars:             5
```

```
 size:         2,020
------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
logor           float  %9.0g                 Log odds ratio
or              float  %9.0g                 Odds ratio
caseodds        float  %9.0g                 Case exposure odds
caseprev        float  %9.0g                 Case exposure prevalence
sdinflor        float  %9.0g                 SD of influence for log OR
------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved
. * Detectable OR by number of cases *;
. powercal ncases, power(0.9) alpha(0.01) delta(logor) sdinf(sdinflor);
Result to be calculated is nunit in variable: ncases
. line or ncases if ncases<=2000, yscale(log);
```

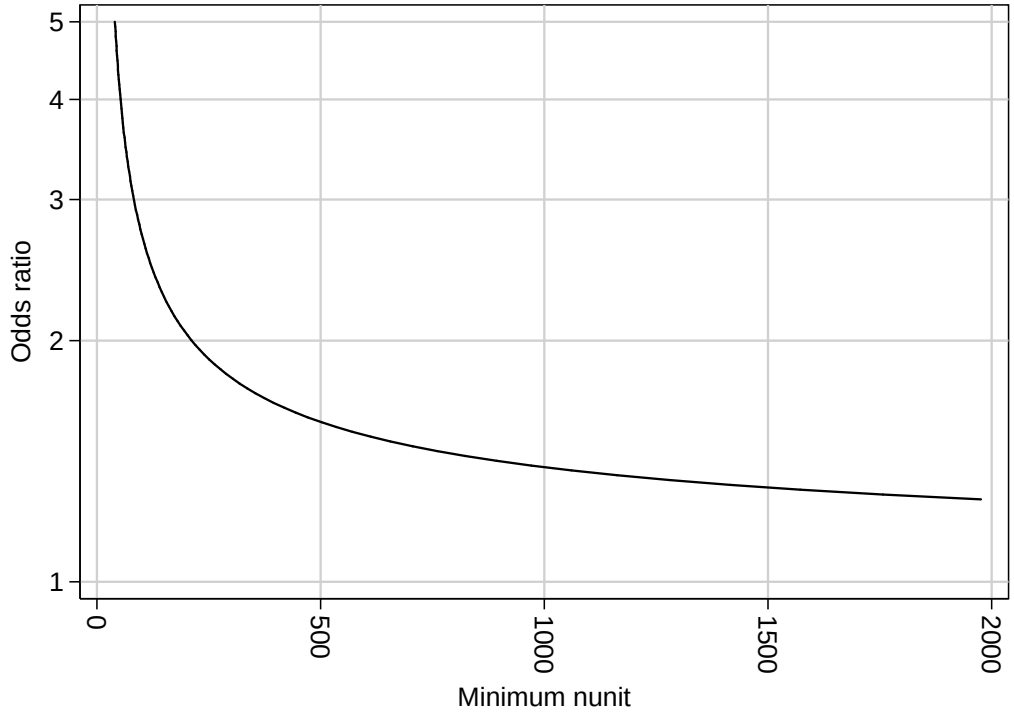

Figure 3. Detectable odds ratios as a function of number of cases

Using the same data set, we can also calculate, for each odds ratio, the significance level attainable with 90% power, using 100 cases and their 200 controls. This is done as in the example below, creating Figure 4. Note that the significance level is plotted on a reverse log scale, so that, the higher a point on the curve is, the more convincing is the significance level that we can expect. Odds ratios between 2 and 3 are likely to be "significant" at the conventional 5% and 1% levels. However, higher odds ratios are more likely to attain significance levels that might convince the skeptics, in view of the problems of multiple comparisons and publication bias. (See Colhoun *et al.*, 2003, for a discussion of these problems in genomic epidemiology.)

```
. * Significance level by odds ratio *;
. powercal alphamin, power(0.9) delta(logor) sdinf(sdinflor) nunit(100);
Result to be calculated is alpha in variable: alphamin
. line alphamin or,
>   yscale(log reverse) ylab(1 0.05 1e-1 1e-2 1e-3 1e-4 1e-5 1e-6 1e-7)
>   xscale(log) xlab(1 1.25 1.5 2(1)5);
```

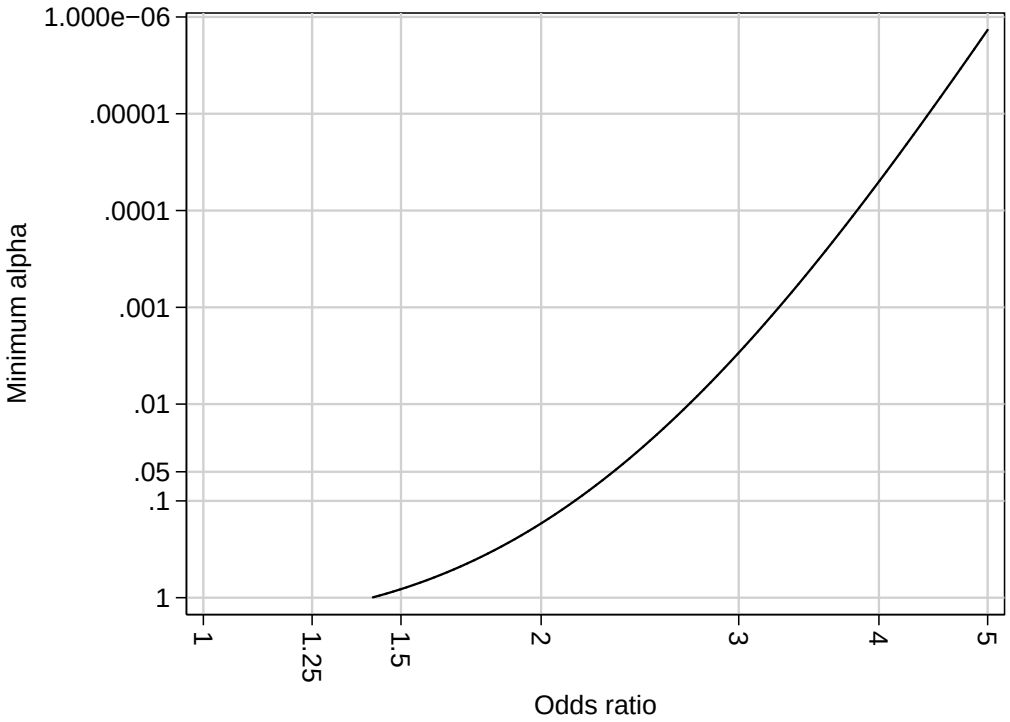

Figure 4. Significance level for 90% power with 100 cases as a function of odds ratio

Acknowledgements

I would like to thank Stanislav Kolenikov of the New Economic School, Moscow, Russian Federation, for posting some very useful formulas about the lognormal distribution on his website at *http://www.komkon.org/˜tacik/*.

References

Aitchison J. and J. A. C. Brown. 1963. *The Lognormal Distribution*. Cambridge, UK: Cambridge University Press.

Clayton D. and P. M. McKeigue. 2001. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* **358**: 1356-1360.

Colhoun, H. M. P. M. McKeigue and G. Davey Smith. 2003. Problems of reporting genetic associations with complex outcomes. *Lancet* 361: 865-872.

Hampel F. R. 1974. The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**: 383-393.

Hampel F. R., E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel. 1986. *Robust statistics. The approach based on influence functions*. New York: Wiley.

Huber P. J. 1981. *Robust statistics*. New York: Wiley.

McCullagh P. and J. A. Nelder. 1989. *Generalized Linear Models*. 2nd ed. London: Chapman & Hall.