The Stata Journal (2001) **1**, Number 1, pp. 1–22

Generalized power calculations for generalized linear models and more

Roger Newson
King's College London, UK
roger.newson@kcl.ac.uk

Abstract. The `powercal` package can compute any one of the 5 quantities involved in power calculations from the other 4. These quantities are power, significance level, detectable difference, sample number, and the standard deviation (SD) of the influence function, which is equal to the standard error multiplied by the square root of the sample number. `powercal` can take arbitrary expressions (involving constants and/or scalars and/or variables) as input, and calculates the output as a new variable. The user can therefore plot input variables against output variables, and this often communicates the tradeoffs involved better than a point calculation as output by the `sampsi` command. General formulas are given for calculating the SD of the influence function when the detectable difference is a linear combination of link functions of subpopulation means for an outcome variable distributed according to a generalized linear model (GLM). This general case includes a very broad range of special cases, where the parameters to be estimated are differences between subpopulation proportions, arithmetic means and algebraic means, or ratios between subpopulation proportions, arithmetic means, geometric means and odds. However, `powercal` is not limited to GLMs, and can even be used with rank methods.

Keywords: st0001, power, alpha, significance level, detectable difference, detectable ratio, sample number, standard deviation, influence function, sample design, generalized linear model, proportion, arithmetic mean, algebraic mean, geometric mean, odds

1 Introduction

When statisticians are not making their living producing confidence intervals and p-values, they are often producing power calculations, or urging their colleagues to involve them at the design stage, so that they can produce power calculations. The traditional tool for doing this in Stata is `sampsi`, which has several limitations. First, `sampsi` can only output power and sample size, and requires the detectable difference and desired significance level to be input. Second, it is only designed to output power and sample size for a limited range of parameters (differences between subpopulation means and proportions) for a limited range of designs (sampling in parallel from two or fewer populations). Third, `sampsi` outputs only point calculations, and does not produce plotted power curves, which often communicate the tradeoffs involved better than point calculations.

Because of these limitations, I wrote the `powercal` package, which is a low-level programming tool for calculating any of the 5 quantities involved in power calculations. These quantities are power, significance level, detectable difference, number of sample units, and the standard deviation (SD) of the per-unit influence function, defined as the standard error multiplied by the square root of the number of sampling units. Each one of these 5 quantities can be calculated as output from the other 4 as input. The input quantities can be specified by the user as expressions involving constants and/or scalars and/or variables, and the output quantity is calculated as a new variable. The user can therefore list and plot the input and output variables used by `powercal`. Detectable differences are defined very broadly, and may be logarithms of ratio parameters or other transformed parameters. Sample units are also defined very broadly, and each unit may be a cluster, or a set of primary units sampled in a defined ratio from subpopulations.

`powercal` is therefore a very comprehensive package. The price of its general usefulness is that the user may need to know some formulas, especially to calculate the SD of the influence function. However, this article also gives a guide to the derivation of such formulas. These usually follow a standard pattern, especially if the differences to be estimated are linear combinations of parameters from generalized linear models (GLMs). However, the usefulness of `powercal` is not limited to GLMs, and extends to other statistics for which a Central Limit Theorem applies, including many rank statistics.

2 The powercal package

2.1 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`]

2.2 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

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

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.

2.3 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 SD 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 SD 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 8 `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`.)

3 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 $\Upsilon(X;\theta,F) = X - \theta$.)

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 (SD) 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 SD 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 SD 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 for δ is defined (approximately) by $\hat{\delta} \pm G^{-1}(1-\alpha/2) \times \mathrm{SE}(\hat{\delta})$, 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

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

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

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

or, equivalently,

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

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 maxima 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

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

then the minima and maxima are defined as follows:

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

The operator $\lceil x \rceil$ represents the minimum integer no less than x, as calculated by the `ceil()` function in Stata 8. 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.

3.1 Formulas for the SD of the influence function

The parameter σ is usually an input parameter, 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 sampling 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 SD of the influence function is simply the SD 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

$$\delta = \sum_{j=1}^{K} a_j \eta_j - \omega \tag{14}$$

where ω and the a_j are constants. The contrast δ is assumed to be zero under a null hypothesis to be tested, and ω is usually (but not always) zero. 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

$$\hat{\delta} = \sum_{j=1}^{K} a_j \hat{\eta}_j - \omega \tag{15}$$

The standard error of $\hat{\eta}_j$ is

$$\mathrm{SE}\left(\hat{\eta}_j\right) \;=\; \sigma_j / \sqrt{n m_j} \tag{16}$$

where σ_j is the SD 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 SD of the per-PSU influence function is equal to

$$\sigma_j \;=\; \frac{d\eta_j}{d\mu_j}\sqrt{\phi V(\mu_j)} \tag{17}$$

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

$$\mathrm{SE}\left(\hat{\delta}\right) \;=\; \sqrt{\sum_{j=1}^{K} a_j^2 \left[\mathrm{SE}\left(\hat{\eta}_j\right)\right]^2} \tag{18}$$

It follows that the SD of the per-CSU influence function of δ is derived from the SDs of the per-PSU influence functions of the η_j by the formula

$$\sigma \;=\; \sqrt{n} \times \mathrm{SE}\left(\hat{\delta}\right) \;=\; \sqrt{\sum_{j=1}^{K} \frac{a_j^2}{m_j}\sigma_j^2} \tag{19}$$

Table 1: Some commonly used link functions for generalized linear models.

Link function	$\eta(\mu)$	$d\eta/d\mu$	*Interpretation of* η
Identity	μ	1	Arithmetic mean of Y
Power $r \neq 0$	μ^r	$r\mu^{r-1}$	Power-$1/r$ algebraic mean of Y^r
Log	$\ln(\mu)$	$1/\mu$	Log arithmetic mean of Y
Logit	$\ln\left[\mu/(1-\mu)\right]$	$1/\mu + 1/(1-\mu)$	Log odds of binary Y

Table 2: Some commonly used variance functions for generalized linear models.

Family	$V(\mu)$	*Interpretation of* ϕ
Normal	1	Variance
Gamma	μ^2	Squared coefficient of variation
Bernoulli	$\mu(1-\mu)$	Always 1
Poisson	μ	Variance/mean ratio

Table 1 gives some commonly used link functions for generalized linear models, with formulas for the link function η as a function of a subpopulation mean μ of a variable Y and for its derivative for use in Equation (17), and interpretations in words for the link η. Table 2 gives some commonly used variance functions for generalized linear models, which assume that the variance of a subpopulation with mean μ is equal to $\phi V(\mu)$, together with an interpretation in words of the dispersion parameter ϕ. Each variance

Table 3: Standard deviations of influence functions for some variance-link combinations.

Family	*Link*	σ_j	*Typical interpretation of* δ
Normal	Identity	$\sqrt{\phi}$	Difference between arithmetic means
Normal	Log	$\sqrt{\phi}/\mu_j$	Log ratio between arithmetic means
Gamma	Identity	$\mu_j\sqrt{\phi}$	Difference between arithmetic means
Gamma	Log	$\sqrt{\phi}$	Log ratio between arithmetic means
Poisson	Identity	$\sqrt{\phi\mu_j}$	Difference between incidence rates
Poisson	Log	$\sqrt{\phi/\mu_j}$	Log ratio between incidence rates
Bernoulli	Identity	$\sqrt{\mu_j(1-\mu_j)}$	Difference between proportions
Bernoulli	Log	$\sqrt{(1-\mu_j)/\mu_j}$	Log ratio between proportions
Bernoulli	Logit	$\sqrt{1/\mu_j + 1/(1-\mu_j)}$	Log ratio between odds

function applies to a distributional family, from which it derives its name. There are many other possible link functions and variance functions, and more comprehensive tables can be found in the Appendices of Hardin and Hilbe (2001).

Table 3 gives formulas for the SD of the influence function for the jth subpopulation derived according to Equation (17) for some common combinations of variance and link functions. The σ_j can be entered into Equation (19) to derive a SD of the per-CSU influence function for the contrast δ of (14), whose typical informal interpretation in words is given in the right-hand column. Again, there are many more possible combinations, some of which are mentioned in Hardin and Hilbe (2001). In particular, we may plan to transform the outcome data prior to analysis. If we use the log transformation and a generalized linear model with the identity link, then the parameter δ will be a log ratio between geometric means, or (in other words) a difference between arithmetic mean logs. If we use a power-r transformation and a generalized linear model with the power-$1/r$ link, then δ will be a difference between power-r algebraic means, where the power-r algebraic mean of a variable Y is defined as $[\mathrm{E}(Y^r)]^{1/r}$. Note that these links can be combined with any variance function.

4 Examples

The examples in the help file for `powercal` are designed to work both under Stata 7 and under Stata 8, and are described in detail in the Adobe Acrobat manual `powercal.pdf`, distributed with the `powercal` package. In this paper, we give more advanced examples, demonstrating the power of Stata 8 graphics.

4.1 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 SD 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 (defined as the SD/mean ratio) 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 SD of the natural logs can be calculated from the coefficient of variation or the qth tail ratio by the formulas

$$\mathrm{SD}_{\log} \;=\; \sqrt{\ln\left(\mathrm{CV}^2+1\right)} \;=\; -\ln(r_q)/\left[2\Phi^{-1}(q)\right] \tag{20}$$

where $\mathrm{SD}_{\log}$ is the SD 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 Subsection 3.1, the log outcomes are distributed according to a generalized linear model with an identity link function and a Normal variance function, the $\eta_j = \mu_j$ are subpopulation arithmetic mean logs (or logs of geometric means), and the dispersion parameter ϕ is the variance of the log outcomes, usually assumed to be the same in all subpopulations. Therefore, from Table 3, the σ_j are the within-subpopulation SDs of the log outcomes. We wish to know the SD 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 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 (20). By (19), the SD of the per-CSU influence function of δ is then given by

$$\sigma = \mathrm{SD}_{\log} \times \sqrt{2} \tag{21}$$

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 (with $2 \times 50 - 2 = 98$ degrees of freedom). 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.

```
. 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 logratio=log(2)*(_n/_N)
. lab var logratio "Log GM ratio"
. gene gmratio=exp(logratio)
. lab var gmratio "GM ratio"
. powercal power, alpha(0.01) delta(logratio) sdinf(sdlog*sqrt(2)) ///
>   nunit(50) tdf(98)
Result to be calculated is power in variable: power
. line power gmratio, ///
>   yscale(range(0 1)) ///
>   ylab(0(0.05)1, grid gmin gmax angle(0)) yline(0.8 0.9, lpattern(shortdash))
>  ///
>   xscale(log range(1 2)) xlab(1(0.1)2, grid gmin gmax)
```

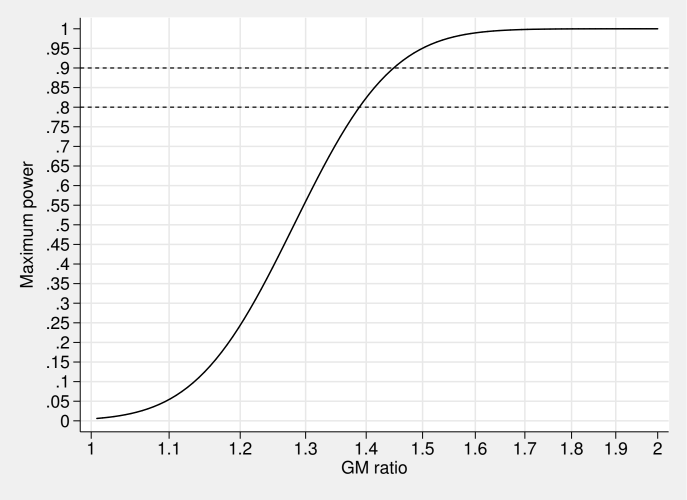

Figure 1: Power to detect geometric mean ratios.

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, which states that, 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.

```
. 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, ///
>   ylabel(, angle(0) grid gmin gmax) yline(1, lpattern(shortdash)) ///
>   ytitle("Detectable GM ratio") ///
>   xlab(0(10)100, grid gmin gmax)
```

4.2 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

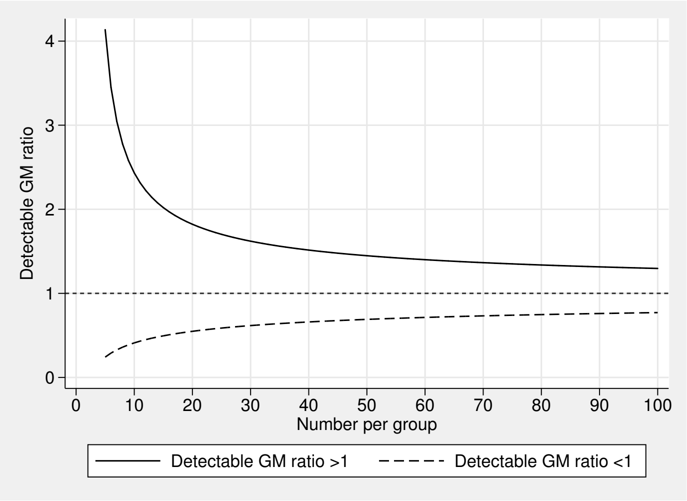

Figure 2: Detectable geometric mean ratios as a function of number per treatment group.

"with the disease" and "without the disease"), and then measure, in each subject, the exposure, which might be the presence of a genetic pattern 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 Subsection 3.1, 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 link function 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 $\phi = 1$. From the bottom row of Table 3, the SD of the per-PSU influence function of the log odds η_j is

$$\sigma_j = \sqrt{1/E_j + 1/(1-E_j)} \qquad (22)$$

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 SD of the per-CSU influence function is then given by the formula (19). 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-control studies, 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 Subsection 3.1, we have $a_1 = 1$ and $a_2 = -1$. In this case, the SD of the per-CSU influence function is given, according to (19), by

$$\sigma = \sqrt{\frac{1}{E_1} + \frac{1}{1-E_1} + \frac{1}{m_2}\left[\frac{1}{E_2} + \frac{1}{1-E_2}\right]} \tag{23}$$

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 OR for the odds ratio, then 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

$$E_1/(1-E_1) = \mathrm{OR} \times E_2/(1-E_2) \tag{24}$$

and then calculate the hypothesized case exposure prevalence E_1 from the case exposure odds $E_1/(1-E_1)$. Note that, if we have an estimate for the control exposure E_2, then the SD of the per-case influence function of the log odds ratio, given by (23), is dependent on the log odds ratio itself. This is in contrast to the case with lognormal geometric mean ratios, where σ is independent of δ and is given by (21).

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 (24) and (23) to calculate the SD of the influence function of the log odds ratio in `sdinflor`. We then use `powercal` to calculate the minimum numbers of cases to detect each odds ratio with 90% power at significance levels $p \leq 0.01$ and $p \leq 0.001$, respectively, and plot these odds ratios against those minimum numbers of cases, suppressing odds ratios which require 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.

```
. 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:         2,424 (99.8% of memory free)
-------------------------------------------------------------------------------
              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 ncases01, power(0.9) alpha(0.01) delta(logor) sdinf(sdinflor)
Result to be calculated is nunit in variable: ncases01

. powercal ncases001, power(0.9) alpha(0.001) delta(logor) sdinf(sdinflor)
Result to be calculated is nunit in variable: ncases001

. lab var ncases01 "Minimum cases (alpha=0.01)"

. lab var ncases001 "Minimum cases (alpha=0.001)"

. line or ncases01 if ncases01<=2000 || line or ncases001 if ncases001<=2000 ,
> ///
>   yscale(log range(1 5)) ylabel(1 1.5 2(1)5, angle(0) grid gmin gmax) ///
>   xlab(, grid gmin gmax) xtitle("Minimum number of cases") ///
>   legend(label(1 "Alpha=0.01") label(2 "Alpha=0.001"))
```

Using the same data set, we can also calculate, for each odds ratio, the significance levels attainable with 50% or 90% power, using 100 cases and their 200 controls. This

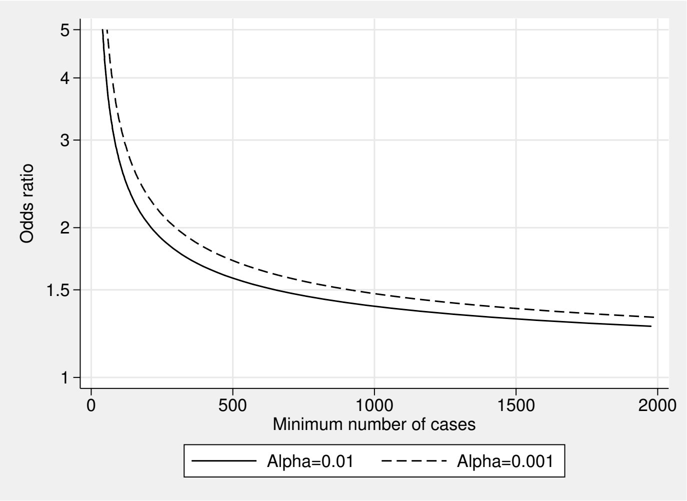

Figure 3: Detectable odds ratios as a function of number of cases.

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 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 at 90% power, or at the 0.1% level with 50% power. 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 Section 35.7 of Kirkwood and Sterne (2003) for a discussion of these problems and Colhoun et al. (2003) for their importance in genomic epidemiology.)

```
. * Significance level by odds ratio *
. powercal alpha50, power(0.50) delta(logor) sdinf(sdinflor) nunit(100)
Result to be calculated is alpha in variable: alpha50

. powercal alpha90, power(0.90) delta(logor) sdinf(sdinflor) nunit(100)
Result to be calculated is alpha in variable: alpha90

. line alpha50 or || line alpha90 or, ///
>   yscale(log range(1e-9 1) reverse) ///
>   ylab(1 0.05 1e-1 1e-2 1e-3 1e-4 1e-5 1e-6 1e-7 1e-8 1e-9, ///
>     angle(0) format(%8.2g) grid gmin gmax) ///
>   yline(0.05 0.01 0.001, lpattern(shortdash)) ///
>   xscale(log) xlab(1 1.25 1.5 2(1)5, grid gmin gmax) ///
>   legend(label(1 "Power=0.50") label(2 "Power=0.90"))
```

4.3 Example 3. Somers' D and ranksum tests

The methods of the `powercal` package are not limited to generalized linear models, and may equally well be used if we plan to analyse the data using rank statistical methods. For instance, we might want to measure typical differences in blood pressure between men and women, and we might expect the distribution to be positively skewed and possibly unequally variable between the sexes. Traditionally, this would be done using

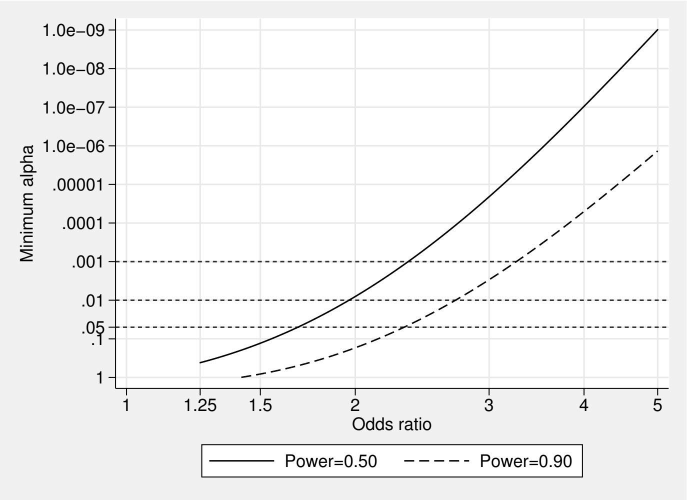

Figure 4: Significance levels for 50% and 90% power with 100 cases by odds ratio.

a Mann-Whitney-Wilcoxon ranksum test (see [R] **ranksum**), which produces a p-value but no confidence interval.

Today, most statisticians would argue that confidence intervals are more informative than p-values alone, even if rank methods are used. In Newson (2002), it is argued that there are at least three possible parameters corresponding to the so-called "nonparametric" ranksum test, namely Somers' D, the Hodges-Lehmann median difference and the Hodges-Lehmann median ratio, and that confidence intervals can be calculated for any of them using the `somersd` package, downloadable from SSC. Somers' D of blood pressure with respect to male gender is defined as the difference between two probabilities, namely the probability that a randomly-sampled male has a higher blood pressure than a randomly-sampled female and the probability that a randomly-sampled female has a higher blood pressure than a randomly-sampled male. Power formulas are more easily defined for Somers' D than for the other two parameters. This is because Somers' D is closely related to Kendall's tau and has a very well-behaved influence function, for which the Central Limit Theorem works very well at low sample numbers, whereas influence functions for medians are very unpredictable. See Hampel et al. (1986) and Huber (1981) for discussion on the influence functions of medians, and Kendall and Gibbons (1990) for discussion of the Central Limit Theorem as applied to Kendall's tau.

Although Somers' D is well-behaved, it is difficult to understand for non-statisticians, who would usually like to be able to convert it to a scale of median differences or ratios, as this would probably be more useful for making monetary or other practical decisions. Unfortunately, there is no unique conversion formula. However, if an outcome variable Y (such as blood pressure) has a Normalizing transformation $g(Y)$, which is Normally distributed within each of two subpopulations being compared (such as males and females), and if $g(Y)$ has mean μ_A and variance ϕ_A in Population A and has mean

μ_B and variance ϕ_B in Population B, then the Somers' D of Y, with respect to a binary variable X equal to 1 for Population A and 0 for Population B, is given by

$$D_{YX} \; = \; 2\Phi\left[(\mu_A - \mu_B)/\sqrt{\phi_A + \phi_B}\right] - 1 \qquad (25)$$

where $\Phi(\cdot)$ is the standard Normal cumulative distribution function. Under these assumptions, Somers' D has the same sign as $\mu_A - \mu_B$, and therefore the same sign as the Hodges-Lehmann median difference, but the conversion curves between the scales depend both on the function $g(Y)$ and on the sum of the subpopulation variances.

As stated in Subsection 3.1, the key to a power calculation formula for a parameter is a formula for the SD of its influence function. Somers' D is defined as a ratio of two U-statistics, in the terminology of Hoeffding (1948). Analytical standard error formulas can therefore be derived from the theory introduced there and discussed further in Serfling (1980), subject to making distributional assumptions. However, we will not use these methods here, but instead use as a pilot study the data set `bpwide`, distributed with official Stata, which contains one observation for each of 120 fictional patients and data on their genders and blood pressures. These blood pressures are in unstated units, but that is not a problem for us, as Somers' D is scale-invariant. (See [R] **sysuse** for more information about the datasets shipped with official Stata.) Instead of using the SD of the influence function for Somers' D itself, we will use the SD of the influence function for the hyperbolic arctangent (or z-transform) of Somers' D, as this transformation is variance-stabilizing, making the SD of the parameter influence function less dependent on the value of the parameter itself. The formulas are discussed in detail in the manual `somersd.pdf`, which is distributed with the `somersd` package, and is a post-publication update of Newson (2000).

In the following advanced example, we load the `bpwide` data and use `somersd` together with the `parmby` program from the `parmest` package (also downloadable from SSC and discussed in Newson (2003)). The results from these programs are used to calculate the SD of the influence function of the z-transformed Somers' D, for input into `powercal`. The variables created by `powercal` are plotted in Figures 5 and 6.

```
. sysuse bpwide, clear
(fictional blood-pressure data)
. gene byte male=1-sex
. lab var male "Male patient"
. parmby "somersd male bp_before, tr(z) td", norestore ///
>   escal(N) rename(es_1 N)
Command: somersd male bp_before, tr(z) td
Somers' D with variable: male
Transformation: Fisher's z
Valid observations: 120
Degrees of freedom: 119
Symmetric 95% CI for transformed Somers' D

------------------------------------------------------------------------------
             |              Jackknife
        male |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   bp_before |   .3086041    .110782     2.79   0.006     .0892446    .5279636
```

```
Asymmetric 95% CI for untransformed Somers' D
            Somers_D    Minimum    Maximum
bp_before  .29916667  .08900846   .4838229

. list N parm estimate stderr min* max* p, clean noobs

      N        parm   estimate      stderr       min95       max95           p
>
    120   bp_before   .3086041   .11078202   .08924464   .52796357   .00621746
>

. scal sdinf=stderr[1]*sqrt(N[1])

. disp _n as text "SD of influence function for z-transformed Somers' D: " ///
>   as result sdinf

SD of influence function for z-transformed Somers' D: 1.2135563

. drop _all

. set obs 1000
obs was 0, now 1000

. gene int npat=_n

. lab var npat "Number of patients"

. foreach X in 05 01 001 0001 {
  2.   powercal detz`X', power(0.9) alpha(0.`X') sdinf(sdinf) ///
>     nunit(npat) tdf(npat-1)
  3.   gene detd`X'=exp(2*detz`X')
  4.   replace detd`X'=(detd`X'-1)/(detd`X'+1)
  5.   lab var detz`X' "z-transformed Somers' D (P<=0.`X')"
  6.   lab var detd`X' "Somers' D (P<=0.`X')"
  7. }
Result to be calculated is delta in variable: detz05
(1 missing value generated)
(999 real changes made)
Result to be calculated is delta in variable: detz01
(1 missing value generated)
(998 real changes made)
Result to be calculated is delta in variable: detz001
(2 missing values generated)
(998 real changes made)
Result to be calculated is delta in variable: detz0001
(2 missing values generated)
(997 real changes made)

. line detd* npat, ///
>  xlab(0(100)1000, grid gmin gmax) ///
>  ylab(0(0.05)1, angle(0) grid gmin gmax) ///
>  ytitle("Detectable Somers' D")

. more

. graph export figseq5.eps, replace
(file figseq5.eps written in EPS format)

. foreach X in 10 15 20 30 {
  2.   scal z=0.5*log((1+0.`X')/(1-0.`X'))
  3.   powercal alpha`X', power(0.9) delta(z) sdinf(sdinf) ///
>     nunit(npat) tdf(npat-1)
  4.   lab var alpha`X' "Alpha (Somers' D = 0.`X')"
  5.   format alpha`X' %8.2g
  6. }
Result to be calculated is alpha in variable: alpha10
Result to be calculated is alpha in variable: alpha15
Result to be calculated is alpha in variable: alpha20
Result to be calculated is alpha in variable: alpha30
. line alpha* npat, ///
>   xlab(0(100)1000, grid gmin gmax) ///
>   yscale(log reverse) ///
>   ylab(1 0.05 0.01 1e-3 1e-4 1e-5 1e-6 1e-7 1e-8 1e-9 1e-10 1e-11, ///
>     format(%8.2g) angle(0) grid gmin gmax) ///
>   ytitle("Minimum alpha")
. more
. graph export figseq6.eps, replace
(file figseq6.eps written in EPS format)
```

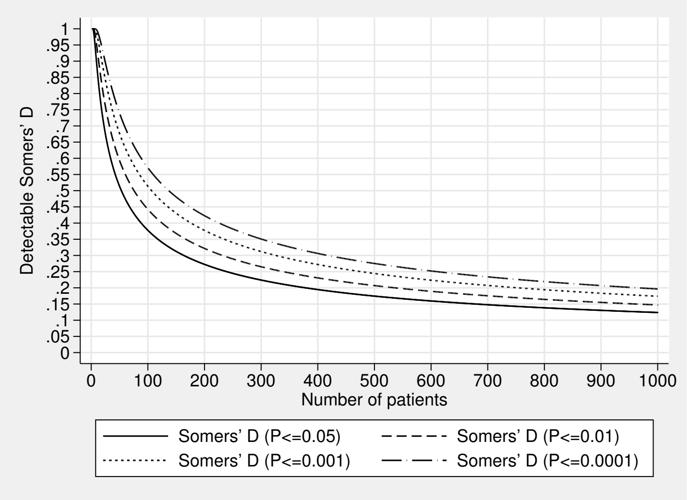

Figure 5: Detectable Somers' D by number of patients for 90% power.

We first load the `bpwide` data, then add a variable `male` indicating male gender, and then use `parmby` and `somersd` to estimate Somers' D and to create a second data set in memory, with 1 observation and data on the sample number, estimate, standard error, confidence limits and p-values for the z-transformed Somers' D. We find that the untransformed Somers' D is 0.29916667, so it is about 30% more likely for a man to have a higher blood pressure than a woman than *vice versa.* The standard error stored in the variable `stderr` is multiplied by the square root of the sample number stored in the variable `N` to give the SD of the influence function, which is stored in the scalar `sdinf` and equal to 1.2135563 z-units. We then create a third data set in memory, with 1000 observations (one for each possible sample number from 1 to 1000), and a variable `npat`, containing the number of patients. Then we use `powercal` in a loop to add to this data set 4 new variables `detz05`, `detz01`, `detz001` and `detz0001`, containing z-transformed Somers' D values detectable with 90% power at p-values 0.05, 0.01, 0.001 and 0.0001, respectively, and use the hyperbolic tangent or inverse z transform to derive detectable untransformed Somers' D values in `detd05`, `detd01`, `detd001` and `detd0001`, respectively. These are line-plotted against `npat` to create Figure 5. After this, we use `powercal` in another loop to add to the data set 4 new variables `alpha10`, `alpha15`, `alpha20` and `alpha30`, containing the significance level attainable with 90% power,

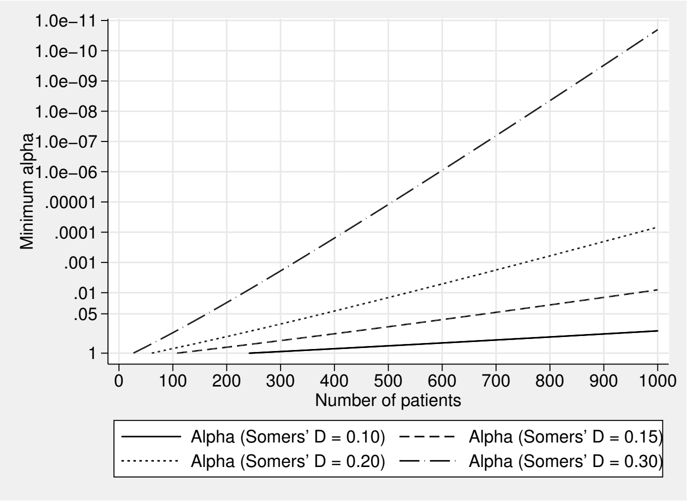

Figure 6: Attainable significance levels by number of patients for 90% power.

assuming population Somers' D values of 0.10, 0.15, 0.20 and 0.30, respectively. These are line-plotted against `npat` in Figure 6. Note that the alpha-values are plotted on a reverse log ordinate, so that the higher they are, the more statistically significant they are. The reverse log ordinate makes the attainable alpha-curves very nearly linear in the number of patients, indicating that the attainable p-value decreases approximately exponentially as patient numbers (and presumably costs) are increased.

5 Acknowledgements

I would like to thank Stanislav Kolenikov of the University of North Carolina, Chapel Hill, for posting some very useful formulas about the lognormal distribution on his website at *http://www.komkon.org/˜tacik/*.

6 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, NY: Wiley.

Hardin, J. and J. Hilbe. 2001. *Generalized Linear Models and Extensions.* College Station, TX: Stata Press.

Hoeffding, W. 1948. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* 19: 293–325.

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

Kendall, M. G. and J. D. Gibbons. 1990. *Rank Correlation Methods.* 5th ed. Oxford, UK: Oxford University Press.

Kirkwood, B. R. and J. A. C. Sterne. 2003. *Essential Medical Statistics.* 2nd ed. Oxford, UK: Blackwell Science.

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

Newson, R. 2000. snp15: `somersd` – Confidence intervals for nonparametric statistics and their differences. *Stata Technical Bulletin* 55: 47–55. In *Stata Technical Bulletin Reprints*, vol. 10, 312–322. College Station, TX: Stata Press. Post-publication update downloadable from Roger Newson's website at *http://www.kcl-phs.org.uk/rogernewson/*.

—. 2002. Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *The Stata Journal* 2(1): 45–64. Pre–publication draft downloadable from Roger Newson's website at *http://www.kcl–phs.org.uk/rogernewson/*.

—. 2003. Confidence intervals and p-values for delivery to the end user. *The Stata Journal* 3(3): 245–269. Pre–publication draft downloadable from Roger Newson's website at *http://www.kcl–phs.org.uk/rogernewson/*.

Serfling, R. J. 1980. *Approximation Theorems of Mathematical Statistics.* New York, NY: Wiley.

About the Author

Roger Newson is a Lecturer in Medical Statistics at King's College London, UK, working principally in asthma research. He wrote the packages `powercal`, `parmest` and `somersd`.