Supplementary Information to “Applying Generalized Funnel Plots to Help Design Statistical Analyses” by Aisbett, Drinkwater, Quarrie and Woodcock

Online Resource 1: Error rates and sample sizes for one-sided and equivalence tests.

This supplement presents the calculations behind *meraglim.shinyapps.io/genSSize*.

As in standard sample size calculations, we assume an anticipated sample distribution if the hypothesis of interest is true (which in standard NHST is typically taken to have mean Cohen’s d) and an anticipated sample distribution if the alternate hypothesis is true. We will make the usual assumption that the alternate mean is the value on the boundary of a one-sided test, allowing the Type I error rate to be equated with the alpha level of the test. Also, as in standard sample size calculations, we will ignore variability in the sample variance; for exact computations refer to Shieh (2016). We initially consider t-tests, then describe how the computations are adapted for correlations.

Denote the cumulative *t*-distribution with *k* degrees of freedom by and use to denote the percentile (so that). Denote the mean of the anticipated sample distribution if the hypothesis is true by , the alpha level of a one-sided tests by *α* and the standard error of the mean effect size by SE. Finally, suppose that effects less than -*Δ* and greater than *Δ* are practically meaningful.

Now suppose that we are either testing for effects less than so that , or are testing for effects larger than so that . In both cases, the Type II error rate is

, ()

so that, with equality only if. If we rearrange (1) to get

.  ()

This is the standard expression used in NHST sample size calculations except the effect magnitude has been replaced by the difference magnitude . The relationship between sample size and both SE (assuming fixed variance) and the degrees of freedom *k* means that sample size appears on both sides of the inequality. A recursive approach must be taken to obtain an exact solution, or the *t-*distribution may be approximated by a normal distribution. Our R implementation instead uses a parameter (set by default to 38) to compute the *t-*tests. When users have selected a test and an approximate sample size this parameter should be adjusted to gain a more accurate estimate. However, in practice this rarely changes estimates much and in any case sample size calculations involve many assumptions and approximations.

In the case of a two-group design with pooled variance and group sizes and for , it is well known that and Substituting into (2) gives the total sample size *n* to achieve required power level, . In a one-group design, so (2) gives as required sample size .

Equivalence involves two tests: that effects are either greater than *Δ*, or less than . The anticipated effect size *m* for the hypothesis of interest must satisfy. The Type I error rate is well-known to be α if both tests are at that level. The Type II error rate is the sum of the errors from the two tests, since the tails of the sample distribution can extend into meaningful regions in both the positive and negative directions. From (1) the error rate is therefore

. ()

When the Type II error rate is twice that of a single one-sided test. In the general case, it is clear from (3) that for any given , a target Type II error rate of *β* can be achieved when

and

.

Take and Then, as in the manipulation that derived (2), we see that (3) will be satisfied if

and ,

or equivalently, if

. ()

Now increases with since is negative whenever . Thus, the first of the terms in the brackets in (4) increases as *m* increases from -*Δ* while the second term decreases. The terms are equal when *m* = 0so the first term is the smaller for negative *m* and the second term is at least as small for non-negative *m.*  It follows that the Type II error rate for the equivalence test will be at most *β* if

. ()

Because of how we derived this inequality, it may under-estimate the standard error needed to achieve adequate power.

The inequality (5) can be applied as before to compute sample sizes as a function of the test levels and the anticipated effect size *m*. For instance, given a two-group design with equal sized groups and pooled variance , and we can apply (5) to derive the total sample size:

.

Shieh (2016) gives SAS and R code for power computations under other designs.

Note that for a given anticipated effect size *m* both the non-superiority and non-inferiority tests may each separately achieve the required power, but there is not enough power to satisfy both at the same time. Given a non-meaningful effect then => so that if inequality (2) is satisfied for a non-superiority test (test on ) then it will be satisfied for a non-inferiority test (test on Similarly, if , then satisfaction of (2) for a non-inferiority test ensures that non-superiority is also satisfied. So, for example, if and

then both non-inferiority and non-superiority tests have adequate power; but taken together they may not.

For simplicity, *meraglim.shinyapps.io/genSSize* does not explicitly flag the region where (5) is not satisfied but where both non-inferiority and non-superiority tests have adequate power at the largest of the user-specified alpha levels. Instead, when the anticipated effect size *m* is less than 0 only the non-superiority test is presented as being satisfied, and vice-versa.

Correlation sample size calculations assume bivariate normal distributions, with all computations performed after applying the Fisher transformation. That is, correlation is transformed to the (approximately) normally distributed random variable that has standard error . The equations discussed earlier are applied with this SE and with the normal distribution approximated by a *t*-distribution with *k* = 500. For example, using (2), the sample size *n* required to test that a correlation is greater than margin *Δ* given anticipated correlation is calculated to be

, where .

Note that the app *meraglim.shinyapps.io/genSSize* presents results in terms of correlations. It achieves this by applying the inverse transformation to the labels on the horizontal axis of the plot and to the horizontal co-ordinate of positions obtained by clicking on the plot.

# References:

Shieh G. Exact power and sample size calculations for the two one-sided tests of equivalence. PLoS ONE. 2016; 11(9): e0162093.