# SAMPLE SIZES FOR MAGNITUDE-BASED INFERENCES ABOUT CLINICAL, PRACTICAL OR MECHANISTIC SIGNIFICANCE

**Will G. Hopkins, FACSM**
**AUT University, Auckland, NZ**
**will@clear.net.nz**

**Med Sci Sports Exerc 38(5), S528, 2006 (Abstract #2746)**

## Abstract

**PURPOSE**. The traditional method of sample-size estimation based on statistical significance is not appropriate for a study designed to make an inference about real-world significance, which requires interpretation of magnitude of an outcome. I present here two new methods for estimating sample size for such studies, based on (a) acceptable error rates for a clinical or practical decision arising from the study and (b) adequate precision for a mechanism-related inference. **METHODS**. For (a) I devised two new types of error: deciding to use an effect that is actually harmful (a Type 1 clinical error), and deciding not to use an effect that is actually beneficial (a Type 2 clinical error). I then constructed a spreadsheet to calculate sample sizes for chosen values of Type 1 and 2 errors, for chosen smallest beneficial and harmful values of various outcome statistics and designs (changes or differences in means in controlled trials or cross-sectional studies, correlations in cross-sectional studies, relative risks in cohort studies, and odds ratios in case-control studies), and for chosen values of other design-specific statistics (error of measurement, between-subject standard deviation, proportion of subjects in each group, and incidence of disease or prevalence of exposure). The calculations are based on the usual assumption of normality of the sampling distribution of the outcome statistic. For (b) I reasoned that precision is adequate when the uncertainty in the estimate of an outcome statistic (represented by its confidence interval) does not extend into values that are substantial in both a positive and a negative sense when the sample value of the statistic is zero or null. Sample sizes are then derived from the spreadsheet by choosing equal Type 1 and 2 clinical errors (e.g., 5% for a 90% confidence interval). The sample sizes for both methods can be compared with those based on the traditional method, included in the spreadsheet. Also included are confidence limits and quantitative and qualitative chances of benefit and harm for the "decision value" and any other values of the outcome statistic. **RESULTS**. Sample sizes for Type 1 and 2 clinical errors of 1% and 20% are ~10% smaller than those for adequate precision with a 90% confidence interval, which in turn are only one-third of traditional sample sizes (for Type I and II statistical errors of 5% and 20%). Confidence limits and clinical chances provided by the spreadsheet are fully consistent with Type 1 and 2 clinical errors. **CONCLUSION**. Researchers can now justify and use sample sizes for studies aimed at making inferences about magnitudes. The sample sizes can be much smaller than those based on statistical significance.

## Key Points

- Some folks base their sample size on those of similar studies. Other folks study as many subjects as time or resources permit.
- Which is OK, but you should use a sample big enough to get a clear outcome.
- "Clear" is traditionally based on statistical significance.
- But some of us now make magnitude-based inferences about the clinical, practical or mechanistic significance of an outcome.
- I present here a spreadsheet for estimating sample sizes for this new approach to inferences. And for the old.
- For either approach, you have to input the smallest important value of the outcome or effect statistic.
- You also have to input maximum acceptable rates of error.
- Some designs need a bit extra information.
- The sample sizes based on the new approach are about one-third of those based on the old, for reasonable confidence about outcomes that are trivial.
- Smaller samples will still give you a clear outcome, if it turns out to be larger.
- Larger samples will give you more confidence about the magnitude.

**"You need a sample big enough to get a clear outcome in terms of real-world significance."**

## How the Spreadsheet Works

- Use the section of the spreadsheet for the design of your study: crossover, correlations, case-control, etc.
- Input the smallest important value of the effect you are studying.
  - **You have to come up with this value**, even for sample size based on statistical significance.
  - There are defaults for some standardized dimensionless measures: see the image of the spreadsheet at right.
  - You can use one smallest value for benefit and another for harm, if you like.
- For sample size based on **clinical or practical significance**…
  - Input a maximum acceptable chance of deciding to use something that is actually harmful.
    - I call this the Type 1 clinical error rate. My default is 1%.
  - Now input a maximum acceptable chance of deciding not to use something that is actually beneficial
    - I call this the Type 2 clinical error rate. My default is 20%.
  - The Type 1 and 2 clinical error rates are analogous to–but quite different from–the Type I and II error rates that you need for sample size based on statistical significance.
    - The usual defaults for these error rates are 5% and 20%.
    - See the lower half of the spreadsheet for sample sizes based on statistical significance.
- For some designs you input one or more extra bits of information: within-subject SD (error of measurement), between-subject SD, proportion of subjects in the groups, and incidences and prevalences of diseases or exposures.
- The required sample size is then shown.
  - The calculations are based on the usual assumption of normality of the sampling distribution of the outcome statistic.
  - The sample sizes are a bit less than a third of those based on statistical significance with Type I and II errors of 5% and 20%.
- Also shown are confidence limits for the true effect, and chances that the true effect is beneficial and harmful when you observe the "decision value" or "critical value".
  - You can change the observed value to see how it affects the confidence limits and chances of benefit and harm.
- For sample size based on mechanistic significance…
  - Use equal Type 1 and 2 clinical error rates.
  - My defaults are 5% (not shown in the image).
    - These error rates ensure that "the true effect can't be both substantially positive and negative".
    - In other words, the 90% confidence interval for the outcome cannot overlap substantially positive and negative values in the worst-case scenario of the observed effect being zero or null.
  - The sample sizes are about a third of those based on statistical significance.
- BONUS. The larger the true value of the effect, the smaller the sample for a clear outcome. To see how larger true effects affect sample size, put the suspected true value in the cell for the smallest harmful effect.

Type 1: rate (%) of using something you shouldn't, because it is actually harmful.

Type 2: rate (%) of not using something you should, because it is actually beneficial.

The chance that an effect that in reality is null will produce something that you will decide is beneficial. You want this chance to be low, and the only way to get that is to have a large disparity between the Type 1 and 2 clinical errors and a relatively high Type 2 rate.

If you want to use the smallest standardized (Cohen) change of 0.20, you must also express the typical error as a fraction of the between-subject SD.

This default is the smallest harmful change, with opposite sign.

Also known as the standard error of measurement. Estimated from a reliability study with conditions comparable to the crossover. If you have a retest correlation coefficient r, the typical error = SD*√(1-r), where SD is the between-subject standard deviation. The typical error is also the SD of the change scores divided by √2.

The default value is 0.20 of the between-subject SD, which is the smallest important standardized difference in the mean, according to Cohen. If you want to use a standardized difference, the between-subject SD has to be 1.00.

The default value is 0.10, which is the smallest important correlation, according to Cohen.

This value must be positive but it can be greater or less than 1.00.
The default value is 1.10, which represents a 10% greater rate or risk of something.

This default is 1/(smallest harmful relative rate).

Proportion (%) of unexposed or control subjects who end up with the disease (or other condition).

For a prospective cohort, the proportion (%) of unexposed subjects who will end up with the disease/condition. For an intervention, the proportion of subjects in the experimental (exposed) group. (Set to 50% for equal groups.)

Type I: rate (%) of null effect turning up statistically significant.

Type II: rate (%) of smallest effect not turning up statistically significant.

Change the **blue** numbers.
Answers are in **plum** and **red**. Some **plum** numbers are copies of **blue** numbers, not answers.
You can change a **plum** number to see what happens, but "undo" to retain the formula in that cell.
Annotations shown here in **green** are in the corresponding cells of the spreadsheet as comments.