SPORTSCIENCE · sportsci.org Latest issue
Perspectives / Research Resources This issue

Sample-size Estimation for Various Inferential Methods

Will G Hopkins

Sportscience 24, 17-27, 2020 (sportsci.org/2020/MBDss.htm)
Institute for Health and Sport, Victoria University, Melbourne, Australia. Email. Reviewer: David Rowlands, Massey University, Wellington, NZ.

I present here a spreadsheet for estimating sample size in studies using the magnitude-based decision method (MBD) or other inferential methods. The MBD sample size provides acceptable uncertainty defined either by error rates or the width of the compatibility interval. Sample sizes with the default error rates or width are approximately one-third those of the traditional method of null-hypothesis significance testing (NHST), which is included in the spreadsheet. The spreadsheet also provides estimates of sample size for superiority testing and equivalence testing. This article includes explanations of numerous specific issues related to sample-size estimation. I recommend using sample size for MBD, if the granting agency or journal allows it; otherwise use sample size for superiority testing, because sample size for equivalence testing is impractically large, and significance testing should be retired. Remember to increase sample size appropriately for drop-outs, clustering of subjects, measures with low validity, comparison of effects in subgroups, moderators, mediators, individual differences or responses, and multiple effects. If you can access only tens of subjects, do an intervention, preferably as a crossover, with a reliable dependent variable. KEYWORDS: clinical significance, compatibility limits, equivalence, NHST, research design, reliability, smallest important effect, superiority, statistical power, Type-I error, Type-II error, validity.

Reprint pdf · Reprint docx · Spreadsheet · Slideshow

Author's note. This article is a revised version of a previous article (Hopkins, 2006), reformatted with linked contents and including new material on sample sizes for superiority and equivalence testing, post-hoc justification of sample size, and a better method for sample size on the fly. Explanatory comments in the spreadsheet have also been updated.

Introduction

We study a sample of subjects to find out about an effect in a population. The bigger the sample, the closer we get to the true or population value of the effect. We don't need to study the entire population, but we do need to study enough subjects to get acceptable accuracy for the true value.

"How many subjects?" is a question I am often called on to answer, usually before a project is submitted for ethics approval. Sample size is an ethics issue, because a sample that is too large represents a needless waste of resources, and a sample that is too small will also waste resources by failing to produce a clear outcome. If the study involves exposing subjects to pain or risk of harm, an appropriate sample size is ethically even more important. Applications for ethical approval of a study and the methods section of most manuscripts therefore require an estimate of sample size and a justification for the estimate.

Free software is available at various sites on the Web to estimate sample size using the traditional approach based on statistical significance. However, my colleagues and I now avoid all mention of statistical significance in our publications, at least in those I coauthor. Instead, we make a decision about the importance of an effect, based on the uncertainty in its magnitude, using the magnitude-based decision method (MBD), formerly magnitude-based inference (MBI). See the article (Hopkins, 2020) and In-brief item in this issue for recent developments with MBD/MBI.

I therefore devised two new approaches to sample-size estimation for studies in which inferences are based on magnitudes. The new approaches are easily adapted to estimate sample size for superiority testing and equivalence testing. In this article I explain all these approaches, and I provide a spreadsheet for the estimates. I also explain various other issues in sample-size estimation that need to be understood or taken into account when designing a study.

In research, we make an inference or decision about the magnitude of an effect, usually about whether the magnitude is important or not. Whichever way the decision goes, we could be wrong, so there are two kinds of error. All methods of estimating sample size are based on keeping these error rates acceptably low.

Sample Size for Statistical Significance

In the traditional approach to null-hypothesis significance testing (NHST), you need a sample size that would produce statistical significance for an effect most of the time, if the true value of the effect were the smallest important value. Stating that an effect is statistically significant means that the observed value of the effect falls in the range of extreme values that would occur infrequently (<5% of the time, for significance at the 5% or 0.05 level) if the true value were zero or null. The value of 5% defines the so-called Type-I error rate: the chance that you will declare a null effect to be significant. "Most of the time" is usually assumed to be 80%, a number that is sometimes referred to as the power of the study. A power of 80% can also be re-expressed as a Type-II error rate of 20%: the chance that you will fail to get statistical significance for the smallest important effect. I have included this traditional approach to sample-size estimation in the spreadsheet accompanying this article and checked that it gives the same sample sizes as other tools (e.g., Dupont and Plummer's software).

NHST works best when you use the sample size as estimated, and when the values of any other parameters required for the calculation (e.g., error of measurement in a pre-post controlled trial, incidence of disease in a cohort study) turn out to be correct. In such rare cases you can interpret a statistically significant outcome as clinically or practically important and a statistically non-significant outcome as clinically or practically trivial. When the sample size is different from that calculated, statistical and clinical significance are no longer congruent. In any case, I think Type-I and Type-II errors of 5% and 20% lead to decisions that are too conservative (Hopkins, 2007a).

Sample Size for Magnitude-Based Decisions

MBD provides a more realistic approach to real-world importance of effects, but it needs its own method of sample-size estimation. In 2006 I devised two approaches. I did an extensive literature search but was unable to find anything similar at that time.

The new methods for estimating sample size are based on (a) acceptable error rates for a clinical or practical decision arising from the study,

and (b) adequate precision for the effect magnitude, which can also be expressed in terms of error rates. See the slideshow accompanying this article for figures illustrating these two approaches.

For (a) I devised two new types of error: a decision to use an effect that is actually harmful (a Type-1 clinical error), and a decision not to use an effect that is actually beneficial (a Type-2 clinical error). I then used statistical first principles to derive formulae to calculate sample sizes for chosen values of Type-1 and Type-2 errors (e.g., 0.5% and 25% respectively), for chosen smallest beneficial and harmful values of outcome statistics in various straightforward designs (crossovers, controlled trials, and so on), and for chosen values of other design-specific statistics (error of measurement, between-subject standard deviation, and so on). The formulae, which are based on the same assumption of normality of the sampling distribution of the effect or a related test statistic that underlies NHST sample sizes, were readily incorporated into a spreadsheet.

For (b) I reasoned that precision is adequate when the uncertainty in the estimate of an outcome statistic (represented by its compatibility interval) does not extend into values that are substantial in both a positive and a negative sense. The interval needs to be narrowest 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 Type-2 clinical errors (e.g., 5% for a 90% compatibility interval, or 2.5% for a 95% compatibility interval). Sample sizes for Type-1 and Type-2 clinical errors of 0.5% and 25% are almost identical to those for adequate precision with a 90% compatibility interval, which in turn are only one-third of traditional sample sizes for the usual default Type-I and Type-II statistical errors of 5% and 20%. For adequate precision with a 95% compatibility interval, the sample sizes are approximately half those of the traditional method.

The above explanations of sample-size estimation for MBD are essentially Bayesian, in that they involve limiting chances that the true effect has beneficial and harmful or substantial threshold values. But MBD can also be expressed in frequentist terms as several one-sided interval hypothesis tests (Aisbett et al., 2020; Hopkins, 2020), and the limiting chances can be expressed as expected error rates or alphas for those tests. From this perspective, the 0.5% risk of harm is the alpha level and Type-II error rate of the one-sided test on the harm threshold, which is called a non-inferiority test (in the positive direction). The 25% chance of benefit is the Type-II error rate for this non-inferiority test, assuming an expected effect on the benefit threshold and the sample size given by the MBD calculator. For non-clinical MBD, the 5% chance of a substantial effect is the Type-II error rate for a non-inferiority test on the substantial threshold, assuming an expected effect on the opposite threshold and the MBD sample size. Janet Aisbett (personal communication) has also pointed out that sample-size estimation in MBD is equivalent to controlling Type-III error rates in NHST, in the sense that the MBD error rates are the rates for mistaking a harmful, beneficial or substantial effect of one sign for a beneficial, harmful or substantial effect of the opposite sign, respectively.

Included in the spreadsheet for estimating sample size are compatibility limits and quantitative and qualitative chances of benefit and harm for the estimated "decision" (or any chosen) values: observed values greater than the decision value will lead you to decide that the effect is clinically beneficial. (The decision values are analogous to the "critical" values of the traditional method of sample-size estimation, above which observed values will be statistically significant.) The chances of benefit and harm for the decision value serve as a check on the accuracy of the formulae I devised to estimate the sample sizes: you will see that the clinical chances provided by the spreadsheet are the same as the Type-1 and Type-2 clinical errors. The compatibility limits also serve as a check, but you will have to change the level of the limits from 90% first to 99% and then to 50% to see that the lower and upper limits are the smallest importants for harm and benefit. With Type-1 and Type-2 errors set to 5%, the 90% limits are of course the same as the smallest importants. Thus, the compatibility intervals fall just within the trivial range of effect values, which leads to a compelling frequentist justification of minimum acceptable sample size for MBD: when the study is done, harmful and beneficial (or substantial negative and positive) values of the effect will not be simultaneously compatible with the data and analytical model, for the chosen levels of compatibility.

Also included in the spreadsheet are panels of cells for outcomes of studies of mean changes and differences when the true effect is zero (or

any chosen value). For the sample size given by the default Type-1 and Type-2 errors of 0.5% and 25%, you will see that the chances of deciding to use a null effect (i.e., obtaining at least possibly beneficial) are appreciable (up to 17%, depending on the design), when the sample size is minimum desirable. For non-clinical effects, clear possibly substantial outcomes occur more frequently, but an effect is decisively substantial only when it is at least very likely, which occurs <1% of the time. These error rates and those with other chosen values of the true effect match closely the error rates obtained by simulation for standardized effects in a controlled trial (Hopkins & Batterham, 2016).

Sample Size for Superiority and Equivalence Testing

In an article showing how magnitude-based decisions are equivalent to several hypothesis tests (Hopkins, 2020), I was obliged to defend the above methods of sample-size estimation against the assertion of Aisbett et al. (2020) that sample size would be better based on superiority testing (also known as minimum-effects testing) and/or equivalence testing. With such tests, the researcher wants a high probability of deciding that an effect is substantial or trivial, when the true magnitude of the effect is respectively some substantial expected value greater than the smallest important or some trivial value smaller than the smallest important. I stated in the article that the MBD sample size is that of superiority testing for the reasonable expectation of a borderline small-moderate true effect, but the sample sizes for equivalence are unrealistically large for most researchers (e.g., 16× the MBD sample size for the reasonable expectation of a true trivial effect equal to half the smallest important). Equivalence testing is an option in a meta-analysis, because the effective sample size is the sum of the sample sizes in all the included studies, albeit with some reduction due to clustering of subjects into studies (explained below).

The MBD spreadsheet can be used to estimate the superiority and equivalence sample sizes, as follows. Set the Type-1 and Type-2 errors to 5%. For superiority, put the expected substantial true value into the cell for the smallest beneficial effect. Now put the smallest important value (with the same sign as the expected substantial value) into the cell for the smallest harmful effect. If you make the expected substantial value 3× the smallest important (borderline small-moderate), you will get the same sample size as for MBD. For equivalence, put the expected trivial true value (positive or negative) into the cell for the smallest harmful effect. Now put the smallest important value with the same sign into the cell for the smallest beneficial effect. If you make the expected trivial value half the smallest important, you will get 16× the MBD sample size.

You might notice that the compatibility intervals with the sample sizes for superiority and equivalence fall precisely in the range defined by the values you have inserted in the smallest-important cells, which is consistent with the way superiority and equivalence are defined by hypothesis tests. The chances of magnitude of the true effect can be interpreted according to the values in the smallest important cells, but the cells on the far right showing chances of clear outcomes for chosen true effects are difficult to interpret.

Interestingly, the sample size for NHST also produces compatibility intervals that fall precisely between two threshold values: a 95% CI on the null side of the critical value just touches the zero, and a 60% CI away from the null just touches the smallest important. It follows that the spreadsheet or other tool for estimating sample size for NHST can be adapted easily to estimate sample size for superiority testing, equivalence testing, and MBD (Aisbett et al., 2020). NHST usually involves a two-sided test of the zero hypothesis, so the Type-I error first needs to be set to 10% to achieve a 90% CI that would touch the zero, whereas the Type-II error is set to 5% to achieve a 90% CI that would touch the smallest important. Now set the smallest important value in the spreadsheet or other tool to the following values: for superiority testing, the expected substantial effect minus the smallest important; for equivalence testing, the smallest important minus the expected trivial effect; and for non-clinical MBD (which is the same as for clinical MBD), twice the smallest important (the positive smallest important minus the negative smallest important).

Specific Sample-Size Issues

Whether you use the spreadsheet for the traditional or new approaches, there are several important sample-size issues you should know about when designing a study. Some of these are implicit in the spreadsheet, but you will need to take others into account yourself.

Sample Size in Similar Studies

Sample-size estimation is challenging for the average researcher, so mistakes are common.

Check your estimate by comparing it with sample sizes in published studies that have measures, subjects, and design similar to yours.

You can even justify a sample size on the grounds that it is similar to those in similar studies that produced clear or statistically significant outcomes, but be aware that effects are clear or significant in many studies because the effects are big enough to make them so. See how wide the compatibility interval is in these studies, using my spreadsheet (Hopkins, 2007a) to generate it, if necessary; if your effect turns out to be smaller but with a compatibility interval of similar width, will your effect be clear or will you need a larger sample?

Choice of Smallest Important Effect

All methods for estimation of sample size need a value for the smallest important effect. The estimates are sensitive to the value: sample size is proportional to the inverse of the square of the smallest important, so halving it results in a quadrupling of sample size. Your justification of sample size must therefore include a justification of choice of the smallest important effect. For details of smallest important effects for all the different kinds of outcome, see Linear Models and Effect Magnitudes (Hopkins, 2010). For details of smallest importants for athlete performance, see a slideshow focusing on medal-winning enhancements presented at the performance analysis conference in 2016 and/or Appendix 1 of Assessing an Individual (Hopkins, 2017a). Here is a brief summary from these links…

A standardized difference or change of ±0.20 (i.e., Cohen's d, 0.20 of the between-subject SD) can be used for **continuous test scores or measures** only when it is not possible to work out a difference or change in the measure that would be associated with a smallest important difference or change in health, wealth or competitive performance of your subjects. In most settings, the SD should be the true or pure SD_P, not the observed SD_O, which is inflated by the typical or standard error of measurement (within-subject SD) e: $SD_O^2 = SD_P^2+e^2$. Hence, the smallest difference or change is $0.2SD_P = 0.2\sqrt{(SD_O^2–e^2)}$ or $0.2SD_O\sqrt{r}$, where $r = SD_P^2/SD_O^2$ is the intraclass or retest correlation. In other words, if the observed SD is used to define the smallest important difference or change, it should be multiplied by the square root of the retest correlation. The time-frame of the error of measurement or retest correlation should reflect the time-frame of the effect to be studied. If you are interested in acute differences or changes, the typical error or retest correlation should come from a short-term reliability study that effectively measures technical error only. If instead you are interested in stable differences or changes over a defined period (e.g., six months), then the smallest important change in the mean (or difference the mean, in a cross-sectional study) should come from the pure between-subject SD over such a period.

For single **Likert or visual-analog scales,** rescale the values to range from 0 to 100 (representing "percent of full-scale deflection"); the smallest important is then ±10.

Smallest effects for **competition times or distances of solo athletes** are ±0.3 of the within-athlete competition-to-competition variability in time or distance (equivalent to an extra medal won or lost every 10 competitions); for test measures directly related to competition performance, work out what 0.3 of the variability corresponds to.

Analyze **match play** as chances of winning using the logistic-regression form of generalized linear modeling; the smallest important is then ±10% in an otherwise even match (i.e., 55% vs 45%, equivalent to an extra win or loss every 10 matches).

The smallest important **hazard or count ratio** is 0.9 and its inverse, 1.11 (equivalent to one extra occurrence in every 10); smallest risk and odds ratios are the same, for low incidence or prevalence (<10%).

Effect of Research Design

The following assertions are easily verified with the spreadsheet. Non-repeated measures studies (cross-sectional, prospective, case-control) usually need 100s or even 1000s of subjects. Repeated-measures interventions (crossovers and controlled trials) usually need fewer subjects (scores or sometimes 100s), depending on the reliability of the dependent variable, as explained below. Crossovers need less than parallel-group controlled trials (down to one quarter), provided reliability does not worsen too much during the washout period. So, if you have limited access to subjects or limited time or resources, you should choose a design and research question to accommodate the number you can investigate.

Drop-outs

Sample-size estimates for prospective studies and controlled trials should be inflated by 10-30% to allow for drop-outs, depending on the demands placed on the subjects, the duration of the

study, and incentives for compliance.

Clustering of Subjects

A good example of clustering is a design where the subjects are players from a number of teams. If players within a team tend to have similar effects (i.e., different from those of other teams), the effective sample size is less than the total number of players. To account for such clustering, the analysis should include a random effect for the clusters, and sample size in theory needs to be increased by a factor of 1+r(c-1), where r is the intra-cluster correlation coefficient and c is the mean cluster size. It follows that you should keep the cluster size as small as possible or include as many clusters as possible.

In practice, r is difficult to determine. The formula for r is (between)/(between+within), where between and within are the pure between-cluster variance and the within-cluster variance respectively. As such, r would need to be estimated in an exploratory study–usually an impractical option. For a repeated-measures design, r is the intra-cluster correlation for change scores, so the exploratory study would have to be done with the intended interventions–again, impractical.

Bottom line: with clustering, sample size needs to be greater than the usual estimate, but you can't be expected to estimate it before you do the study. You should therefore initially study at least as many clusters as will give you the usual estimate of sample size, then either do more clusters using sample size on the fly, or if you can't access more clusters, do a post-hoc justification of sample size.

Expected Magnitude of the Effect

A larger true effect needs a smaller sample size. You can understand this assertion by considering sample size for acceptable uncertainty: the compatibility interval for a trivial effect has to be sufficiently narrow not to overlap small positive and negative values, whereas the compatibility interval for a large positive or negative effect can be much wider before it overlaps small negative or positive values. But the width of the compatibility interval is inversely proportional to the square root of the sample size, so the wider compatibility interval for larger effects implies a smaller sample size. The spreadsheet has instructions on how to estimate sample size for a true effect expected to be larger than the smallest important: replace the smallest important beneficial effect with the expected larger effect.

Sample Size "On the Fly"

In this approach, also known as a group-sequential design, you study a series of cohorts, accumulating sample size all the while, until you get a clear or significant outcome. This approach is a practical way to deal with the various uncertainties in the estimation of sample size; it is also ethically superior to using a fixed sample size, because it reduces waste of resources and risk to subjects.

In the original version of this article, I provided the following crude approach to estimating the approximate sample size for a second cohort, when the first produces an unclear outcome. Assume that the observed effect in the second cohort will be the same as in the first cohort, then see how much narrower the compatibility interval needs to be for a clear outcome with the combined sample size for this effect. The width of the compatibility interval is inversely proportional to the square root of the sample size, so some simple maths provides an estimate of the number of extra subjects. Note, however, that this sample size will give only a 50% chance of a clear outcome, so you may need to repeat this process several times.

An approach to estimate the sample size for only one extra cohort with a much lower risk of an unclear outcome than with the above method is to assume the magnitude of the effect observed with the first cohort is the true value, then to estimate the minimum desirable sample size for such an effect (by inserting this value of the effect in the cell for smallest beneficial change). For repeated-measures designs, you will need to insert the within-subject SD (typical error), estimated with the effect in your study using the panel of cells in the top right of the spreadsheet.

When statistical significance is used to terminate sampling, the group-sequential approach is known to produce biased outcomes and inflated error rates, but software is available to avoid these problems (Rogers et al., 2005). Errors and bias need to be investigated for the above two methods to terminate sampling when an effect becomes clear. The second approach is bound to be better.

Unavoidably Small Sample Size

A sample size smaller than the minimum desirable is ethically defensible, if the true effect is likely to be large enough for the outcome to be clear. You can also argue that an unclear outcome with a sample size that isn’t way too small will still set useful limits on the likely magnitude

of the effect and will therefore be worth publishing, because it will contribute to a meta-analysis. To obtain a value for the smallest true effect that your sample size will estimate with acceptable precision (i.e., practically guaranteed a clear outcome), change the value of the smallest important beneficial effect in the accompanying spreadsheet until it gives your sample size. Provide this value and its compatibility interval in a proposal, ethics application and Methods section of a manuscript. Use the compatibility limits to comment on the uncertainty for the effect that your study will establish, if you end up observing a trivial effect, a small effect, and so on.

If your study involves repeated measurement (monitoring, crossovers, pre-post controlled trials), one way to offset an unavoidably small sample size is to take more than one measurement at each time point of interest, then average the measurements at each point. The error of measurement of the mean of n measurements is the error divided by $\sqrt{n}$, and the required sample size is proportional to the square of the error (see the section on reliability below), so taking three measurements pre and three measurements post an intervention will decrease the required number of subjects to one-third (provided the error doesn't increase between the pre and post measurements). See the In-brief item When N is <10 for more on this and other strategies for coping with small samples.

"Post-hoc" Justification of Sample Size

If you did a study with sample size determined by available participants or resources, you might find that a reviewer or editor of the submitted manuscript requests a sample-size justification. For effects with a clear outcome, you can state that your sample size was already adequate, but for unclear effects, you can estimate the sample sizes that would make them clear using either or both of the two methods explained above for sample size on the fly. You can and probably should also state the minimum desirable sample size provided by the spreadsheet. To do that, or to use the second (and better) of the two on-the-fly methods, a repeated-measures design requires an estimate of the within-subject SD, which you can get by inserting your estimate of the effect and its compatibility limits in the panel of cells at the top right of the spreadsheet. Use that SD to estimate the usual minimum desirable and the on-the-fly sample sizes.

Even if you used the correctly estimated minimum desirable sample size, you could get an inconclusive outcome, thanks to sampling variation. The likelihood of such an outcome, which I have estimated by simulation, is at most ~10%. (Interested academics can download a zip file of spreadsheets showing the simulations. The spreadsheets can be tweaked to show that increasing the sample size by ~25% makes the likelihood of an inconclusive outcome negligible.) If you find yourself in this position, estimate the required extra sample size using the methods for sample size on the fly.

Effect of Validity

For non-repeated measures designs, sample size depends on validity of the dependent variable. This principle follows from the fact that the random error represented by less-than-perfect validity increases the uncertainty in the outcome statistic, so more subjects are needed for acceptable uncertainty. From first principles, the sample size is proportional to $1/v^2 = 1+e^2/SD^2$, where v is the validity correlation coefficient, e is the error of the estimate, and SD is the between-subject standard deviation of the criterion variable in the validity study. Sample size thus needs to be doubled when the validity correlation is 0.7 and quadrupled when it is 0.5. Such adjustments are not included in the spreadsheet.

Validity of a predictor variable has the same effect on sample size as validity of the dependent variable in a non-repeated measures design. However, the effect of less-than-perfect validity also manifests itself as a reduction in the magnitude of the effect of the predictor, the reduction being proportional to v and v^2 for correlations and slopes respectively, where v is the validity correlation for the predictor–hence the need for a larger sample size. The so-called correction for attenuation is therefore a factor of $1/v$ (or $1/\sqrt{r}$, if reliability error is the only source of validity error). In contrast, validity and reliability of a dependent variable affect the uncertainty of a difference or change in a mean, but have no effect on its expected magnitude.

Effect of Reliability

With controlled trials and other repeated-measures designs for mean effects, sample size is sensitive to reliability of the dependent variable, again because of the effect of error on uncertainty. From statistical first principles, sample size is proportional to $(1–r) = e^2/SD^2$, where r is the test-retest reliability correlation coefficient, e is the error of measurement, and SD is the observed between-subject standard deviation. Thus

sample sizes of only a few subjects are theoretically possible for measures of sufficiently high reliability, although you should always have at least 10 subjects in each group to reduce the chance that the sample substantially misrepresents the population. This effect of reliability on sample size is implicit in the spreadsheet, because you have to enter the error of measurement (the within-subject SD) to get the sample size.

The estimate of measurement error should come from a reliability study of duration similar to that of the intervention. The resulting sample size may still be an underestimate, because any individual responses to the treatment will effectively inflate the error of measurement and thereby widen the compatibility interval for the treatment effect. In any case, it's often hard to find reliability studies with a dependent variable and time between trials comparable with those in your intended study, but you can often find comparable crossovers or controlled trials. I have therefore included a panel of cells in the sample-size spreadsheet to estimate within-subject SD from such studies. The published studies needn't have the same kind of intervention, but try to find some with similar time between trials and similar subjects, because the approach is based on the assumption that the error in the published study or studies is similar to what will be in your study. It's also assumed that individual responses to the treatment in your study will be similar to those in your study. This assumption may be more realistic and conservative than the usual approach of using the error from a reliability study, in which there are of course no individual responses. You could address this issue in your Methods section where you justify sample size, if you use this approach.

Group Sample Size for Group Comparisons

With designs involving comparison of groups (e.g., a parallel-groups controlled trial), make the groups of equal size to give the smallest total size. If the size of one group is limited only by availability of subjects, a larger number of subjects for the comparison group will increase the precision of the outcome, but more than five times as many subjects in the comparison group gives no further practical increase in precision. You can check this assertion with the spreadsheet.

Comparison of an Effect in Subgroups

When you want to compare an effect between independent subgroups, a surprising consequence of statistical first principles is that you will need *twice* as many subjects in *each* subgroup to get the same precision of estimation for the comparison as for either subgroup alone, representing a four-fold increase in sample size. Thus, for example, a controlled trial that would give adequate precision with 20 subjects would need 40 females and 40 males for adequate precision of the comparison of the effect between females and males. Comparisons of effects in subgroups therefore should not be undertaken as a primary aim of a study without adequate resources.

Gender subgroups are a special concern, because researchers who have difficulty recruiting enough subjects of one gender sometimes recruit a small proportion of the other gender and analyze the outcome without regard to gender. This approach is misguided. If you do not adjust for gender, you bias the mean effect towards that of the larger group. But when adjusting for gender, stats packages average the separate effects for the males and females. The resulting effective sample size is actually *less* than that of the larger group, when less than ~25% of the subjects are in the smaller group. Download a simple spreadsheet I devised to illustrate this point. Conclusion: try to recruit the minimum desirable sample size for at least one gender. If you include a smaller sample size of the other gender, analyze the genders separately. Compare the genders with a third analysis, but the comparison may be unclear. This conclusion applies also to ethnicity and other subgroups of subjects differing in a subject characteristic that could modify the effect.

Modifiers and Mediators

The above rule about quadrupling sample size for subgroup analyses applies also to the sample size needed to estimate the linear modifying effect of a continuous predictor, such as a measure of habitual training, when its effect is evaluated as the effect of 2 SD of the predictor (Hopkins, 2010). With such analyses, you are effectively comparing the effect for a group of subjects who are 1 SD above the mean with the effect for subjects 1 SD below the mean.

Adjustment of an effect to the mean value of a moderator can actually *reduce* the sample size required for the effect itself, when the moderator has a substantial effect (because it explains otherwise unexplained variance). The most important example is adjustment to the mean value of the dependent variable at baseline in a crossover or controlled trial. The reduction in sample

size depends on the relative magnitudes of the within- and between-subject SDs, and the sample-size spreadsheet takes this dependency into account. Note that you should adjust for this and other potential moderators, even if their effects are unclear.

A potential mediator of a treatment effect in a crossover or controlled trial is analyzed by including its change score as a main-effect predictor in the linear model. As such, its required sample size is twice that of the mean effect, or four times if the mediator is included as an interaction with the group effect in a controlled trial (implying a potentially different mechanism in control and experimental groups).

Individual Responses

In a controlled trial, the magnitude of individual responses should be determined before and after the subject characteristic(s) that might help to explain them have been included as modifiers. The magnitude of individual responses is expressed as a standard deviation (SD_{IR}) free of measurement error (e.g., ±2.6% around the treatment's mean effect of 1.8%). The sample size for adequate precision in the estimate of SD_{IR} in the worst-case scenario of zero change in the mean and zero SD_{IR} is ~$6.5n_{\Delta}^2$, where n_{Δ} is the sample size required for adequate precision in the change in the mean. See an In-brief item in the 2018 issue of Sportscience for the derivation of this formula (Hopkins, 2018b).

The conclusion is that sample size for adequate precision of individual responses is impractically large. Researchers should aim instead for the more practical sample size for adequate precision of potential effect modifiers and mediators that might explain individual responses. The sample size for effect modifiers and mediators is "only" 4× the sample size for adequate precision of the mean change, as explained above. The standard deviation for individual responses is still worth estimating, and for sufficiently large values it will be clear. See sample size on the fly to determine how much larger your sample would need to be to get a clear effect for SD_{IR}, if it is unclear. For more on the neglected but increasingly important issue of individual responses, see the full article on individual responses (Hopkins, 2018a) an item on "individual responses made easy" (Hopkins, 2015), and links from the most recent article on controlled trials in this journal (Hopkins, 2017b).

Multiple Effects

When you investigate more than one effect in a study, there is inevitable inflation in the chances of making errors. For example, imagine you studied two independent effects and found chances of harm and benefit of 0.4% and 76% for one effect and 0.3% and 56% for the other. If you decide to use both effects, the chance of doing harm overall is 0.7%, which exceeds the default threshold of 0.5%. Opting to use only the most important or pre-planned effect would keep the chance of harm below 0.5%, but you would thereby fail to use an effect that has a chance of benefit of either 56% or 76%, which is way above the default threshold of 25% and represents potential waste of a beneficial effect.

You could have avoided this scenario by using a sample size that kept the overall Type 1 and 2 errors to <0.5% and <25%. For the worst case of independent effects that are on the borderline for making a decision one way or the other, the spreadsheet provides the sample size when you set the Type 1 and 2 errors to 0.5/n% and 25/n%, where n is the number of independent effects. (These values are approximations; exact values are $100[1 - [1\text{-}e/100]^{1/n}]$, where e is the Type 1 or 2 percent error, but the simpler formulae are accurate enough.) The same formulae apply when estimating sample size with Type I and II statistical errors. For two effects the spreadsheet shows that sample size needs to increase by nearly 50%, and for four effects the sample size needs to be doubled. If the effects are not independent, for example in a study where you intend to choose the best of three or more treatments, sample size usually does not need to be increased to the same extent. Exactly how big it should be is difficult to estimate, so err towards studying too many subjects rather than too few.

Case Series

Sample size for a case series is not included in the spreadsheet. A case series is aimed at establishing *norms* of specific measures to allow confident characterization of future cases relative to the norms. (*Cases* can also refer to normal subjects, if the aim is to characterize a subject characteristic, such as a skill.) Assuming the measure or an appropriate transform is normally distributed, norms are established with a mean and SD estimated with adequate precision. The uncertainty in the mean needs to be less than the default of 0.2 SD, which is achieved with a sample size one-quarter that of a cross-sectional study, or about 70 subjects for 90% compatibility limits. This sample size also gives uncertainty of ×/÷1.15 for the SD, which is sometimes used as

the smallest important difference in an SD. Smaller sample sizes establish noisier norms, which result in less confident characterization of future typical cases but acceptable characterization of future unusual cases. Larger samples are needed to characterize percentiles accurately, especially when the measure is not normal distributed.

Single-subject Studies

The number of repeated observations in a single-subject study is analogous to the sample size for a sample-based study and can be estimated using the same procedures. The smallest important effect used in the calculation should be the same as for a sample-based study, because the effects that matter for a single subject are still the same as for subjects in general.

The simplest example is a set of repeated measurements pre and post an intervention, and you assess the change in the mean. Assume that the measurements within the pre cluster (and within the post cluster) are independent, that is, that they differ only because of random error of measurement (the within-subject SD or typical error). The analysis is then effectively the same as when you compare the means of two groups of subjects: groups are the pre and post measurements, and the between-subject SD is the typical error. The example I provided in the In-brief item in the current issue for a measure of solo athletic performance shows a sample size of 120 measurements: 60 pre and 60 post! With only a few repeated measurements, trivial or small observed mean changes in the individual athlete would be inconclusive.

Sample size to account for and estimate a trend in an individual athlete's tests is too difficult to estimate from equations. The workbook for monitoring an individual includes a spreadsheet that allows simulation of the measurements, using which you can determine whether the number of measurements you can take will be adequate. Read the article accompanying the workbook, and see below for more on using simulation for sample size.

Measurement Studies

These are performed to estimate measures of validity or reliability of any measures and factor structure of psychometric inventories. Sample size for such studies is not included in the spreadsheet, but it shows a dependence on magnitude similar to that for the other designs. Extremely high reliability or validity (observed error much less than the smallest important effect) can be characterized with as few as 10 subjects, because the upper compatibility limit for the true error is still negligible. More modest observed validity or reliability (correlations ~0.7-0.9; errors of measurement of ~2-3× the smallest important effect) need samples of 50-100 subjects for reasonable compatibility that the true values of validity or reliability aren't substantially higher or lower than the observed values. Studies of diagnostic tests require hundreds of subjects to ensure adequate sampling of the various subject characteristics that can modify diagnostic accuracy. Studies of factor structure need hundreds and preferably a thousand or so subjects, because the factors are derived from pairwise correlations amongst all the items in the inventory, so there is massive inflation of error due to multiple effects.

If you are planning a repeated-measures study, in principle you should first do a reliability pilot study aimed at determining the error of measurement (within-subject SD) for estimation of the required sample size. There are other reasons for doing a pilot study, including feasibility of the protocol and familiarization of the researchers and the participants, but the sample size required for adequate precision of estimate of the error of measurement is generally impractically large. In any case, it is better to estimate the error from published studies, as explained above in the section on reliability. When you have done the study, you can estimate the error using the same approach with your effect and provide it in your manuscript.

Simulation for Sample Size

Simulation can be used to determine sample size for complex designs or analyses, especially those involving non-linear models or combinations of repeated measurements or other correlated dependent variables. You make reasonable assumptions about errors and relationships between the variables. You then generate data sets of various sizes using appropriately transformed random numbers to represent the errors and relationships. Finally you analyze the data sets to determine the sample size that gives acceptable width of the compatibility interval. An advantage of this approach is that you have to consider carefully the nature of the data and the intended analysis before you begin, which could lead to improvements in the design. It also provides the ideal vehicle for a *sensitivity analysis*, in which you explore how changes in parameters

and errors affect the outcome statistic. The article on understanding statistics via simulation (Hopkins, 2007b) will get you started on doing simulations with a spreadsheet.

Conclusions

I recommend MBD sample-size estimation justified either in Bayesian or frequentist terms. If a grant application or manuscript is being submitted to an agency or journal that is hostile towards MBD, you should show an estimate for superiority testing with an expected small-moderate effect. This estimate is the MBD sample size for clinical error rates of 0.5% and 25% (or non-clinical error rates of 5%), so you can also state in a grant application that, if the effect is not decisively beneficial or harmful (or decisively substantial, for a non-clinical effect), your sample size will provide precision (compatibility limits) that should be adequate for publication: the effect will not be compatible with benefit and harm (or substantially positive and negative, for a non-clinical effect). Try to avoid justifying sample size with NHST, as you will need 3× as many subjects, and we are supposed to "retire statistical significance" (Amrhein et al., 2019). Sample sizes for equivalence testing are far too large for most researchers.

Remember to increase sample size appropriately for drop-outs, clustering of subjects, measures with low validity, comparison of effects in subgroups, moderators, mediators, individual differences or responses, and multiple effects. If you can access only tens of subjects, do an intervention, preferably as a crossover, with a reliable dependent variable.

Acknowledgements: Thanks to Janet Aisbett and the reviewer, Dave Rowlands, for providing corrections and suggestions for improvement.

References

Aisbett J, Lakens D, Sainani KL. (2020). Magnitude based inference in relation to one-sided hypotheses testing procedures. SportRxiv, https://osf.io/preprints/sportrxiv/pn9s3/.

Amrhein V, Greenland S, McShane B. (2019). Retire statistical significance. Nature 567, 305-307.

Hopkins WG. (2006). Estimating sample size for magnitude-based inferences. Sportscience 10, 63-70.

Hopkins WG. (2007a). A spreadsheet for deriving a confidence interval, mechanistic inference and clinical inference from a p value. Sportscience 11, 16-20.

Hopkins WG. (2007b). Understanding statistics by using spreadsheets to generate and analyze samples. Sportscience 11, 23-36.

Hopkins WG. (2010). Linear models and effect magnitudes for research, clinical and practical applications. Sportscience 14, 49-58.

Hopkins WG. (2015). Individual responses made easy. Journal of Applied Physiology 118, 1444-1446.

Hopkins WG. (2017a). A spreadsheet for monitoring an individual's changes and trend. Sportscience 21, 5-9.

Hopkins WG. (2017b). Spreadsheets for analysis of controlled trials, crossovers and time series. Sportscience 21, 1-4.

Hopkins WG. (2018a). Design and analysis for studies of individual responses. Sportscience 22, 39-51.

Hopkins WG. (2018b). Sample size for individual responses. Sportscience 22, i-iii.

Hopkins WG. (2020). Magnitude-based decisions as hypothesis tests. Sportscience 24, 1-16.

Hopkins WG, Batterham AM. (2016). Error rates, decisive outcomes and publication bias with several inferential methods. Sports Medicine 46, 1563-1573.

Rogers MS, Chang AMZ, Todd S. (2005). Using group-sequential analysis to achieve the optimal sample size. BJOG An International Journal of Obstetrics and Gynaecology 112, 529-533.

Published July 2020.