SPORTSCIENCE · sportsci.org Latest issue

Perspectives / Research Resources This issue

Estimating Sample Size for Magnitude-Based Inferences

Will G Hopkins

Sportscience 10, 63-70, 2006 (sportsci.org/2006/wghss.htm)

Sport and Recreation, AUT University, Auckland 0627, New Zealand. Email. Reviewers: Greg Atkinson, Research Institute for Sport and Exercise Sciences, Liverpool John Moores University, Liverpool L3 2ET, UK; Alan M Batterham, School of Health and Social Care, Teesside University, Middlesbrough TS1 3BA, UK.

Sample-size estimation based on the traditional method of 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 a spreadsheet using two new methods for estimating sample size for such studies, based on acceptable uncertainty defined either by the width of the confidence interval or by error rates for a clinical or practical decision arising from the study. The new methods require sample sizes approximately one-third those of the traditional method, which is included in the spreadsheet. The following issues are also addressed in this article: choice of smallest effect, sample size with various designs, sample size "on the fly", dealing with suboptimal sample size, effect of validity and reliability of dependent and predictor variables, sample size for comparison of subgroups, sample size for individual differences and responses, sample size when adjusting for subgroups of unequal size, sample size for more than one important effect, the number of repeated observations in single-subject studies, sample sizes for measurement studies and case series, and estimation of sample size by simulation.

KEYWORDS: clinical significance, confidence limits, research design, reliability, smallest worthwhile effect, statistical power, Type 1 error, Type 2 error, validity.

Reprint pdf · Reprint doc · Spreadsheet · Slideshow.ppt · Slideshow.pdf

Update Jan 2018. I previously asserted that adequate precision for the estimate of the standard deviation representing **individual responses** in a controlled trial was similar to that for the subject characteristics that potentially explain the individual responses. That assertion was incorrect. In an In-brief item in the 2018 issue of this journal (Hopkins, 2018), I show that the required sample size in the worst-case scenario of zero mean change and zero individual responses is $6.5n^2$, where n is the sample size for adequate precision of the mean. The bullet point on individual responses has been updated accordingly. 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 in the updated bullet point for analyses of subgroups and continuous moderators and a new bullet point for mediators. The standard deviation for individual responses should still be assessed, and for sufficiently large values it will be clear.

Update June 2017. The spreadsheet now takes into account the reduction in sample size that occurs when the control treatment in a crossover and the **pretest in a controlled trial is included as a covariate,** which it always should be. The usual error variance is reduced by a factor $1–e^2/(2SD^2)$, where SD is the observed between-subject SD and e is the typical (standard) error of measurement. When SD >> e (a highly reliable measure) there is no reduction, but at the other extreme, SD = e (i.e., there are no real differences between subjects–a very unreliable measure, with intraclass or retest correlation = 0), the error variance and therefore sample size is reduced by up to one half, depending on the degrees of freedom of the t statistics in the remaining formulae.

Update April 2016. Sample sizes for designs where the dependent variable is a **count** of something have now been updated to include crossovers and controlled trials. The estimates are based by default on the normal approximation to the Poisson distribution, whereby the observed between-subject SD of the counts is the square root of the mean count (the expected SD when the counts in each subject arise from independent events). The estimates also allow for "over-dispersion" and "under-dispersion" of the counts. With over-dispersion, underlying real differences between subjects' counts produce an observed between-subject SD greater than the square root of the mean count. With under-dispersion, which is less common, the observed SD is less than expected, possibly because of sampling variation rather than any real under-dispersion in the counts.

This panel in the spreadsheet is configured for smallest effects defined by a ratio of the counts, the default being 0.9 or its inverse 1.11. For smallest effects defined by standardization, just use the earlier panel highlighted in yellow, according to which a smallest effect of 0.20 requires ~272 subjects (136+136) for a group comparison or parallel-groups controlled trial (or a similar number for a crossover and 4x as many for a pre-post controlled trial).

Update October 2015. I have added a comment cell with extra information about smallest changes and differences in means of continuous variables in crossovers, controlled trials, and group comparisons. In particular, I now indicate how to take into account error of measurement when using standardization, according to which the smallest difference or change is 0.2 of the between-subject standard deviation (SD). 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 e: $SD_O^2 = SD_P^2 + e^2$. Hence, the smallest difference or change is $0.2SD_P = 0.2\surd(SD_O^2 – e^2)$ or $0.2SD_O\surd r$, 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.

Update August 2014. Cells for calculating the rate of various kinds of magnitude-based outcome when the true effect is null worked previously for clinical outcomes but did not give correct rates for non-clinical outcomes. These cells have been simplified and updated to allow estimation of rates for any true value. The effect of changing the sample size on the observed change required for a clear outcome has now also been added.

Updates June 2013. Within-subject SD (typical or standard error of measurement) is needed to estimate sample size for crossovers and pre-post controlled trials, but it's often hard to find reliability studies with a dependent variable and time between trials comparable with those in your intended study. However, you can often find comparable crossovers or controlled trials, so I have devised a panel 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 or 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.

Updates June 2011. A panel for a count outcome is now added to the spreadsheet. The smallest important effect is shown as a count ratio of 1.1, as explained in the article on linear models and effect magnitudes in the 2010 issue of Sportscience.

The panel for event outcomes now allows inclusion of smallest beneficial and harmful effects as risk difference, odds ratio and hazard ratio (in addition to the risk ratio that was there originally). The calculations for the event outcomes are based on assumption of a normal distribution for the log of the odds ratio, and the sample sizes for risk difference, hazard ratio and risk ratio are computed by converting the smallest effects for these statistics into odds ratios.

Sample-size estimation when there is repeated measurement of a dependent variable representing a count or an event is not yet included in the spreadsheet.

There is now a bullet point on the issue of the sample size needed in a reliability pilot study.

The reviewer of these updates (Greg Atkinson) suggested I include a comment about sample size for equivalence studies, which are aimed at showing that two treatments are practically equivalent. To put it another way, what is the sample size for acceptable uncertainty in the estimate of the difference in the effects of the two treatments? My novel approaches to sample-size estimation address precisely this question.

Update June 2008: a bullet point on likelihood of an inconclusive outcome with an optimal sample size; also, slideshow now replaced with an updated version presented at the 2008 annual meeting of the American College of Sports Medicine in Indianapolis (co-presented by Stephen W Marshall, who made useful suggestions for changes to some slides).

Update Mar 2008: advice on how to estimate a value for the smallest effect that a suboptimal sample size can estimate adequately now added to appropriate bullet point; also more in the bullet point on choosing smallest effects and their impact on sample size.

Update Nov 2007: a bullet point on sample size for adequate characterization of effect modification; that is, the sample size to determine the extent to which the effect differs in subgroups or between subjects with different characteristics.

Updates to Oct 2007: a bullet point on estimation of sample size when you have more than one important effect in a study and you want to constrain the chance of error with any of them; a paragraph reconciling 90% confidence intervals with Type 1 and 2 errors of 0.5% and 25%; a minor addition to the bullet point on sample size on the fly; other minor edits.

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 ethical approval. Sample size is an ethical 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 an inference about the importance of an effect, based on the uncertainty in its magnitude. See the article by Batterham and Hopkins (2005a) for more. I have therefore devised two new approaches to sample-size estimation for studies in which inferences are based on magnitudes. In this article I explain the traditional and new 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.

While preparing a talk on sample-size estimation in 2008, I realized that there is a kind of unified theory that ties together all methods of sample-size estimation, as follows. In research, we make inferences about effects. The inference results in a decision or declaration about the magnitude of the effect, usually the smallest magnitude that matters. Whatever way the decision goes, we could be wrong, so there are two kinds of error. We estimate a sample size that keeps both error rates acceptably low.

Sample Size for Statistical Significance

According to this traditional approach, 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 worthwhile 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 deal with the choice of the value of this effect later.

The traditional approach 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, and when other effects are estimated from the same data, statistical and clinical significance are no longer congruent. In any case, I have found that Type I and II errors of 5% and 20% lead to decisions that are too conservative

(Hopkins, 2007). Some other approach is needed to make inferences about the real-world importance of an outcome and to estimate sample sizes for such inferences.

Sample Size for Magnitude-Based Inferences

I have been aware of this problem for about 10 years, during which I have devised two approaches that seem to be suitable. Two years ago I did an extensive literature search but was unable to find anything similar, although it is apparent that a Bayesian approach can achieve what I have achieved and more (e.g., Joseph et al., 1997). However, I have yet to see the Bayesian approach presented in a fashion that researchers can access, understand, and use. A recent review of sample-size estimation was entirely traditional (Julious, 2004).

I have worked my approaches into a spreadsheet that hopefully researchers *can* use. I have included the traditional approach and checked that it gives the same sample sizes as other tools (e.g., Dupont and Plummer's software). 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. I presented these methods as a poster at the 2006 annual conference of the American College of Sports Medicine (Hopkins, 2006a).

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 constructed a spreadsheet using statistical first principles to calculate sample sizes for chosen values of Type 1 and 2 errors (e.g., 0.5% and 25% respectively), for chosen smallest beneficial and harmful values of outcome statistics in various straightforward designs (changes or differences in means in controlled trials or cross-sectional studies, correlations in cross-sectional studies, risk ratios 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 or its log transform.

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, or 2.5% for a 95% confidence interval). Sample sizes for Type 1 and 2 clinical errors of 0.5% and 25% are almost identical to those for adequate precision with a 90% confidence interval, which in turn are only one-third of traditional sample sizes for the usual default Type I and II statistical errors of 5% and 20%. For adequate precision with a 95% confidence interval, the sample sizes are approximately half those of the traditional method.

Perceptive readers may wonder if there is a problem with providing 90% confidence intervals in a paper and using them to make calls about effects being clear, while at the same time making a decision to use an effect only if the chance of harm is <0.5% (which is equivalent to a 99% rather than a 90% confidence interval not overlapping into harmful values). Although the sample sizes estimated by both methods are practically identical, there will indeed be occasions when an effect is conclusive by one method but inconclusive by another. An effect can also be clear and trivial on the basis of a 90% confidence interval but decisive and clinically useful on the basis of chances of benefit and harm. It is easy to generate these scenarios using the spreadsheet for confidence limits and clinical chances (Hopkins, 2007).

Included in the spreadsheet are confidence limits and quantitative and qualitative chances of benefit and harm for any chosen values of the outcome statistic. The default values shown in the spreadsheet are the calculated "decision" 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 confidence limits and chances of benefit and harm for the decision values serve as a check on the accuracy of the formulae I devised to estimate the sample sizes. You will see that the confidence limits and clinical chances provided by the spreadsheet are fully consistent with the Type 1 and 2 clinical errors.

Also included are outcomes of studies for the estimated or any other sample size when the true effect is null (zero for differences in means, zero for correlation coefficients, 1.0 for rate ratios). For the sample size given by the default Type 1 and 2 errors of 0.5% and 25%, you will see that the chances of deciding to use a null effect are appreciable (up to 17%). Fortunately, for smaller sample sizes this figure declines rapidly. The chance of observing non-trivial outcomes that appear to be clear is the 10% you would expect for 90% confidence limits with a true null effect, when the sample size is optimal. This figure may seem high, but it is less problematic when you express these non-trivial outcomes with their full probabilities. As can be seen from the spreadsheet, only ~2.2% of the outcomes would be "likely [or probably] non-trivial", and <0.1% would be "very likely non-trivial". Thus 7.8% of the 10% would be "possibly [or maybe] non-trivial", which seems acceptable. With suboptimal sample sizes the "likely non-trivial" outcomes balloon out to a maximum of 17%, so you will need to be cautious about borderline clear outcomes when your sample size is much smaller than it ought to be. Of course, if you use more than the estimated sample size, the error rates are smaller.

General 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 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 justify a sample size on the grounds that it is similar to those in **similar studies** that produced clear outcomes, but be aware that effects are clear in many studies because the effects are substantial. See how wide the confidence interval is in these studies, using my spreadsheet (Hopkins, 2007) to generate it, if necessary; if your effect turns out to be smaller but with a confidence interval of similar width, will your effect be clear or will you need a larger sample?
- All methods for estimation of sample size need a value for the **smallest important effect**. The estimates are sensitive to the value: 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 most continuous measures the default can be Cohen's thresholds of 0.20 for a standardized difference or change in means and a correlation of 0.10. In observational studies the resulting sample size is ~270 for the defaults of my default methods. A reasonable default for a hazard, risk or odds ratio in an intervention is ~1.10-1.20, because a 10-20% change in the incidence of an injury or illness would affect one or more groups in a community, however rare the condition. A risk ratio of this order is quantifiable in a well-controlled large-scale intervention, but expert epidemiologists consider that biases inherent in most cohort and case-control studies effectively set the smallest *believable* risk ratio in such studies to ~3.0 (Taubes, 1995). This limitation is bad news for public health but good news for researchers who can’t afford huge sample sizes. Smallest effects for measures directly related to the performance of solo athletes are ~0.5 of the competition-to-competition variability in performance (Hopkins, 2004; Hopkins, 2006b); the resulting sample sizes are usually many times larger than most researchers use.
- Sample size **depends on the design**. Non-repeated measures studies (cross-sectional, prospective, case-control) usually need hundreds of subjects. Repeated-measures interventions (crossovers and controlled trials) usually need scores of subjects. Crossovers need less than parallel-group controlled trials (down to one quarter), provided reliability does not worsen too much during the washout period. These assertions are easily verified with the spreadsheet. 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.
- To take account of any **clustering of subjects**, you can in theory inflate sample size by a factor of 1+*r*(*c*-1), where *r* is the intracluster correlation coefficient and *c* is the mean cluster size. It follows that you should keep the cluster size as small as possible. The formula for *r* is (between)/(between + within), where be-

tween and within are the pure between-cluster variance and the within-cluster variance respectively. As such, *r* is difficult to guestimate and would need to be estimated in an exploratory study. For a repeated-measures design the r is for change scores, so the exploratory study would have to be done with the intended interventions–usually an impractical option.

- 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.
- A **larger true effect** needs a smaller sample size. You can understand this assertion by considering sample size estimated via acceptable uncertainty. The confidence interval for a trivial effect has to be sufficiently narrow not to overlap small positive and negative values, whereas the confidence 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 confidence interval is approximately inversely proportional to the square root of the sample size, so the wider confidence interval for larger effects implies a smaller sample size. When you have to use a small sample size, it follows that you will still get a clear outcome, if the true effect is sufficiently large. On the other hand, if the outcome is unclear, you will find it more difficult to publish the work. The spreadsheet has instructions on how to estimate sample size for larger effects.
- The relationship between effect magnitude and sample size makes it possible to determine **sample size "on the fly"**, whereby you study a series of cohorts of subjects until you get a clear outcome. This approach, also known as a group-sequential design, 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. When statistical significance or lack of it 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. (See Rogers et al., 2005) The extent of error and bias when adequate precision and acceptable clinical error rates are used to terminate sampling needs to be investigated. Meanwhile, estimate the approximate sample size for an additional cohort by assuming the true value of the effect is the value in subjects already assayed, then see how much narrower the confidence interval needs to be for a clear outcome with this effect. The width of the confidence interval is inversely proportional to the square root of the sample size, so some simple maths will provide an estimate of the number of extra subjects. Note that this sample size will give only a 50% chance of a clear outcome, so you may need yet another cohort.
- An unavoidably **suboptimal sample size** (i.e., smaller than the size estimated for acceptable errors with the smallest important effect) 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 effect your sample size will estimate with acceptable confidence, change the value of the smallest important effect in the accompanying spreadsheet until it gives your sample size. Provide this value and its confidence interval in a proposal, ethics application and Methods section of a manuscript. Use the confidence interval to comment on the "useful limits" in the proposal or ethics application, if you end up observing a trivial effect.
- Even **optimal** sample sizes can produce **inconclusive** outcomes, thanks to sampling variation. The likelihood of such an outcome, which I have estimated by simulation, is at most ~10%. For the approaches based on statistical and clinical significance, this maximum occurs with small sample sizes and apparently when the true value is equal to the critical and decision value respectively, while for the confidence-interval approach it occurs when the true value is null. Interested academics can download a zip file (9 MB) 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.
- 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.

- With controlled trials and other repeated-measures designs, 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 standard deviation) to get the sample size.
- The **estimate of measurement error** used to estimate sample size in a repeated-measures intervention has to 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 confidence interval for the treatment effect. Sample size on the fly is one way to allow for individual responses.
- **Validity of a predictor variable** in any design 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 manifests itself as a reduction in the magnitude of the effect of the predictor, the reduction being proportional to v, 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.
- 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.
- When you want to compare an outcome 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, 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. This sample size and rule applies also to the **linear modifying effect of a continuous predictor**, such as height of subjects, when its effect is evaluated as the effect of 2SD of the predictor (Hopkins, 2010); that is, the effect for a group of subjects who are 1SD above the mean minus the effect on subjects 1SD below the mean.
- 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).

- In a controlled trial, the **magnitude of individual responses** needs to be determined 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 $\sim 6.5n_{\Delta}^2$, 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, 2018). The standard deviation for individual responses is still worth estimating, and for sufficiently large values it will be clear. See the above bullet point about 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 articles on controlled trials in this journal (Batterham and Hopkins, 2005b; Hopkins, 2003; Hopkins, 2006c; Hopkins, 2015; Hopkins, 2017).
- Researchers who have difficulty recruiting enough **subjects of one sex** sometimes recruit a small proportion of the other sex and analyze the outcome without regard to sex. This approach is misguided. If you do not adjust for sex, you bias the mean effect towards that of the larger group. But to adjust for sex, you 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 30% of the subjects are in the smaller group. Download a simple spreadsheet I devised to illustrate this point. Conclusion: use subjects of one sex only, or aim for proportions of females and males in the sample that come close to their proportions in the population. This conclusion applies to other subgroupings.
- 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.
- 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% confidence 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.

- 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. Sample size in principle should be increased to take account of autocorrelation between repeated observations, but it is reasonable to assume that the model in the analysis removes most of the autocorrelation from the residuals and therefore that the sample size need not be increased substantially. 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.
- **Measurement studies**, which characterize validity and reliability of any measures and factor structure of psychometric inventories, are not included in available software for estimating sample size. Sample size for such studies shows a dependence on magnitude similar to that for the other designs. Very high reliability or validity (observed error << smallest important effect) can be characterized with as few as 10 subjects, because the upper confidence 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 confidence 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 usually need hundreds of subjects, because the alpha reliability of the factors is usually modest.
- Sample size for a **reliability pilot study** aimed at determining error of measurement for estimation of sample size in a repeated-measures main study. Sample size in the main study is inversely proportional to the square of the error of measurement. It follows that uncertainty in the error of measurement estimated in the pilot study is magnified into uncertainty in the sample size needed for the main study. For example, to limit the uncertainty in the estimate of sample size in a repeated-measures study to no more than ±20% (or a ×/÷ factor of 1.20), the uncertainty in the estimate of error has to be ±9.5% (×/÷√1.20). If "uncertainty" is 90% confidence limits, the spreadsheet for confidence limits (Hopkins, 2007) shows that the sample size for the reliability study has to be 174, which is unrealistically high. The smaller sample sizes of <50 that researchers often use in reliability studies is justifiable only if the resulting estimate of sample size in the main study turns out to be ~10-20, because the uncertainty in the estimate of such small sample sizes (e.g., ×/÷1.70 if the pilot study had 20 subjects) can be accommodated by increasing the sample size in the main study by ~5-10 subjects.
- Use of **simulation** 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 confidence 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.

In conclusion, it is important to point out that the approaches to sample-size estimation described here provide estimates based on inferences about a population mean effect. When the effect is an intervention, the outcome for an individual receiving the intervention will be different from the mean effect and will depend on individual responses to the intervention. To calculate chances of benefit and harm for the individual, we therefore need a sample size that characterizes individual responses adequately. As yet there is no spreadsheet and, as far as I know, no published formulae for this purpose.

I have created a slideshow to summarize most of the above principles, which you can download in Powerpoint or PDF format. You should view the slideshow as a full-screen presentation, especially for those slides explaining the statistical basis of the traditional and new approaches. The spreadsheet itself has extensive comments.

References

Batterham AM, Hopkins WG (2005a). Making meaningful inferences about magnitudes. Sportscience 9, 6-13

Batterham AM, Hopkins WG (2005b). A decision tree for controlled trials. Sportscience 9, 33-39

Hopkins WG (2003). A spreadsheet for analysis of straightforward controlled trials. Sportscience 7, sportsci.org/jour/03/wghtrials.htm (4447 words)

Hopkins WG (2004). How to interpret changes in an athletic performance test. Sportscience 8, 1-7

Hopkins WG (2006a). Sample sizes for magnitude-based inferences about clinical, practical or mechanistic significance (Abstract 2746). Medicine & Science in Sports & Exercise 38, S528-S529

Hopkins WG (2006b). Magnitude matters. Sportscience 10, 58

Hopkins WG (2006c). Spreadsheets for analysis of controlled trials, with adjustment for a subject characteristic. Sportscience 10, 46-50

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

Hopkins WG (2010). Linear models and effect magnitudes for research, clinical and practical applications. Sportscience 14, 49-57

Hopkins WG (2015). Individual responses made easy. Sportscience 19, i

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

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

Joseph L, du Berger R, Belisle P (1997). Bayesian and mixed Bayesian/likelihood criteria for sample size determination. Statistics in Medicine 16, 769-781

Julious SA (2004). Tutorial in biostatistics: sample sizes for clinical trials with Normal data. Statistics in Medicine 23, 1921-1986

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

Taubes G (1995). Epidemiology faces its limits. Science 269, 164-169

Updated, reviewed and published Apr 2007.
Latest update Jan 2017.