A Spreadsheet for Deriving a Confidence Interval, Mechanistic Inference and Clinical Inference from a P Value

Will G Hopkins

Sport and Recreation, AUT University, Auckland 0627, New Zealand. Email. Reviewers: Stephen W Marshall, Departments of Epidemiology and Orthopedics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599; Weimo Zhu, Kinesiology & Community Health, University of Illinois at Urbana-Champaign, Urbana, IL 61801.

The null-hypothesis significance test based only on a p value can be a misleading approach to making an inference about the true (population or large-sample) value of an effect statistic. Inferences based directly on the uncertainty in the true magnitude of the statistic are more comprehensible and practical but are not provided by statistical packages. I present here a spreadsheet that uses the p value, the observed value of the effect and smallest substantial values for the effect to make two kinds of magnitude-based inference: non-clinical (or mechanistic) and clinical. For a non-clinical inference the spreadsheet shows the effect as unclear if the confidence interval, which represents uncertainty about the true value, overlaps values that are substantial in a positive and negative sense; the effect is otherwise characterized with a statement about the chance that it is trivial, positive or negative. For a clinical inference the effect is shown as unclear if its chance of benefit is at least promising but its risk of harm is unacceptable; the effect is otherwise characterized with a statement about the chance that it is trivial, beneficial or harmful. The spreadsheet allows the researcher to choose the level of confidence (default, 90%) for mechanistic inferences and the threshold chances of benefit (default, 25%) and harm (default, 0.5%) for clinical inferences. The spreadsheet can be used for the most common effect statistics: raw, percent and factor differences in means; ratios of rates, risks, odds or counts; correlations (sample size is required). Inferences about standard deviations are also provided. The calculations are based on the same assumption of a normal or t sampling distribution that underlies the calculation of the p value for these statistics. KEYWORDS: clinical decision, confidence limits, null-hypothesis test, practical importance, statistical significance.

Reprint pdf · Reprint doc · Spreadsheets: Bayesian (MBI) · Frequentist

Update Oct 2022. I have modified the Bayesian spreadsheet to allow estimation of chances of true magnitudes for a standard deviation (SD). The modifications might be useful for researchers wishing to assess the sampling uncertainty in the magnitude of a measurement error. This error is usually derived via the SD of change scores or the residual in a mixed model, so it can never be negative. That part of the spreadsheet (Panel 5) therefore provides chances that the SD is trivial and chances that it is substantial positive, but not substantial negative. Note that the smallest important and other magnitude thresholds for an SD are half those of differences or changes in means for the same variable and subjects (Smith & Hopkins, 2011).

I have also provided instructions on estimating chances when the SD comes from a random effect in a mixed model, which could be useful for researchers wishing to assess the sampling uncertainty in individual responses or in the SD representing heterogeneity (tau) in a meta-analysis In such cases, the sampling distribution of the variance is assumed normal, hence negative values of variance (and by convention, the SD) can occur and are meaningful, hence chances of substantial negative variance (and negative SD, or factor SD <1.00) can be estimated (Panels 1 and 2). However, the calculations require a p value derived from the variance's standard error, which some stats packages do not provide (e.g., R).

Finally I have clarified some instructions and comments in the spreadsheet, and I restored the word *inference* in places, along with *decisions*, because I think use of the term *magnitude-based inference* (MBI) is acceptable again, since publication of the article on sampling uncertainty in *Frontiers in Physiology*.

Update Jan 2020. Forthcoming publications will show that the decision process in magnitude-based inference is equivalent to several frequentist interval hypothesis tests. For consistency with those tests, the Bayesian terms in the spreadsheet describing the probability that the true effect has a substantial or trivial magnitude (possibly, likely, very likely, and most likely) have been replaced in a frequentist version of the spreadsheet with frequentist terms describing compatibility of the data and model with those magnitudes (ambiguously, weakly, moderately, and strongly compatible). The word "chances" in this article can also be interpreted as "level of compatibility". The original spreadsheet is for researchers who prefer the Bayesian interpretation and who opt for a prior belief sufficiently vague or weakly informative to make no practical difference to the compatibility interval and decision. To incorporate a more informative prior, first read this article and then use the Bayes tab in this spreadsheet.

Update March 2018. I have removed the paragraph describing what I thought was an efficient method for constraining the error rates with multiple inferences. The Bonferroni method is now incorporated into the spreadsheet, with an accompanying explanatory comment that includes a suggestion below for reporting multiple inferences.

One of the first resources I provided for exercise scientists at A New View of Statistics 10 years ago was a spreadsheet to convert a p value (which all stats packages provide) into a confidence interval (which many didn't at that time). Five years ago I updated the spreadsheet to estimate something that no stats packages provide directly: the chances that the true value of the statistic is trivial or substantial in some positive and negative sense, such as beneficial and harmful. In this article I present an update that uses these chances to generate magnitude-based inferences. The article will serve as a peer-reviewed reference for citing the spreadsheet or the two kinds of magnitude-based in-

ference described herein: *mechanistic* and *clinical*. The article can also be regarded as part of a developing trend towards acknowledgement of the importance of magnitude in making inferences. Many journals now instruct authors to report and interpret effect sizes, but there is a need for instructions on how to incorporate sampling uncertainty into the interpretation. This article and the accompanying spreadsheet address that need.

The spreadsheet is aimed at helping researchers focus on precision of estimation of the magnitude of an effect statistic instead of the misleading null-hypothesis test based on a p value, when using data from a sample to make an inference about the true, population or infinite-sample value of the effect. The confidence interval represents uncertainty in the estimate of the true value of the statistic–in plain language, how big or small the true effect could be. Alan Batterham and I have already presented an intuitively appealing vaguely Bayesian approach to using the confidence interval to make what we call magnitude-based inferences (Batterham & Hopkins, 2005, 2006): if the true value could be substantial in both a positive and negative sense, the effect is unclear; otherwise it is clear and is deemed to have the magnitude of the observed value, preferably qualified with a probabilistic term (*possibly trivial, very likely positive,* and so on). We chose a default level of 90% for the confidence interval, which is consistent with an unclear effect having >5% chance of being positive and >5% chance of being negative. This approach is now included in the spreadsheet as a mechanistic inference. When an effect is unclear, the spreadsheet instructs the user to get more data. The spreadsheet allows the user to choose levels for the confidence interval other than 90% and to set values for chances defining the qualitative probabilistic terms. The qualitative terms and the default values are: *most unlikely, <0.5%; very unlikely, 0.5-5%; unlikely, 5-25%; possibly, 25-75%; likely, 75-95%; very likely, 95-99.5%;* and *most likely*, >99.5%.

In our article about magnitude-based inferences, Batterham and I did not distinguish between inferences about the clinical or practical vs the mechanistic importance of an effect. I subsequently realized that there is an important difference, after publishing an article last year on two new methods of sample-size estimation

(Hopkins, 2006a). The first new method, based on an acceptably narrow width of the confidence interval, gives a sample size that is appropriate for the mechanistic inference described above. The other method, based on acceptably low rates of making what I described as Type 1 and Type 2 clinical errors, can give a different sample size, appropriate for a decision to use or not to use an effect; such a decision defines a clinical (or practical) inference.

The meaning and wording of an inference about clinical utility differ from those of a mechanistic inference. It is in the nature of decisions about the clinical application of effects that the chance of using a harmful effect (a Type 1 clinical error) has to be a lot less than the chance of not using a beneficial effect (a Type 2 clinical error), no matter how small these chances might be. For example, if the chance of harm was 2%, the chance of benefit would have to be much more than 2% before you would consider using a treatment, if you would use it at all. I have opted for default thresholds of 0.5% for harm (the boundary between most unlikely and very unlikely) and 25% for benefit (the boundary between unlikely and possibly), partly because these give a sample size about the same as that for an acceptably narrow 90% confidence interval. An effect is therefore clinically unclear with these thresholds if the chance of benefit is >25% and the chance of harm is >0.5%; that is, if the chance of benefit is at least promising but the risk of harm is unacceptable. The effect is otherwise clinically clear: beneficial if the chance of benefit is >25%, and trivial or harmful for other outcomes, depending on the observed value. The spreadsheet instructs the user whether or not to use the effect and, for an unclear effect, to get more data. Thresholds other than 0.5% and 25% can also be chosen.

I invite you to explore the differences between statistical, mechanistic and clinical inferences for an effect by inserting various p values, observed values and threshold important values for the effect into the spreadsheet. Use the kind of effect you are most familiar with, so you can judge the sense of the inferences. You will find that statistically significant and non-significant are often not the same as mechanistically or clinically clear and unclear. You will also find that a mechanistic and a clinical infer-

ence for the same data will sometimes appear to part company, even when they are both clear; for example, an effect with a chance of benefit of 30% and chance of harm of 0.3% is mechanistically possibly trivial but clinically possibly beneficial. With a suboptimal sample size an effect can be mechanistically unclear but clinically clear or vice versa. These differences are an inevitable consequence of the fact that thresholds for substantially positive and negative effects are of equal importance from a mechanistic perspective but unequal when one is a threshold for benefit and the other is a threshold for harm. To report inferences in a publication, I suggest we show 90% confidence intervals and the mechanistic inference for all effects but indicate also the clinical inference for those effects that have a direct application to health or performance.

With its unequal values for clinical Type 1 and Type 2 errors, a clinical inference is superficially similar to a statistical inference based on statistical Type I and II errors. The main difference is that a clinical inference uses thresholds for benefit and harm, whereas a statistical inference uses the null rather than the threshold for harm. Which is the more appropriate approach for making decisions about using effects with patients and clients? I have no doubt that a study of a clinically or practically important effect should be designed and analyzed with the chance of harm up front. Use of the null entails sample sizes that, in my view, are too large and decisions that are therefore too conservative. For example, it is easy to show with my spreadsheet for sample-size estimation that a statistically significant effect in a study designed with the usual default Type I and II statistical errors of 5% and 20% has a risk of harm of less than one in a million, and usually much less. Thus there will be too many occasions when a clinically beneficial effect ends up not being used because it is not statistically significant. Statistical significance becomes less conservative with suboptimal sample sizes: for example, a change in the mean of 1 unit with a threshold for benefit of 0.2 units is a moderate effect using a modified Cohen scale (Hopkins, 2006c), but if this effect was only just significant ($p = 0.04$) because of a small sample size, the risk of harm would be 0.8%. Supraoptimal sample sizes can produce a different kind of problem: statistically significant effects that are

likely to be clinically useless. Basing clinical decisions directly on chances of benefit and harm avoids these inconsistencies with clinical decisions based on statistical significance, although there is bound to be disagreement about the threshold chances of benefit and harm for making clinical decisions.

Depending on the clinical situation, some researchers may consider that 0.5% for the risk of harm is not conservative enough. I ask them to consider that, in other situations, 0.5% may be *too* conservative. For example, an athlete would probably run a 2% risk of harm for a strategy with an 85% chance of benefit, which would be the outcome in a study with a suboptimal sample size that produced a p value of 0.12 for an observed enhancement in performance of 3.0 units (e.g., power output in percent), when the smallest important threshold is 1.0 unit. This example demonstrates that the threshold for an acceptable risk of harm may need to move with the chance of benefit, perhaps by keeping a constant ratio for odds of benefit to harm. Table 1 shows chances that all have approximately the same odds ratio (~50) and that could represent thresholds for the decision to use an effect in studies with sample sizes that turn out to be suboptimal or supraoptimal. The highest thresholds in the table (>75% for benefit and <5% for harm) are consistent with the common-sense decision to use an effect that is likely to be beneficial, provided it is very unlikely to be harmful.

Table 1. Combinations of threshold chances of benefit and harm with the same approximate odds ratio (~50) for decisions to use effects when sample size is suboptimal, optimal and supraoptimal.		
Chances of		
benefit	harm	Sample size
>75	<5	suboptimal
>50	<1.5	
>25	<0.5	optimal
>10	<0.2	supraoptimal
>5	<0.1	

Mechanistic and clinical inferences evidently require the researcher to find answers to sometimes difficult questions. Greg Atkinson has called attention to some of these questions in an article in the current issue of this journal (Atkinson, 2007). Is it appropriate to base a mechanistic decision on the way in which a symmetrical confidence interval overlaps substantially positive and negative values? Is it appropriate to base a clinical decision on chances of benefit and harm? What is the appropriate default level for a confidence interval? What are appropriate thresholds for chances of benefit and harm for a clinical decision? What are appropriate threshold values of benefit and harm for the effect statistic? I have attempted to answer these questions by reading and by reflecting on experience. Another difficult issue neither Greg nor I have addressed is the dollar value of benefit, the dollar cost of harm, and the dollar cost of using an effect, all of which need to be factored somehow into a clinical decision. For me, defaulting to an inference based on a null-hypothesis test in the face of these difficulties is not an option. In any case, using an hypothesis test to make a clinical decision requires answers to the same or similar difficult questions about thresholds for magnitude, thresholds for Type I and II errors, dollar value and dollar costs.

Researchers who champion statistical significance should be wary of any sense of security in the conservatism of their inferences. Confounding and other biases arising from poor design or experimentation can make a harmful effect appear beneficial with a vanishingly small p value or chance of harm. Even when there are no biases, all the inferences described in this article relate only to a population mean effect, not to effects on individuals. An effect could therefore be beneficial on average but harmful to a substantial proportion of subjects, yet the p value and chance of harm will again be vanishingly small with a large enough sample. Larger representative samples are needed to adequately characterize such individual differences or responses, and also to quantify any harmful (or beneficial) side effects of experimental treatments. It seems to me that researchers, reviewers and editors should be less concerned about getting p<0.05 and more concerned about reducing biases, characterizing individual responses and quantifying side effects.

The spreadsheet now includes a Bonferroni type of adjustment for the inflation of the chance of making at least one error when you make inferences about more than one effect. Although I agree with others (e.g., Perneger,

1998) who have advised against reducing the p value for declaring statistical significance with each inference, I think it is appropriate to highlight effects that are still clear with the Bonferroni adjustment. For example, with five independent effects, show 90% confidence intervals, but highlight non-clinical effects that would still be clear at the 98% level (100-10/5=98), and clinical effects that would still be clear with <0.1% risk of harm (0.5/5=0.10). You can highlight such effects in a table by showing them in **bold**, with an explanatory footnote.

You should be aware of the upward bias in magnitude that occurs when choosing the largest of several effects with overlapping confidence intervals. The accompanying supplementary spreadsheet uses simulation to show that, for two effects of about equal magnitude, the bias is ~0.5 of the average standard error, or about one-sixth the average 90% confidence interval. Bias increases with more effects but decreases as true differences between the effects increase.

Finally, some technical issues… I devised the formulae for confidence intervals and chances of benefit and harm using the same statistical first principles that underlie the calculation of p values. The effect statistic or its appropriate transformation is assumed to have either a t sampling distribution (raw, percent or factor differences between means) or a normal sampling distribution (rate, risk or odds ratios and correlation coefficients). The central-limit theorem practically guarantees that this assumption is not violated for all but the smallest sample sizes of the most non-normally distributed data. The log transformation is used for factor effects and ratios. Percent effects need to be converted to factor effects, as explained in the spreadsheet. The Fisher (1921) z transformation is used for correlations, and a sample size rather than a p value is required. There are also panels for calculating the confidence interval for a standard deviation using the chi-squared distribution and for comparing two standard deviations using the F distribution.

In the unlikely event that you want to make a clinical inference for a difference in means or a ratio for which you have confidence limits but no p value, download the spreadsheet for combining independent groups. Insert the values of the effect, the confidence limits and the small-est important effect into the sheet for >2 groups, with a weighting factor of 1 for the effect (Hopkins, 2006b).

Acknowledgment: Patria Hume and Darrell Bonetti contributed to the idea of using higher thresholds for risk of harm with higher chances of benefit.

References

Atkinson GA. (2007). What's behind the numbers? Important decisions in judging practical significance. Sportscience 11, 12-15. http://sportsci.org/2007/ga.htm

Batterham AM, Hopkins WG. (2005). Making meaningful inferences about magnitudes. Sportscience 9, 6-13. http://sportsci.org/jour/05/ambwgh.htm

Batterham AM, Hopkins WG. (2006). Making meaningful inferences about magnitudes. International Journal of Sports Physiology and Performance 1, 50-57.

Fisher RA. (1921). On the probable error of a coefficient of correlation deduced from a small sample. Metron 1, 3-32.

Hopkins WG. (2006a). Estimating sample size for magnitude-based inferences. Sportscience 10, 63-67. http://sportsci.org/2006/wghss.htm

Hopkins WG. (2006b). A spreadsheet for combining outcomes from several subject groups. Sportscience 10, 51-53. http://sportsci.org/2006/wghcom.htm

Hopkins WG. (2006c). Spreadsheets for analysis of controlled trials, with adjustment for a subject characteristic. Sportscience 10, 46-50. http://sportsci.org/2006/wghcontrial.htm

Perneger TV. (1998). What's wrong with Bonferroni adjustments. BMJ 316, 1236-1238.

Smith TB, Hopkins WG. (2011). Variability and predictability of finals times of elite rowers. Medicine and Science in Sports and Exercise 43, 2155-2160.

Response to Reviewer's Comments

Weimo Zhu made the following comment: "Overall, you have addressed a very important topic and have provided the field with a possible solution to the problem." He also made numerous useful suggestions I have complied with.

Weimo wanted me to remove "P value" from the title, because I use more than just the P

value to derive inferences. This suggestion is reasonable, but the title is nevertheless correct as it stands, and I think there is a need for emphasis on the P value. I have added extra words in the Abstract to make it clear that the observed value and smallest substantial values of the effect are also needed. He also suggested that I should cite and acknowledge the works by "early pioneers" [for the vaguely Bayesian interpretation of the confidence interval]. Alan Batterham and I could find no earlier reference for the way in which we use the confidence interval to declare an effect clear or unclear. Alan is usually very thorough with literature searches.