SPORTSCIENCE sportsci.org
Perspectives / Research Resources

A Spreadsheet for Combining Outcomes from Several Subject Groups

Will G Hopkins

Sportscience 10, 51-53, 2006 (sportsci.org/2006/wghcom.htm)
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; Glen E Fincher, Sport Sciences, Ashland University, Ashland, Ohio 44805.

Data analysis that fails to account for independent groups defined by a subject characteristic (e.g., sex) or by a design characteristic (e.g., treatment order) can result in bias, confounding, and loss of precision in the outcome. Combining the outcomes from separate analyses of the groups is a robust approach to the problem that is easily achieved with the spreadsheet presented here. Differences in the outcome between groups represent the effect of the characteristic on the outcome, while the mean of the outcomes represents the outcome adjusted appropriately for the characteristic. The spreadsheet calculates confidence limits for the differences and for the mean from the confidence limits for the outcome in each group. It also presents magnitude-based inferences for the differences and mean. There are separate cells in the spreadsheet for outcomes represented by means or other normally distributed statistics, relative rates (risk, odds and hazard ratios) or other log-normally distributed statistics, and correlation coefficients. KEYWORDS: confidence limits, confounding, covariate, inference, modeling.

Reprint pdf · Reprint doc · Spreadsheet · Commentary by Glen Fincher

Update May 2009. Corrected an error in the estimation of chances of benefit and harm when combining correlation coefficients as differences (2 groups) and as custom effects (>2 groups).

Update Oct 2008. Understanding the difference between fixed effects and random effects will help you decide when you can use this spreadsheet and what to do when you can't use it. See the note at the end of this article.

I have also added correlation coefficients to the spreadsheet for >2 groups. Analysis of correlations requires the Fisher transformation, and as with all non-linear transformations other than the log, estimation of the magnitude of the back-transformed effect requires specification of some reference value. A cell is included for this purpose.

Update Oct 2007. The spreadsheet now includes customizable clinical and mechanistic inferences (Hopkins, 2007).

A subject characteristic such as sex or a design characteristic such as order of treatments in a crossover divides a sample into two or more groups. When you study the effect of something on the subjects–either as an intervention in an experimental design or as a relationship between measures such as health, activity and performance in a non-experimental design–the magnitude of the effect will always differ to some extent between the groups. Failure to account for the groups in the analysis may therefore lead you to the wrong conclusion about the effect.

The wrong conclusion can be the result of bias, confounding, and imprecision. *Bias* arises when the proportion of subjects in the groups is not the same as the proportion in the population. For example, if you have twice as many males as females in the sample, the mean without regard to sex will be biased towards males. *Confounding* occurs when the proportion of subjects in each group differs for different values of another characteristic in the analysis. For example, if you analyze for the effect of age without taking sex into account, and there are more males amongst the older subjects, the value of the effect you get for age will be partly an effect of sex; the effect of age is then said to be confounded by sex. Finally the difference in the magnitude of the effect between the groups leads to *imprecision*, because the difference turns up as unexplained residual error that

makes a wider confidence interval for the effect. For example, if the difference in an effect between males and females is twice the magnitude of the error within either group, the residual error in an analysis without regard to sex will be about 40% greater than with sex in the analysis, and the confidence interval will be correspondingly wider. (I derived this estimate from first principles and checked it using simulation in another spreadsheet.)

The spreadsheet that accompanies this article provides a method of analysis that avoids these problems with groups. The linear modeling procedures in statistical packages provide another method, but statistical packages have their own problems: they are hard to use, the best are expensive, and all are a long way from presenting outcomes in a clinically or practically meaningful way. Furthermore, analysis of variance–the standard approach to including a subject characteristic representing groups–accounts for different means between the groups, but it does not allow for different standard deviations in the groups. Analysis of variance therefore leads to the wrong inferences when there are unequal numbers of subjects in the groups, because it effectively uses the equal-variances form of the t statistic instead of the unequal-variances form to derive confidence limits or p values. The same problem applies to groups in repeated-measures ANOVA in the not infrequent situation of different errors of measurement in the groups. (These assertions are based on simulations I documented in a conference presentation.) Mixed modeling solves this particular problem by allowing estimation of various standard deviations, but mixed modeling is beyond the reach of most researchers; hence this article and spreadsheet.

The spreadsheet works by combining outcomes from separate analyses for each group. Differences in the outcome between groups represent the effect of the grouping variable on the outcome. The mean of the outcomes across the groups represents the outcome adjusted appropriately for the effects of the grouping. The spreadsheet calculates confidence limits for differences and for the mean using the confidence limits for the outcome in each group. If you also enter a value for the smallest clinically or practically important effect, the spreadsheet presents magnitude-based inferences for the differences and mean using the approach of Batterham and Hopkins (2005). There is also a cell showing a summary qualitative outcome, as described in the article on controlled trials in this issue (Hopkins, 2006).

There are separate grids of cells in the spreadsheet for various kinds of outcome statistic: means or other statistics with sampling distributions of the t statistic; percent and factor effects with sampling distributions of the t statistic after log transformation; relative rates (risk, odds and hazard ratios) or other log-normally distributed statistics; and correlation coefficients, whose sampling distribution is normal after z transformation (Fisher, 1921). In the case of t-distributed statistics, I used the Satterthwaite (1946) approximation to calculate the degrees of freedom for the combined error variances derived from the confidence limits for each group.

The approach of combining separate analyses in this manner is robust to the effect of any differences in error between the groups, because a different error is automatically generated in each analysis. This approach also effectively accounts for all interactions between the grouping variable and other predictors, so you do not have to worry about how to deal with the interaction terms in a full analysis. However, you must be careful about the way you adjust for any covariates in the separate analyses. For example, if age is a covariate, you should use the grand mean age to adjust the outcome in the separate analyses for females and males, if you want to control for age in the subsequent comparison or averaging of the outcome for females and males.

To use the spreadsheet, you first perform the analysis of interest separately for each group with either a statistical package or another spreadsheet. The spreadsheets I describe in an article in the current issue of this journal are suitable for the various kinds of controlled trial (Hopkins, 2006). For each group you then enter the values of the outcome and its confidence limits. (If your stats package provides only p values, generate confidence limits from them using another spreadsheet at newstats.org.) If the outcome is a difference in means, you also enter its associated error degrees of freedom.

Percent differences in means have to be converted to factor differences and analyzed as

such when any of the percents or their combined values are greater than 10%. The factor corresponding to x% is 1+x/100; for example, 7%→1+7/100=1.07, 23%→1.23, 165%→2.65, and -23%→1-23/100=0.77. Convert the derived factor and confidence limits back into percents if they are less than about 1.5 (50%) or greater than about 0.5 (-50%), but otherwise report them as factors.

When you have only two groups (e.g., females and males), use the first sheet in the spreadsheet. Enter outcomes for more than two groups in the second sheet, which allows you to combine the groups in various ways (e.g., the difference between an outcome in one sport and the mean of the outcome in two others). These "custom" combinations are derived via weighting factors that you insert for each group and that must add to exactly zero or one. Examples are explained in the spreadsheet. An error message appears when the weights are invalid.

The spreadsheet should get frequent use to adjust for order effects in a crossover. Proper design of a crossover involves a Latin square to define the smallest number of groups of subjects such that all subjects in a given group get the treatments in the same sequence, and overall every treatment follows every other treatment the same number of times. To adjust for any order effect, perform separate analyses for the treatment effects in each group, then average across all groups using the spreadsheet. This adjustment is especially important if there is a substantial order effect and the number of subjects is not the same in all groups. With only two groups, the magnitude of the order effect is half the difference in the outcome between the groups (use weights of 0.5 and -0.5), but you will have to figure out the appropriate sign. With more than two groups the algebra required to work out the order effects is too complex.

It is important to understand that the confidence limits for each group are combined by assuming the groups are independent–that is, by assuming the random variation you would get with repeated sampling in one group has no correlation with that of any other group. Groups represented by different subjects generally meet this requirement, but an important exception is controlled trials in which randomization to the groups has been balanced on the pre-test value of the dependent variable. For such data, the outcomes are not independent unless you adjust for the pre-test value by including it as a covariate in the analysis. Without such adjustment, the confidence interval provided by the spreadsheet for differences between groups will usually be too wide and the confidence interval for the mean of the groups will usually be too narrow. In the absence of any other information, there is no way to combine non-independent outcomes to avoid this problem. Your only option is to analyze the original data using repeated-measures modeling of some kind.

One of the reviewers raised the question of whether there is any loss of precision when combining separate analyses compared with a single full analysis. We put it to the test for a binary outcome variable with an assumed binomial distribution, and we got identical confidence limits by the two methods. Separate analyses of a normally distributed variable with only a few subjects in each group could lead to loss of precision compared with the full analysis, but only if you made the unjustifiable assumption in the full analysis that the groups had the same error. For example, if you had five subjects in each group, the degrees of freedom for the error in the full analysis under the assumption of equal errors would be 9, whereas in the combined separate analyses the degrees of freedom would be at most 8. It is easy to show with the t statistic that the 90% confidence interval in the case of 8 degrees of freedom is only 6% wider than with 9 degrees of freedom. If the errors were different, there would be less than 8 degrees of freedom (as given by the Satterthwaite approximation), and the confidence interval would be substantially wider. But if the errors *were* different, you *must* use different errors in the full analysis, and such an analysis gives an outcome identical to that via combining separate analyses.

If your study is a Latin-squares balanced crossover of, say, four treatments, and you have only two subjects on each of the four treatment sequences, there will be a serious loss of precision unless you make the reasonable assumption of equal error for a given comparison of treatments in each of the four groups. Use the crossover spreadsheet for the analysis, which is based on this assumption, or if you can access mixed modeling, use it to adjust for the order effect and model some extra error for treatments and/or trials where there are substantial

changes in the mean.

Update: Analysis of Fixed vs Random Effects

The spreadsheet is designed to compare or combine the levels of a **fixed** effect. Examples of fixed effects (*and their levels*) are sex (*male, female*), a specific treatment effect (*caffeine, placebo*), and identified sports (*American football, rugby union, rugby league*). For the spreadsheet to give correct answers, values in each level must also be independent of those in the other levels, which will usually be the case if the subjects in each level are different from the subjects in the other levels.

Do not use this spreadsheet to combine values of levels of a **random** effect. The levels of a random effect represent a sample drawn randomly from some population, such as all possible team sports, rather than specific identified sports that would be the same if you drew another sample. An even simpler example of a random effect is the identity of the subjects in any of the usual sample-based studies.

A good check on whether an effect is fixed or random is to consider the various ways you could combine the levels. If the mean and the standard deviation of the values are the only sensible combinations you can think of, chances are the effect is random. If you can imagine combining the levels in other ways, as shown in the examples in the spreadsheet, and if a standard deviation of the values of the levels doesn't quite make sense, chances are the effect is fixed.

The confidence interval for the mean of the levels of a random effect has to take into account the uncertainty arising from the fact that a different sample would give different levels of the random effect, each with different values. You can generate the confidence interval for the mean with the CONFIDENCE function in Excel, which uses the formula you learn about in Stats 101 classes: an appropriate value of the t statistic times the standard deviation divided by the square root of the sample size.

Often there is an approach to the analysis of data that takes away the worry about whether an effect is fixed or random. For example, when you compare the means of two groups of subjects with an unpaired t statistic, the identity of the subjects is a random effect and the identity of the groups is a fixed effect, but who cares? In other situations a good working knowledge of fixed and random effects–and skill with an advanced stats package–allow you to take into account non-uniformity of error and multiple levels of repeated measurement or other clustering of observations in complex data. Appropriate use of random effects also allows you to estimate individual responses to a treatment as a standard deviation, although my controlled-trial spreadsheets provide such estimates correctly in straightforward designs. For more information on fixed and random effects, see the slideshow on repeated measures.

This update was reviewed by Stephen Marshall.

References

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

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

Hopkins WG (2006). 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

Satterthwaite FW (1946). An approximate distribution of estimates of variance components. Biometrics Bulletin 2, 110-114

Published Dec 2006