

A Spreadsheet for Analysis of Straightforward Controlled Trials

Will G Hopkins

Sportscience 7, sportsci.org/jour/03/wghtrials.htm, 2003 (4447 words)

Will G Hopkins, Sport and Recreation, Auckland University of Technology, Auckland 1020, New Zealand. [Email](#).

Reviewer: Alan M Batterham, Department of Sport and Exercise Science, University of Bath, Bath BA2 7AY, UK.

Spreadsheets are a valuable resource for analysis of most kinds of data in sport and exercise science. Here I present a spreadsheet for comparison of change scores resulting from a treatment in an experimental and control group. Features of the spreadsheet include: the usual analysis based on the unequal-variances unpaired t statistic; analysis following logarithmic, percentile-rank, square-root, and arcsine-root transformations; plots of change scores to check for uniformity of the effects; back-transformation of the effects into meaningful magnitudes; estimates of reliability for the control group; estimates of individual responses; comparison of the groups in the pre-test; and estimates of uncertainty in all effects, expressed as confidence limits and chances the true value of the effect is important. Analysis of straightforward crossover trials based on the paired t statistic is provided in a modified version of the spreadsheet. KEYWORDS: analysis, crossover, design, intervention, psychometric, randomized, statistics, transformation, t test.

[Reprint pdf](#) · [Reprint doc](#) · [Slideshow](#) · [Reviewer's Comment](#)

Spreadsheets: [controlled trial](#) and [crossover](#)

[Update](#) on transformations, 7 Nov 2003.

[Minor edits](#), 13 Nov 2003.

[Update](#) on comparison of pre-test values, 27 Nov 2003.

[Slideshow](#) uploaded, 27 Nov 2003.

Amongst the various kinds of [research design](#), a controlled trial is the best for investigating the effect of a treatment on performance, injury or health (Hopkins, 2000a). In a controlled trial, subjects in experimental and control treatment groups are measured at least once pre and at least once during or post their treatment. The essence of the analysis is a comparison between groups of an appropriate measure of change in a dependent variable representing performance, injury, or health. The final outcome statistic is thus a difference in a change: the difference between groups in their mean change due to the experimental and control treatments. Another form of controlled trial is the crossover, in which all subjects receive all control and experimental treatments, with sufficient time following each treatment to allow its effect to wash out. The final outcome statistic in a crossover is simply the mean change between treatments.

Calculating the value of the outcome statistic in controlled trials and crossovers is easy enough. More challenging is making inferences about its true or population value in terms of a p value, statistical significance, confidence limits, and clinical significance. The traditional approach is repeated-measures analysis of variance (ANOVA). In a [slideshow](#) I presented at a conference recently, I pointed out how this approach can lead to the wrong p value and therefore wrong conclusions about the effect of the treatment in controlled trials. I also explained how the more recent approach of mixed modeling not only gives the right answers but also permits complex analyses involving additional levels of repeated measurement in addition to covariates for subject characteristics. Mixed modeling is available only in advanced, expensive, and user-unfriendly statistical packages, such as the Statistical Analysis System (SAS), but straightforward analysis of

controlled trials either by repeated-measures ANOVA or mixed modeling is equivalent to a t test, which can be performed on an Excel spreadsheet.

In the last few years I have been advising research students to use a spreadsheet for their analyses whenever possible and to consult a statistician only when they need help with more complex analyses. Spreadsheets I have devised for various analyses ([reliability](#), [validity](#), [assessment of an individual](#), [confidence limits and clinical significance](#)) seem to have facilitated this process. Although it may be instructive for students to devise a spreadsheet from scratch, they probably reach a more advanced level more rapidly by using a spreadsheet that already works: anytime they want to learn more about the calculations, they need only click on the appropriate cells. Use of a pre-configured spreadsheet surely also reduces errors and saves more time (the student's and the statistician's). I have therefore devised a spreadsheet for analysis of straightforward controlled trials, and I have modified it for analysis of crossovers.

Controlled Trials

Features of the spreadsheet for controlled trials include the following...

- **The usual analysis of the raw values** of the dependent variable, based on the unequal-variances unpaired t statistic.

The data in the spreadsheet are for one pre- and two post-treatment trials, and the effects are the differences in the three pairwise changes between the trials (post1-pre, post2-pre, and post2-post1). You can easily add more trials and effects, including parameters for line or curve fits to each subject's trials—what I call [within-subject modeling](#) (Hopkins, 2003a). As for the role of the unequal-variances t statistic, see the section on [uniformity of residuals](#) at my stats site (Hopkins, 2003b) and also the [slideshow](#) I referred to earlier. In short, never use the equal-variances t test, because the variances are never equal. (The variances in question are the squares of the standard deviations of the change scores for each group.)

With three or more groups (for example, a control and two treatment groups), you will have to use a whole new spreadsheet for each pairwise comparison of interest. Enter the data for two groups, save the spreadsheet, resave it with a new name, then replace one group's data with those of the third group. Save, resave, then copy and paste data and so on until you have all the pairwise group comparisons.

The spreadsheet does not provide any adjustment for so-called [inflation of Type 1 error](#) with multiple group comparisons or with the multiple comparisons between more than two trials (Hopkins, 2000b). These adjustments (Tukey, Sidak, and so on) probably don't work for repeated measures, and in any case they are nonsense, for various reasons that I detail at my stats site and in that [slideshow](#). The main reason is that the procedure, which involves doing pairwise tests with a conservatively adjusted level of significance only if the interaction term is significant, dilutes the power of the study for the most important comparison (for example, the last pre with the first post for the most important experimental vs control group). I don't believe in testing for significance anyway, but even if I did, I would be entitled to apply a t test to the most important pre-planned comparison without looking at the interaction and without adjustment of significance for multiple comparisons. Some of the best researchers in our field fail to understand this important point.

- **Analysis of various transformed values** of the dependent variable, to deal with any systematic effect of an individual's pre-test value on the change due to the treatment.

For example, if the effect of the treatment is to enhance performance by a few percent regardless of a subject's pre-test value, analysis of the raw data will give the wrong answer for most subjects. For these and most other performance and physiological

variables, analysis of the logarithm of the raw values gives the right answer. Along with [logarithmic transformation](#), the spreadsheet has [square-root transformation](#) for counts of injuries or events, [arcsine-root transformation](#) for proportions, and [percentile-rank transformation](#) (equivalent to non-parametric analysis) when an appropriate formula for a transformation function is unclear or unspecified (Hopkins, 2003c and other pages).

A dependent variable with a grossly non-normal distribution of values and some zeros thrown in for good measure is a good candidate for rank transformation. An example is time spent in vigorous physical activity by city dwellers: the variable would respond well to log transformation, were it not for the zeros.

Dependent variables with only two values (example: injured yes/no) and [Likert-scale variables](#) with any number of points (example: disagree/uncertain/agree) can be coded as integers and analyzed directly without transformation (Hopkins, 2003b). I now code two-value variables and 2-point scales as 0 or 100, because differences or changes in the mean then directly represent differences or changes in the percent of subjects who, for example, got injured or who gave one of the two responses. Advanced approaches to such data involve repeated-measures logistic regression, but the outcomes are odds ratios or relative risks, which are hard to interpret.

If you have a variable with a lower and upper bound and values that come close to either bound, consider converting the values so that they range from 0 to 100 ("percent of full-scale deflection"), then applying the arcsine-root transformation. Composite psychometric scores derived from multiple Likert scales should behave well under this transformation, especially when a substantial proportion of subjects respond close to the minimum or maximum values.

- **Plots of change scores** of raw and transformed data against pre-test values, to check for outliers and to confirm that the chosen transformation results in a similar magnitude of change across the range of pre-test values.

These plots achieve the same purpose as plots of [residual vs predicted](#) values in more powerful statistical packages. Statisticians justify examination of such plots by referring to the need to avoid heteroscedasticity (non-uniformity of error) in the analysis, which for a controlled trial means the same thing as aiming for uniformity in the effect of the treatment.

Sometimes it's hard to tell which transformation gives the most uniform effect in the plots. Indeed, when there is little variation in the pre-test values between subjects, all transformations give uniform effects and the same value for the mean effect after back transformation. Regardless, your choice of transformation should be guided by your understanding of how a wide variation in pre-test values would be likely to affect the effects.

After applying the appropriate transformation, you may sometimes still see a tendency for subjects with low pre-test values to have more positive change scores (or less negative change scores) than subjects with high pre-test values. This tendency may be a genuine effect of the pre-test value, but it may also be partly or wholly an artefact of [regression to the mean](#) (Hopkins, 2003d). You address this issue by performing an analysis of variance or general linear model with the change score as the dependent variable, group identity as a nominal predictor, and the pre-test value as a numeric predictor or covariate interacted with group. The difference in the slope of the covariate between the two groups is the real effect of pre-test value free of the artefact.

- **Various solutions to the problem of back-transformation** of treatment effects into meaningful magnitudes.

Log-transformation gives rise to percent effects after back transformation. If the percent effects or errors are large (~100% or more, as occurs with some hormones and

assays for gene expression), it is better to back-transform log effects into factors. For example, an increase of 250% is better expressed as a factor of 3.5.

Another approach, which as far as I know is novel, is to estimate the value of the effect at a chosen value of the raw variable. I have included this approach for back-transformation from percentile-rank, square-root, and arcsine-root transformations. See the spreadsheet to better understand what I mean here. Note that I have not included this approach with log transformation, because percent and factor effects are better ways to back transform log effects.

Finally, I have also expressed magnitudes of effects for the raw variable and for all transformations as [Cohen effect sizes](#): the difference in the changes in the mean as a fraction or multiple of the pre-test between-subject standard deviation. You should interpret the magnitude of the Cohen effect sizes using my scale: <0.2 is trivial, 0.2-0.5 is small, 0.6-1.1 is moderate, 1.2-1.9 is large, and 2.0 or more is very large (Hopkins, 2002a). Use Cohen effect sizes for effects that relate to population health, for effects derived from physiological, biomechanical, or other mechanism variables, and for some measures of performance that have no direct association with competitive performance scores (for example, field tests for team-sport athletes). Do NOT use Cohen effect sizes for direct measures of competitive athletic performance or performance tests directly related thereto; instead express the effects as percents or as raw units.

- **Estimates of reliability** in the control group, expressed as typical error of measurement and change in the mean.

The design and analysis of the control group on its own amount to a [reliability](#) study. You can therefore compare measures of reliability derived from the control group with those of published reliability studies or your own pilot reliability study, if you did one. The most important measure of reliability is the typical error: the typical variation (standard deviation) a subject would show in many repeated trials (Hopkins, 2000c and my stats pages on [reliability](#)). The typical error shown in the spreadsheet is simply the standard deviation of the change score in the control group divided by $\sqrt{2}$. If your error differs substantially from that of previous studies, you should try and explain why in your write-up. Note that one obvious explanation for a worse error in your study is a longer time between trials compared with that in reliability studies.

The change in the mean from trial to trial is another measure of reliability. A substantial change usually indicates that the subjects (or the researcher) showed a practice, learning, or other familiarization effect between the trials. Again, you will need to explain any such change in your study. A substantial change in the mean can also help explain a worse typical error than in previous studies: there are usually individual responses to the change (some subjects show more familiarization), and individual responses produce more variation in change scores.

- **Estimates of individual responses** to the treatment, expressed as a standard deviation for the mean effect, in the various forms of the transformed and back-transformed variable.

If subjects differ in their response to the experimental treatment after any appropriate transformation, the standard deviation of the change scores in the experimental group will be greater than that in the control group. It is easy to show from basic statistical theory that the variation in the response between individuals, expressed as a standard deviation, is the square root of the difference in the variances (square of the standard deviations) of the change scores. This estimate of variation is free of error of measurement, although error of measurement obviously contributes to its uncertainty.

The standard deviation representing individual responses is the typical variation in the response to the treatment from individual to individual. For example, if the mean

response is 3.0 units and the standard deviation representing individual responses is 2.0 units, most individuals (about two-thirds) will have a true response somewhere in the region of 1 to 5 (3-2 to 3+2). You can state that the effect of the treatment is typically 3.0 ± 2.0 units (mean \pm standard deviation).

Sometimes the observed difference in the variances of the change scores is negative, either because the experimental treatment somehow reduces variation between subjects (for example, by bringing them up to a common ceiling or plateau), or more likely because sampling variation results in a negative difference purely by chance. I have devised the convention of showing the square root of a negative variance as a negative standard deviation. You should interpret such negative standard deviations as indicative of no individual responses.

If you find that there are substantial individual responses to a treatment, the next step is to find the subject characteristics that predict them. To do that properly, you have to include subject characteristics in the analysis as covariates, which is not possible with the spreadsheet. However, you can do a publication-worthy job by correlating or plotting the change scores with the subject characteristics that you think might predict the individual responses. See the [slideshow](#) on repeated measures for more information on individual responses, including a caution about misinterpreting a poor correlation between change scores.

- **A comparison of pre-test values** of means and standard deviations in the experimental and control groups for all transformations and back-transformations, to check on the balance of assignment of subjects to the groups.

The purpose of a control group is to provide an estimate of the change that occurs in the absence of the experimental treatment. This change is then subtracted from the change in the experimental group to give the pure effect of the experimental treatment. Fine, but suppose the pre-test score affects the change resulting from the experimental and control treatments (example: subjects with lower pre-test scores respond more to either or both treatments). Suppose also that the mean pre-test score differs substantially between the groups—that is, the groups are not balanced with respect to the pre-test score. Under these circumstances, part of the difference between the change scores will be due to the lack of balance, so the analysis will not provide a correct estimate of the experimental effect. It is therefore important to compare the two groups and comment on the difference in the mean pre-test score. The spreadsheet provides you with such a comparison. You can describe the magnitude of the difference between the groups either by noting whether the difference is greater than the smallest worthwhile change (see the last bullet point below) or by interpreting the Cohen effect size for the difference as trivial, small, and so on, when Cohen is appropriate. It is equally important to compare the group mean values of any other subject characteristic that could interact with the two treatments (examples: age, competitive level, proportion of males...), but you will have to re-jig the spreadsheet for such comparisons yourself.

When there *is* a substantial difference between the groups in the pre-test mean value of the dependent variable or of another subject characteristic, make sure you examine the plots of the change scores against pre-test value or create plots to examine the effect of the other subject characteristic on the change scores. If you see a substantial effect of pre-test value on the change scores, perform the analysis of variance or general linear model I described in the section on plots of change scores to deal with regression to the mean. That way you will have adjusted statistically for the differences in the groups in the pre-test. Your stats package should give you the experimental effect as the difference in the change for subjects who would be on the overall mean value of the covariate, because this value corresponds to the expected

value for the population represented by the full sample. You might need an expert to help you here.

You can also use this part of the spreadsheet for a comparison of means (and standard deviations) of any two independent groups without repeated measurements. Just ignore all the analyses related to changes in the means.

- **Estimates of uncertainty** expressed as confidence limits at any percent level for all effects, including the standard deviations representing individual responses.

[Confidence limits](#) define a range within which the true value (that is, the population value) of something is likely to occur, where *likely* is traditionally *with 95% certainty*. I now favor 90% confidence limits, because I believe 95% limits give an impression of too much uncertainty. They are also too easily reinterpreted in terms of statistical significance at the 5% level, which I am also keen to avoid.

The confidence limits for the treatment effects are generated using the same formulae as in my [spreadsheet for confidence limits](#). Confidence limits for the standard deviation representing individual responses are based on one of the methods used in Proc Mixed in SAS. The sampling distribution of the difference of the variances of the change scores is assumed to be a normal distribution. The variance of the sampling distribution of a variance is $2(\text{variance})^2/(\text{degrees of freedom})$. The variance of the difference in the variances is simply the sum of the two sampling variances, because the control and experimental groups are independent.

I have supplied confidence limits for the comparison of the two groups in the pre-test, but I don't think we should use them for controlled trials. After all, any difference between the two groups is real as far as imbalance in the assignment is concerned. What matters is how different the groups *were* for the study, not how different the groups *could be* if we drew a huge sample to get the difference between the population values for the control and experimental groups. But if you are using the spreadsheet only for a comparison of means and standard deviations of a single measurement of subjects in two groups, then of course confidence limits are appropriate.

- **Chances that the true value of an effect is important**, for all comparisons of means.

Important means *worthwhile* or *substantial* in some clinical, practical, or mechanistic sense. You provide a value for the effect that you consider is the smallest that would be important for your subjects. The spreadsheet then estimates the chances that the true value is greater than this smallest important value, and it also shows the chances in a qualitative form (*unlikely*, *possible*, *almost certain*, and so on).

Sometimes you have no clear idea of the smallest important value, especially for physiological or other mechanism variables. In such cases, use Cohen's smallest effect size of 0.2, which is included in the spreadsheet as the default. By inserting values of 0.6, 1.2, or 2.0, you can also estimate the chances that the true value is moderate, large, or very large. Try these different values when you want to say something positive about a smallish effect that has considerable uncertainty arising from large error of measurement, individual responses, or a small sample size. You can usually state something like *the mean effect could be trivial or small, but it is unlikely to be moderate and is almost certainly not large*. Such statements will help get your otherwise apparently inconclusive study into a journal, if the reviewers and editor are enlightened.

The chances are calculated using the same formulae as in the [spreadsheet for confidence limits](#). See my [recent article](#) for more (Hopkins, 2002b).

Crossovers

The example in the spreadsheet is for a control treatment and two experimental treatments, with equal numbers of subjects on each of the six possible treatment sequences. For a study with a single experimental treatment, ignore or delete the columns relating to the second experimental treatment. Add more columns to add more treatments. You can also use the spreadsheet to analyze a time series, in which all subjects receive the same treatment sequence. This design obviously isn't as good as a crossover, but it can still provide valuable information if you do multiple trials before, during and/or after the treatment.

The analyses are based on the paired t statistic. For convenience, the t statistic is generated by comparing a column of change scores with a column of zeros rather than by comparing directly the columns of scores for each treatment. This approach also allows you to include and analyze a column representing any effect derived from within-subject modeling, such as a slope representing a gradual change in a time series, or parameters describing a dose-response relationship when the treatments differ only in dose.

Plots of change scores in a crossover fulfill the same roles as in a controlled trial: to check for outliers and to check that your chosen transformation makes the effect reasonably uniform across the range of subjects' values for the control treatment (the equivalent of the subjects' values for the pre-test in a controlled trial). Beware that regression to the mean will sometimes be responsible partly or entirely for any trend towards experimental-control changes that get more positive for lower control scores, even in the plots with the most appropriate transformation. To see the effect of the control value on the experimental treatment free of this artifact, one of the treatments in the crossover needs to be a repeat of the control treatment. The plot of the experimental-control change scores for one control against the values for the other control shows the pure effect of the control value on the experimental treatment. Fit a line or a curve to the points to quantify the effect, or do an equivalent analysis with mixed modeling.

Lack of a control group in a crossover precludes estimation of reliability, but I have included estimates of the typical error derived from the changes between treatments. This error is still formally a measure of reliability. If you compare your values with those from reliability studies, take into account the possibility that individual responses to one or more of your treatments can inflate the error.

Another potential source of inflation of error in a crossover is a substantial familiarization effect, particularly between the first and second trials. This effect adds to the change scores of subjects who have the control treatment first, whereas it subtracts from those who have it second. If there are equal numbers of subjects in each treatment sequence, there is no net effect on the mean change score (the treatment effect), but the random error increases. Analysis with mixed modeling allows you to estimate the magnitude of the familiarization effect and to eliminate this source of error by including the order of the treatments as a within-subject nominal covariate in the fixed-effects model. You end up with a more precise estimate (narrower confidence interval) for the treatment effect. Mixed modeling also provides a check on the balancing of the assignment. Get your subjects to repeat the control treatment, preferably in a balanced manner, and it will also provide you with estimates of reliability and individual responses.

In conclusion, the spreadsheets should help you analyses your data appropriately. Admittedly, you can't include covariates in the analysis, but the spreadsheet should still help you do the right things when you use a more sophisticated statistical package.

Link to [reviewer's comment](#).

Download the spreadsheets: [controlled trial](#) and [crossover](#).
View the [Slideshow](#).

References

- Hopkins WG (2000a). Quantitative research design. *Sportscience* 4(1), sportsci.org/jour/0001/wghdesign.html (4318 words)
- Hopkins WG (2000b). Getting it wrong. In: *A New View of Statistics*. newstats.org/errors.html
- Hopkins WG (2000c). Measures of reliability in sports medicine and science. *Sports Medicine* 30, 1-15 ([download reprint](#))
- Hopkins WG (2002a). A scale of magnitudes for effect statistics. In: *A New View of Statistics*. newstats.org/effectmag.html
- Hopkins WG (2002b). Probabilities of clinical or practical significance. *Sportscience* 6, sportsci.org/jour/0201/wghprob.htm (638 words)
- Hopkins WG (2003a). Other repeated-measures models. In: *A New View of Statistics*. newstats.org/otherrems.html
- Hopkins WG (2003b). Models: important details. In: *A New View of Statistics*. newstats.org/modelsdetail.html
- Hopkins WG (2003c). Log transformation for better fits. In: *A New View of Statistics*. newstats.org/logtrans.html. (See also the pages dealing with other transformations.)
- Hopkins WG (2003d). Regression to the mean. In: *A New View of Statistics*. newstats.org/regmean.html

Published Oct 2003
[editor](#)
[©2003](#)