SPORTSCIENCE sportsci.org

Perspectives / Research Resources

Understanding Statistics by Using Spreadsheets to Generate and Analyze Samples

Will G Hopkins

Sportscience 11, 23-36, 2007 (sportsci.org/2007/wghstats.htm)

Institute of Sport and Recreation Research, AUT University, Auckland 0627, New Zealand. Email. Reviewers: Bernadette Murphy, Faculty of Health Sciences, University of Ontario Institute of Technology, North Oshawa, Ontario L1H 7L7; Patria Hume, Institute of Sport and Recreation Research, AUT University, Auckland 0627, New Zealand.

The random number and probability distribution functions in Excel allow the user to easily generate samples that simulate data typical of any kind of biomedical study. The act of generating the samples should provide the user with an implicit understanding of fundamental statistical concepts, including variables, probability, independence, sampling variation, linear modeling, random error, fixed effects, random effects, and individual responses. Analysis of the samples, which is essentially an attempt to recover the formulae that generated the samples, should reinforce these concepts and develop others related to statistical inference, including bias, confidence limits, statistical significance, and chances of benefit and harm. The spreadsheets accompanying this article provide examples of generation and analysis of data for reliability and validity studies and for simple and covariate-adjusted comparisons of group means without and with repeated measurement. An example is also given for generation of a binary variable for data simulating events, such as the occurrence of injuries, but the analysis by generalized linear modeling is currently not available in these spreadsheets. KEYWORDS: confidence limits, data analysis, probability, random number, research design, simulation

Reprint pdf · Reprint doc · Spreadsheets

When I began to use an advanced statistical package, I found that I could come to terms with the output by analyzing data with an obvious effect that I had made up. This exercise also gave me a deeper understanding of statistics. In this article I explain how you can go through a similar process with spreadsheets.

In the beginning I used my imagination to get numbers representing an effect, such as a correlation between weight and height or a difference in mean IQ between boys and girls. I soon realized that it would be better to use the stats package itself to generate a known correlation or a known difference between means. If the analysis reproduced the known effect, I could be more confident that I was using the package correctly.

What did I mean by a *known* effect? There is a true value for an effect in a population, but when you do a study, you get only a sample value, which is never exactly the true value. So when you generate a sample containing an effect, the trick is to reproduce the behavior of real samples: the *known* effect is the true or population value, and you make up a sample that gives something like the true value. If you were to make up many samples, every sample would give a different value scattered around the true value. On average, the sample value would be the true value. Or you could make up a very large sample, in which case the sample value should be very close to the true value.

Of course, in real life you can never know the true value of an effect in a population, so when you analyze real data, you can't check if you've got the right answer. But with made-up data, you can. And it's not simply a matter of making lots of samples or a very large sample. With any statistical analysis of a sample, you have to derive the *confidence interval* for the effect. Do this with your made-up sample and you should find that the confidence interval includes the right answer–except that there's a small chance it won't: 10%, if you choose 90% for the level of your confidence interval. If you are confused at this point, you won't be when you build up some simple data from scratch. All these ideas will emerge naturally.

Let's now see how to make samples with a spreadsheet. In the following sections I will work my way through the accompanying workbook, which has a spreadsheet for each section. At the end of each section I list the concepts for which you should have developed an implicit understanding. Within each spreadsheet I have filled cells with the colors of the rainbow to represent the sequence in which the cells were created: red first through to violet and sometimes to grey and white. (My apologies to the color blind.) You'll probably learn more by recreating each cell's formula in an adjacent blank cell or in a new blank spreadsheet rather than by simply clicking your way through the filled cells.

RAND() and Coin Tossing

At the heart of sampling is the notion of a *random* sample of a population. In Excel you access randomness via the function RAND(). Nothing goes in the brackets, by the way: all Excel functions have brackets, whether or not they have an argument. A similar function is PI(), which generates the value of π. If you can't remember a function's exact name, access it and all the other available functions in two ways: either click on the *f*x symbol towards the left-hand end of the lowest menu bar, or select Insert/Function… from the top menu bar. A window will open that you can navigate through to find the function of interest.

So, put or type =RAND() into a cell (in upper or lower case, followed by pressing Enter, which I won't bother to state again) and you will get a number between 0 and 1, such as 0.230592. That's only the first six decimal places of a number that has 15 decimal places, but we don't have to worry about all those extra digits. Each digit is chosen randomly from 0 through 9. The result is a random number between 0 and 1. Put =RAND() into another cell and you will get another similar random number. But notice that the first cell you put =RAND() into has now changed to yet another random number! Whenever you get into and out of any cell, even if you put nothing into it, Excel updates all the values of RAND everywhere in the spreadsheet. To make this updating happen rapidly (which is often useful), hit Ctrl-D ("copy down", so make sure the cursor is not sitting in the first row of the table or in a blank cell immediately underneath a non-blank cell).

A number that ranges between 0 and 1 can be interpreted as a *probability*: values closer to 0 represent more unlikely, and values closer to 1 represent more likely. You can interpret probability as the proportion of times you expect something to happen. So let's have a bit of fun using IF and RAND to simulate coin tosses. Type =IF(RAND()<0.5,"heads","tails") into a cell. Now click on that cell, then click on the little black box in the bottom right-hand corner and drag down for 10 or so cells. You have just simulated 10 or so tosses of a coin. Now do the same again, but change the 0.5 to 0.9 and see what happens when you toss the coin 10 times. To change the value in a lot of cells all at once, put =RAND() in one cell, say G5 as shown, then put =IF(RAND()<G5,"heads","tails") in another cell. To do it using the full functionality of Excel, type =IF(RAND()< then click on cell G5, then hit the "Function 4" (F4) key and $ signs will appear immediately. (Toggle the F4 key and you cycle through all four possible combinations of $ and no $--for advanced users! Mac users: sorry, F4 doesn't work for you.) The $ signs "freeze" the reference to a cell when you do a copy operation, as you will see when you complete this cell and copy it down for 10 or so cells, as above. Now change the contents of G5 to see what happens when you have a weird weighted coin that comes up heads or tails more often than a fair coin.

Concepts: random, probability.

Simple Sample Statistics

I will start with a *sample* that gives nothing more than some simple statistics–the *mean* and *standard deviation* (SD)–for a single continuous variable. Excel has a function, NORMSINV, that converts a probability into a value of a normally distributed variable drawn from a *population* with a mean value of 0 and an SD of 1. You don't have to understand what NORMSINV does; just give NORMSINV the probability generated by RAND, et voila, a value of a variable! The rest of this paragraph is for the clever or curious. NORM and S stand for the standard unit normal probability distribution, and INV stands for the inverse thereof. In fact, it is the inverse of the *cumulative* normal probability distribution: =NORMSINV(p) interprets p as a cumulative probability by returning a value of a normally distributed variable with mean of 0 and SD of 1, such that p is the proportion of the population with values less than the value it returns. (Draw a normal distribution and shade the area under the curve to the left of a value: that's p. Also, p here is not the p value of null-hypothesis tests.) Small values of p give large negative values of the unit normally distributed variable, p=0.5 gives the value zero, and values of p close to 1 give large positive values of the variable. p=0.025 and p=0.975 give -1.96 and 1.96, which you will recognize if you've done stats courses.

Let's generate that sample. Use the layout in the spreadsheet as a guide to label and make a column of RAND values and a column of NORMSINV of those values. (Having made one cell, copy it down by dragging down the little black box in the corner as before, or by highlighting the cell and blank cells below it, then hitting Ctrl-D.) Then make up a few names or copy my column of gender-neutral names to emphasize that we are about to make values of a subject characteristic like peak power.

Next, decide on a mean and SD (e.g., 400 and 50, as shown). The mean is a measure of where the middle of the population falls, and the SD is a measure of the scatter of the values about the middle. Generate an observed value of peak power from these and the NORMSINV value using this equation: peak power = mean + SD*(the NORMSINV value). Use the F4 key to freeze the cells for mean and SD, then copy the cell down to generate your sample of 100 or so. Get a feel for the way the RAND value makes the NORMSINV value, and how the NORMSINV value combines with the mean and SD to make the observed value. It's obvious that the mean is a measure of middle of the numbers, and the SD is a measure of their scatter about the middle.

Each value we have generated is called an *observation*, short for the observed value of a variable. Notice that we generated the observations with only three things: a mean, a standard deviation, and a random number. Although the observations for different subjects are related by the mean and SD, the random number makes it impossible to say anything about the specific value of any one observation, given the value of any other observation. The observations are therefore said to be *independent*. Most analyses are based on the assumption that the values of each subject are independent of the values of all other subjects. Repeated measurements on the same subjects in a data set are, of course, not independent: you get a similar value whenever you re-measure the same subject, as we will see in the next spreadsheet.

Display a frequency histogram of the values as shown on the spreadsheet. (See instructions in the spreadsheet for this unbelievably clunky aspect of Excel.) With bigger and bigger sample sizes, the frequency histogram looks more and more like a classic *normal distribution*. What is a normal distribution, anyway? Just another one of those incredible manifestations of the laws of nature, which here make most variables have the same expected shape for their *sampling distribution*: their distribution of values. What's also incredible is that someone worked out its equation. Of course, real variables do not have perfect normal distributions, but they don't have to for statistical analyses.

Now for something more advanced... Derive the sample mean, sample SD and sample size using the AVERAGE, STDEV and COUNT functions. Click the cursor in a blank cell somewhere, then hit Ctrl-D a few times and watch what happens to the sample values. What you are seeing is *sampling variation*. Notice how the sample values hover around the population value.

Finally, some really advanced statistics… Create a *confidence interval* (or confidence limits) for the mean and for the SD. A confidence interval represents an *inference* or conclusion about the population or huge-sample value of a statistic. When you publish a 90% (say) confidence interval for a statistic, you are saying to the world: "I did this study with a sample, so I can't be sure about the true or population value, but if I were to do the study with a huge sample, the value of the statistic would probably fall in here; by *probably* I mean *there's a 90% chance*."

The confidence interval is defined by the lower and upper confidence limits. Generate these for the mean using the TINV function and for the SD using the CHIINV function. These functions are based on the shape of the sampling distribution of values you get for the mean and the SD. The formula for the confidence limits for the mean is simple enough: mean ± t*SD/√(sample size), where t is the value of a t statistic (given by TINV) appropriate for the level of confidence and the sample size. The formula for the confidence limits for the SD is a bit more complicated, so get the formulae into your spreadsheet by copying my cells. Check that the formulae point to the right cells by double clicking and noting where the colored boxes fall. Drag the colored boxes to the right cells if they aren't right. Click on the cells below ±CL and ×/÷CL to see how I derived the way of showing confidence limits as a single plus-or-minus number for the uncertainty in the mean and a single times-or-divide number for the uncertainty in the SD.

We seldom use confidence limits for a simple sample mean, but confidence limits for some kinds of SD are becoming more common. Regardless, I've got you to do some confidence limits here to show you what happens when you draw repeated samples. Do your population values of the mean and SD fall between their confidence limits? Hit Ctrl-D repeatedly and check each time. Do it until you get a sense that, yes, maybe one time in 10 the true limits do not enclose the true value. That's what a 90% confidence interval means: there's a 90% chance that it encloses the true value.

Play with the data. Try a 50% confidence interval to get a quicker sense of whether it really is 50%. Make the sample a lot smaller or bigger as follows: click the cursor in the far left of a row (not the first row of the sample) to highlight it, drag down to highlight multiple rows, then either right-click Delete… or right-click Copy and right-click Insert Copied Cells. (Do this operation below the level of the figure, to avoid problems.) Notice the effect of a change in sample size on the sampling variation and on the width of the confidence interval, but notice also with repeated use of Ctrl-D that the confidence interval stays true to its level.

Concepts: independent, observation, normal distribution, sampling distribution, population and sample mean and standard deviation, sampling variation, confidence interval or limits, inference.

Reliability and Error of Measurement

Measurement error and the concept of reliability enter naturally at this point. All we do here is tweak the values of a variable to make them more like the values you would measure in reality. The tweaking consists of adding two terms: another RAND, to represent random measurement error, and a constant term to represent the kind of systematic error, offset or *bias* that occurs when there is a learning, familiarization, fatigue or other order effect with repeated measurement on the same subjects. The spreadsheet shows two trials (measurements) on a sample of subjects, and an additional eight trials on one subject. You can change the random and systematic measurement errors for each trial, but the random error is shown as the same for all trials, and the systematic error is shown as disappearing after the first two trials. There is no need to make values for Trials 3 to 10 for more than the first subject, but I have put $ signs in the right places for the first subject in case you do.

Notice how each subject's second measurement is similar to his or her first. If you were to lump all the observations for both trials together, the observations would no longer all be independent: given one observation, you know there is a similar value for another observation, the observation from the same subject. Repeated measurements on the same subjects produce *repeated measures* (here, the variables represent the values in each trial) and each set of repeated measurements makes a *cluster* of observations (here, the pairs of measurements, representing a cluster for each subject).

The spreadsheet demonstrates that random measurement error can be estimated simply by deriving the SD for a single subject's repeated measurements: that's the usual meaning of measurement error. If you thought there was a shift in the mean between the first two or three trials, you would leave those trials out; including them would increase the apparent random error, which you can show in the spreadsheet by comparing the SD for Trials 3 to 10 with that for Trials 1 to 10. Increase the change in the mean in Trial 1 to make the effect more obvious, and use repeated sampling with Ctrl-D.

Most often we don't have enough repeated measurements to estimate the measurement error for each subject, so we assume it is the same for every subject and derive it by dividing the SD of the change scores by √2. This approach to estimation of error also neatly gets around the problem of any systematic change in the mean contributing to the error: by using change scores you estimate the change in the mean and thereby stop it contributing to the estimate of random error.

The √2 comes about as follows: there is independent random error on both trials; when you subtract (or add) two variables with independent random errors, the random error of the result is obtained by adding the variances (squares of the SDs) and taking the square root; with equal errors, the SD of the change scores is therefore √2 times the error we are trying to estimate. Try using NORMSINV(RAND()) to verify that the SD of the sum of two independent random variables is the square root of the sum their variances.

I have included another measure of reliability, the *test-retest correlation* between two trials. The CORREL function provides the value as the usual (Pearson) correlation coefficient. The retest correlation has to be really high–in excess of 0.96–for a measure to be good for assessing individuals, as explained in an article/slideshow (Hopkins, 2004), An *intraclass correlation coefficient* can also be calculated for reliability studies, but for two trials the Pearson and intraclass give practically identical values. When there are three trials or more, the intraclass correlation coefficient is a kind of average of all possible pairs of correlations between trials. Generally you do not need the intraclass correlation, because you should analyze reliability for consecutive pairwise trials to address the problem of changes in the mean and in the error of measurement with repeated testing. If any averaging is to be done, it should be the average of consecutive pairwise estimates, which usually have the same time between trials.

The various reliability statistics and their confidence limits can be generated with a reliability spreadsheet available at my stats site. Practice using that spreadsheet by copying in the numbers generated here.

Finally and most importantly, this spreadsheet embodies the concepts of *linear modeling*. A model is an equation linking *predictor* variables to a *dependent* variable. In a linear model, the predictors are simply added together; indeed, *additive* modeling would be a better term. I have produced each subject's values in each trial using an additive model: a true value for each subject that consists of a constant value (the overall mean) PLUS a value that differs randomly between subjects (characterized by the between-subject SD) PLUS a constant that is the same for each subject within a trial but differs between trials (the systematic error) PLUS an error that differs randomly for every subject (characterized by the random error SD). In this model the overall mean and the constant within a trial are *fixed* effects (because they don't change between some or all subjects) and we estimate them as means or mean differences or changes; the terms that differ from subject to subject or even within a subject are *random* effects (because they change randomly) and we estimate them as SDs.

Statistical analysis is simply an attempt to recover part or all of the model that generated the data. Here we create the model, and we can check if the analysis is working properly. In reality Nature creates the model, and we don't know what it is. We have to assume a model, then try to recover it.

In the past, stats packages focused on the fixed-effects part of the model and used a procedure called *analysis of variance* to estimate these effects. Only one random effect was permissible, although strategies and patches were devised to allow for more. Advanced stats packages allow you to specify and estimate all sorts of random effects, and the approach is called *mixed modeling* to reflect the mix of fixed and random. The computational procedure, estimation by maximum likelihood, is not easy to set up in Excel. The analyses in this and my other spreadsheets are therefore based on ANOVA, but I use strategies to allow for different random effects where required.

Concepts: random error, systematic error, repeated measures, clusters, test-retest correlation, linear model, predictor and dependent variables, fixed effect, random effect, mixed model.

Log Transformation

In the previous section I developed a linear model by adding various terms to make the value of a dependent variable. I then used an analysis that attempted to recover the terms from a sample. You might be surprised to learn that almost all statistical analyses are based on recovering terms from linear models. Are most effects in the real world additive? No, so we have to find ways to make real effects into additive effects if we want to use linear models.

A large number of effects in biomedical science–I would say the majority–are multiplicative; that is, you should think about what's going on in terms of *percent* or *factor* differences or changes. For example, a treatment increases output power by 5% or a factor of 1.05, regardless of the individual's initial power output. Enter the *logarithmic transformation*. Logs turn multiplication into addition, so multiplicative effects become additive effects that can be analyzed with all the procedures developed for linear models. When you've done the analysis, you *back-transform* the outcome, either into a percent or a factor.

The log of a number is the power to which you raise a number called the *base*, which explains why multiplication becomes addition when you take logs. You can use any base for log transformation. Excel has logs for two bases: the LOG function uses base 10 (LOG(10) = 1, LOG(100) = 2,…), while LN uses exponential e and is known as the natural log. Use LN, because the back transformation is available as the function EXP (e to the power of…). If an effect or SD is y after LN transformation, the back transformation to a factor is EXP(y) and to a percent it is 100*(EXP(y)-1). I use 100*LN for the transformation, because the values of effects and SD in transformed units are already approximately percents, which sometimes makes it easy to see what's going on without having to back transform. The approximation is practically perfect when effects are <5%. The back-transformations for 100*LN are EXP(y/100) and 100*(EXP(y/100)-1).

The spreadsheet demonstrates the use of logs with simple numbers, then builds up some data for the effect of a treatment that has the same multiplicative effect for a few subjects. You will see that you get the same answer for the mean effect in the sample, whether you use the raw numbers or their logs (LN or 100*LN), but there is a big difference for the SD of the effect: the zero for the log-transformed variable reflects the fact that the effect is set up as a constant factor. We say that the effect is *uniform* across subjects after log transformation or that it is a uniform factor effect in the original data.

For this idealized example I did not include any error of measurement. When effects are multiplicative, the errors of measurement are also usually uniformly multiplicative; that is, you should think about them as constant percent or factor errors. Log transformation converts such errors into uniform additive errors. Analyses are not trustworthy when the errors are not uniform, so log transformation is important.

Concepts: percent or factor effect, log transformation, back transformation, uniform effect, uniform error.

Percent Error of Measurement

The next spreadsheet generates data for another reliability study, this time with a variable that requires log transformation for linear modeling. You therefore assume that the changes in the mean and the errors of measurement happen as percent or factor effects. For variables that behave in this manner, the differences between subjects are also usually best expressed as a percent or *factor SD*.

A percent SD is also known as a *coefficient of variation*, or CV, and it is usually defined as the SD of the observed values expressed as a percent of the mean. I have converted the between-subject CV to a factor SD (1+CV/100), to emphasize that percent variations and effects are really factor variations and effects. I often use "×/÷" in front of a factor SD, in the same way the we use "±" in front of the usual raw SD. Thus, 400 ×/÷ 1.3 indicates that the values range typically from 400 ÷ 1.3 to 400 × 1.3. The values often fall outside these limits, which are "typical".

Should you report variation as a ± CV or as a ×/÷ factor SD? There is no difference between +30% and ×1.3, but if you subtract 30% off a number you get something different from ÷1.3. The difference is negligible when the percent is <10, but -100% implies the resulting number is 0 (which is nonsense), whereas ÷2 implies the resulting number is half the mean (which is correct). For this reason I advocate showing CV only for CV of up to 20% or so, and I always keep in mind that what is implied by ±CV is really ×/÷(1+CV/100). CV are definitely preferable to factors when the CV are small: you can understand ±5.3% and ±0.86%, but it's hard to get a sense of magnitude from the corresponding factor SD of ×/÷1.053 and ×/÷1.0086.

Now let's generate a sample, which turns out to be a bit complicated. First, get true values by logging the mean and factor SD and combining with a random value of a unit-normally distributed number given by NORMSINV(RAND()), then back-transform. Next generate observed values by logging the true values and the random and systematic errors expressed as factors, combining with another random value given by NORMSINV(RAND()), then back-transform. Now you can log these values, perform the analysis exactly as before, then back-transform.

I have called your attention in the spreadsheet to a discrepancy between the means for each trial calculated directly from the raw data vs the back-transformed means of the log-transformed data. The back-transformed mean gives an unbiased estimate of the population mean we started with, whereas the raw mean consistently overestimates the population mean. Similarly, the CV between subjects calculated from the logs is an unbiased estimate of the CV we used to generate the data, whereas the CV calculated by dividing the raw mean by the raw SD is biased low. The bias isn't noticeable when the between-subject CV is <20% or so, but it is easy to see with a CV of 100% or more. Verify these assertions by doing repeated sampling with Ctrl-D.

You can make the differences more obvious by increasing the sample size, which will reduce the sampling variation. Here's a trick for increasing the sample size: click and drag in the far left of the spreadsheet to highlight the entire rows from the second subject (Ariel) down to the last subject (Kade); copy, then right-click and Insert Copied Cells. Notice that the sample size increases to 39. Repeat this operation as often as you like with as many rows as you like, but never highlight the first row nor highlight beyond the last row.

The estimates derived from the logs are obviously correct, but there's a sense in which the raw mean and SD are also correct–these are, after all, the actual mean and the actual SD of the set of numbers. Even so, the log-derived mean and CV give a better idea of the way in which the original numbers are distributed. When a variable requires log transformation, the simple statistics summarizing the variable should be the back-transformed mean and the back-transformed CV or factor SD.

I've been focusing on the mean and between-subject SD, but the main point of this exercise was to derive the error of measurement as a CV or factor SD. Use Ctrl-D to verify that the sample values of the changes in the mean and the error of measurement hover around the population values in an apparently unbiased way. I haven't bothered with calculations for changes in mean and errors derived directly from the raw numbers, because it's not appropriate to do them that way with these data. However, with real data you have to decide whether to use raw or log-transformed values. You make the decision by viewing the scatter of change scores between two trials plotted on the Y axis against mean scores of the two trials for each subject. The SD of the scatter in the vertical direction is √2 times the error of measurement, so if the scatter is reasonably uniform across the range of mean values, the error is reasonably uniform. When a variable needs log transformation for the reliability analysis, the scatter of change scores of the raw values gets bigger for bigger mean values, but the scatter of change scores of the log values looks reasonably uniform.

Non-uniform scatter, or *heteroscedasticity*, is usually a problem only when the between-subject CV is ~20% or more. Anything less and you can't see any difference in the scatter of the raw or log change scores, and you get little difference in the estimate of the CV via log transformation or via the raw data. You can explore these assertions by generating data and putting them into the spreadsheet for analysis of reliability, where all calculations and plots are generated automatically.

Concepts: percent variation, coefficient of variation, factor SD, ±, ×/÷, heteroscedasticity.

Validity and Error of the Estimate

The next two spreadsheets exemplify another measurement concept, *validity*. In a validity study you establish the relationship between the values of a practical or new measure and the concurrent true or criterion values. The relationship is a *calibration equation* for adjusting or converting the practical measure into the criterion. The analysis also provides an estimate of the remaining or residual random error, known as the *standard error of the estimate*: that is, the error in the estimate of the criterion value from a practical value.

The relationship between the criterion and practical measures is almost always investigated with a simple linear model, a straight line. As with reliability, some measures have percent or factor relationships and errors and so need log transformation before modeling. The spreadsheets show an example of each.

To generate the data, I started with a sample of subjects having values of a practical measure drawn from a population of normally distributed values with a chosen mean and SD. I then generated the criterion values using the equation Criterion= *a* + *b**Practical + *error*, where *a* is an intercept, *b* is a slope, and *error* is a random number drawn from a normal distribution with mean zero and SD equal to the standard error of the estimate. This equation gives the impression that the error is all in the criterion variable, whereas most of the error could be coming from the practical. Nevertheless, the use of an equation in this form results in an unbiased estimate of the criterion from the practical, and the resulting standard error of the estimate is also an unbiased estimate of the random error in the prediction. Failure to understand this issue has resulted in generations of researchers thinking that other kinds of modeling are required when there is error in a predictor variable.

Once you have specified the means and SD of the two variables and the standard error of the estimate, you can calculate directly the correlation between the variables. The formula for the population correlation is a bit complex, but it amounts to the square root of the fraction of variance in the Y variable explained by the X variable. The sample correlation has the same underlying formula, but you can get the value using the CORREL function. It is also possible to generate data for a straight line by specifying the population correlation rather than the population standard error of the estimate, as explained in a later section.

The spreadsheets include graphs or scatter plots of the two variables, with the best-fitting straight lines. Lines of identity on the graphs, and estimates of bias in the practical measures, make sense only when the practical is in the same units as the criterion. You can develop a calibration equation for dissimilar measures, in which case the correlation coefficient can be a useful way to assess the practical measure. A correlation in excess of 0.98 represents the start of really good measures for assessing individuals, also explained in that same slideshow (Hopkins, 2004), but measures with lower validity correlations are still useful for selection of groups of people or for providing additional evidence for making decisions about an individual patient or client.

The spreadsheet for validity with variables having a percent or factor relationship shows graphs for the raw data, for the raw data with log scales, and for the log-transformed data. Having drawn the graph for the raw data, the best way to draw the others is to click-Ctrl-drag on it to make an identical copy, then click on the points on the copy; the data will be outlined in colored boxes, and you can simply drag the boxes to the new data, if required. The straight line in the raw graph becomes a curve when you change the axes to log scales, so double click on it to get the Format Trendline window, click on the Type tab and change the type to Power.

These graphs provide a beautiful illustration of non-uniformity of error with raw data that becomes uniform with log transformation or equivalently, log axes. Play with the values of the between-subject SD and the standard error of the estimate. You will find that non-uniformity becomes noticeable only when the between-subject SD and standard error of the estimate are greater than about 20%. (Non-uniformity becomes less noticeable the smaller the sample size.) To see non-uniformity with small values of the standard error of the estimate you need to view a plot of *residuals vs predicteds*, the last step with this spreadsheet. The residual is the difference between the observed value of the dependent variable (here the criterion) and the value of the dependent value predicted by the observed value. As such, the residuals represent the scatter about the line in the vertical direction, and the standard deviation of the residuals is actually the standard error of the estimate. Uniformity of the residuals is probably the most important assumptions in analysis of linear models, so inspection of a plot of residuals vs predicteds is important. Copy the raw data into the validity spreadsheet at newstats.org for analyses with confidence limits and plots of residuals vs predicteds.

Concepts: validity, calibration equation, standard error of the estimate, residuals vs predicteds.

Comparison of Means in Two Groups

One of the most common statistical analyses is a comparison of the means of two groups, such as the mean of some variable for girls vs boys. The groups are usually independent; that is, knowing the value of any individual in one group gives you no clue about the value of an individual in the other group. Lack of independence could occur if there were clusters in the sample, such as boys and girls drawn from school classes, and your analysis would have to reflect that. Here we will generate data for a simple comparison of means of independent groups and do a simple analysis for *statistical significance* of the difference using the so-called *p value*. I'd prefer *not* to teach you about the p value and statistical significance, but it might be another 50 years before researchers are finally weaned off this confusing approach to making inferences about true values, so here goes…

The *p* in *p value* stands for the probability of getting any value of the effect statistic more extreme than your observed value (whether positive or negative), if in reality there was no effect. A small p value therefore implies that it is unlikely there is no effect, although p does not actually represent the probability of no effect. (The probability of exactly no effect is actually zero, which is one of several fundamental problems with use of statistical significance.) The convention is to treat $p<0.05$ as a threshold for deciding that there is an effect, and the effect is then said to be *statistically significant* at the .05 or 5% level.

Calculate the p value for the comparison of the means in the spreadsheet using the TTEST function. A *t test* is a procedure for generating a p value based on the sampling distribution of the difference between two sample means. The sampling distribution is a t distribution, which is like a normal distribution, except that it isn't quite a normal distribution, because you have to use the sample standard deviation to define its spread, and the additional uncertainty flattens the distribution a little. Follow the prompts to select the two arrays of numbers, then always choose "2" for the 2-tailed option (because the p is for more extreme positive or negative values) and "3" for unequal variances (because the scatter of values in the two groups could be substantially different, and the sample size isn't usually big enough to decide for sure, and even if it were you should still assume unequal variances). If the resulting p value is really small, it will display in decimal exponential notation, for example 1.64E-06 (which means 0.00000164).

To understand the p value better, make the population mean values equal. Now hit Ctrl-D repeatedly and watch the p value. On average, one time in 10 it will be less than 0.10, and one time in 20 it will be less than 0.05. It follows that, if there was really no difference in the means, one time in 10 or 20 you would nevertheless declare there was a statistically significant difference at the 10% or 5% level. But what can you say about the real difference? That's where statistical significance is deficient and where confidence intervals are so much better.

At this point you can get some practice using a spreadsheet I devised to calculate confidence intervals from a p value. See an accompanying article in this issue of Sportscience (Hopkins, 2007a). The spreadsheet also generates *magnitude-based inferences*, which are inferences based on the uncertainty in the magnitude of true value. An effect is unclear if there is a good chance that the true value could be substantial in a positive and negative sense, such as beneficial and harmful; otherwise the effect is clear, and if there is a big enough chance of benefit and small enough chance of harm, you would use it. To calculate the probability that the true value is substantial in a positive or negative sense, you will have to enter the *smallest important value* of the statistic (also known as the smallest worthwhile effect or the least clinically important effect). Only then can the spreadsheet calculate the probability that the true value is substantial in a positive or negative sense.

The value for the smallest worthwhile effect can come from your clinical or practical experience, or you can use defaults that others have worked out. The usual default for most purposes is based on the typical variation between subjects, the standard deviation. Why is the SD involved in a consideration of the magnitude of the difference between means? To understand, draw frequency histograms for the two samples in the spreadsheet. You will see that the histogram for the males in shifted right somewhat relative to that for the females, but there is a lot of overlap. If the histogram was shifted so far that there was only a tiny overlap, you would judge the difference to be large. But the overlap depends on the SD: the only way to get a tiny overlap is to have a shift that is much larger than the SD; conversely, a shift that is only a small fraction of the SD represents almost complete overlap–a trivial difference.

So, work out the difference in means as a multiple or fraction of an SD by dividing it by the SD. The result is a *standardized* difference. Jacob Cohen, a psychologist, suggested a value of 0.20 as the threshold separating trivial from small effects, and he proposed 0.50 and 0.80 as thresholds for moderate and large (Cohen, 1988). I agree with 0.20, but I think thresholds for moderate and large should be 0.60 and 1.20. I also suggest thresholds of 2.0 and 4.0 for very large and extremely large differences in means. The corresponding thresholds for correlations are 0.1, 0.3, 0.5, 0.7 and 0.9. See the page on magnitudes at my stats site for more.

The standard deviations of each group are never exactly the same either in the sample or in reality. Depending on your perspective, you could standardize with the SD of either group; for example, use the females' SD if you are interested in how different the males are in female terms. If subjects were randomized to the two groups for a post-only parallel-groups controlled trial, use the SD of the control group. Otherwise it's appropriate to average them, but it's not a simple matter of adding and dividing by 2. Standard deviations usually have to be turned into variances by squaring them before you process them in any way, as we saw earlier. Here you average the squares of the SD and take the square root, as shown in the spreadsheet. What is *not* appropriate for a simple two-group comparison is to use the SD of all the observations lumped together.

Concepts: statistical significance, p value, t test, magnitude-based inferences, smallest important value, standardized difference.

Means in Two Groups Plus a Predictor

It's one thing to compare the means of a dependent variable in two groups, but suppose the subjects on average differ between the groups in some other way that is related to the dependent variable. *Some other way* is represented by values of another variable called a *predictor* or *covariate*, the latter name referring to the fact that it correlates or covaries with the dependent. In research, we often want to *adjust* or *control for* a predictor or covariate by asking the question: what is the difference in the dependent variable between groups for subjects who have the same value of the predictor? This difference is not contaminated or *confounded* by the predictor.

For an example, the spreadsheet generates values of a predictor, maximum oxygen uptake, and correlated values of a dependent variable, peak power, in a group of females and a group of males. You choose means and SD for both variables and the correlation between them in each group. You generate the individual values of the predictor first, then from them you generate the values of the dependent using this equation: (Y-Ymean)/SDy = r*(X-Xmean)/SDx + *error**$\sqrt{(1-r^2)}$, where Y and X are the dependent and predictor, r is the correlation, and *error* is drawn from a normal distribution with mean = 0 and SD = 1, using NORMSINV(RAND()). This equation also gives an important interpretation of a correlation: if you change the X or predictor variable by one SD for X, the average change in the Y or dependent variable is r times the SD for Y. It would also be easy to generate the Y from the X by specifying a slope (change in Y per change in X) and a residual error (standard error of the estimate) for Y.

We are interested in the difference in mean peak power between the females and males when differences in maximum oxygen uptake are taken into account. You won't understand what's happening until you draw the figure showing the relationship between the predictor and dependent in each group. You should also draw a dashed line in the figure to illustrate the mean difference between the groups at a chosen value of the predictor. The usual chosen value is the overall mean of the predictor in both groups combined, but you can choose any value.

You could easily calculate the predicted (adjusted) value of peak power in each group at the chosen value of maximum oxygen uptake using the FORECAST function. Compare the adjusted difference with the raw difference. Play with the population mean values until you get an adjusted difference that is approximately zero, whereupon you will see that the lines through the points in each group run together. It all makes sense, really.

The confidence interval for the difference between the groups would take too long to develop on this spreadsheet. I have therefore provided a new spreadsheet for this and other purposes with data like these. See the In-brief item in this issue (Hopkins, 2007b) for more. Copy and paste a set of raw values into the spreadsheet then work your way through it, paying special attention to the inferences.

There is an interesting question about which SD you use to gauge magnitudes when you adjust for a covariate. What you are asking is this: how different are females and males who have the same value of the covariate? You therefore use the SD for subjects with a given value of the covariate. Here that SD is the standard error of the estimate, which is the scatter about the line at any given value of the covariate. In the new spreadsheet I have averaged the standard errors of the estimate from each group, by averaging their variances and taking the square root.

It's possible to adjust for two or more covariates simultaneously, using *multiple linear regression* via the function LINEST. Instead of fitting a simple linear model of the form Y = *a* + *b**X + *error*, you fit the more general Y = *a* + *b**X_1 + *c**X_2 +… + *error*. In the analysis you recover the *a, b, c,…* and derive the effects therefrom. Unfortunately LINEST has some illogical features that make it difficult to use, so I won't go any further with it here, but download a no-frills spreadsheet I have devised for instructions, examples and additional information about adjusting bias in various correlations.

The model for multiple linear regression is, of course, just another linear model, but it has a generic look about it. An equation of this form is therefore known as a *general linear model*. The Y has to be a *continuous* variable, such as time in seconds. The predictors (the Xs) can be continuous variables, *ordinal* variables that have integers as values (0, 1, 2,…), or *nominal* variables such as sex, with levels (female, male) rather than numeric values. In stats packages nominal predictor variables are coded into numeric variables having values 0 and 1 for each level (e.g., female=0, male=1) before analysis.

The *a, b, c,…* in the general linear model are known as *coefficients* or *parameters*. An analysis based on recovering values of parameters is therefore known as a *parametric analysis*. Anything else is a *non-parametric analysis*, and in principle you should use such analyses when you can't be sure about the form of the model or whether the assumptions underlying the analysis are justified. In practice you can't be sure about anything, but most data can be coaxed to give trustworthy outcomes with parametric analyses. Non-parametric analyses suffer from loss of precision with small sample sizes, they do not provide estimates of magnitude, and they do not allow you to investigate underlying relationships represented by a model. I therefore strongly advise against their use.

The spreadsheet for the analysis of two groups shows analysis not only for the raw data but also for the data after three kinds of transformation: log, rank and root. *Log transformation*, as already discussed, is for factor effects. *Root transformation* is used for variables representing counts or scores proportional to counts. *Rank transformation* is a last resort for variables that don't analyze properly either in the raw state or after log or root transformation. But what does "properly" mean? Basically the model has to fit the values of the dependent variable uniformly over the range of values of the dependent, because that assumption underlies the linear model. So you need to check for uniformity in some kind of plot related to the individual values and the values generated by the model: either the residuals vs predicteds plot, or for simple models as here, the more intuitive plot of the dependent and predictor, with the modeled straight lines. The scatter of individual values around the lines (or the scatter of residuals for different predicted values) has to be more-or-less uniform: no systematic deviation and no difference in the degree of scatter over the range of values. (The analysis here allows for a difference in the degree of scatter in the two groups, which the usual ANOVA doesn't.) So you should always check the scatter to help make a decision about transformation.

Analysis after rank transformation is sometimes referred to mistakenly as non-parametric and is sometimes regarded as a panacea for unusual data. But it is definitely parametric, and the implication of the model is that differences in rank are linearly related to predictors. It follows that an effect that is uniform with raw data or other transformation cannot be uniform after rank transformation, because you need different effects to get the same change in rank for subjects with different values of the variable. Rank transformation is therefore not a cure-all: you have to check on its suitability.

Concepts: adjusting for a covariate, confounding, multiple linear regression, general linear model, continuous, ordinal, nominal, coefficient, parameter, parametric analysis, non-parametric analysis, rank transformation, root transformation.

Pre-post Controlled Trial

In the previous spreadsheet we compared the means of single observed values in two groups, females and males. The groups can, of course, differ in ways other than the sex of the subjects; in particular, if one group consists of subjects treated in some usual or control manner while the other group consists of subjects treated in a different experimental manner, the data would represent a *controlled trial*. It is unusual to have a controlled trial in which the subjects are measured only once, but there are situations where it is the best design for studying an intervention, as described in the article on controlled trials at this site (Batterham and Hopkins, 2005). Far more common is the kind of controlled trial where subjects are measured before and after the control and experimental treatments. I call it a pre-post parallel-groups controlled trial. What you then compare between the groups is the *change* in the dependent variable. The next spreadsheet generates data for such a study.

The spreadsheet allows for changes in the mean with each trial, a different error of measurement with each trial, and extra variation representing *individual responses* to the treatment. All the terms are added in classic linear-model fashion to give the observed value for a given subject in a given trial. The term representing individual responses should be particularly helpful for understanding what it means to have individual responses to a treatment. It is simply a random normally distributed number with a mean of zero and a given standard deviation. Each subject in the experimental group gets a new value for each post-test. Thus a subject who draws a positive value has a bigger response than the mean or fixed effect of the treatment in that trial, while a subject with a negative value has a smaller response. Depending on the magnitudes of the mean effect and the standard deviation representing individual responses, a substantial proportion of subjects could have a negative response to the treatment. Whatever, one aim of the analysis is to estimate the standard deviation representing individual responses, and I hope it is clear that the estimate should be free of the random error of measurement common to every trial.

I have also built in a covariate to simulate a subject characteristic that correlates with the extra variation representing individual responses. The covariate thereby accounts partly for the individual responses, depending on the strength of the correlation. This effect of a covariate on an outcome is important, because with real data it can help you identify positive responders, non-responders and negative responders to a treatment.

The spreadsheet generating all these data is now getting too complex for you to reproduce in less than a few hours. It is probably more sensible for you to examine the formulae in some of the cells to see how the data are generated, then to paste the data into another spreadsheet for detailed analysis. The spreadsheets for analysis of controlled trials are available in an article published here last year (Hopkins, 2006). See how well it estimates the fixed effects, the individual responses, the covariate, and the error term.

Concepts: controlled trial, individual responses.

Binary Outcome

The last spreadsheet generates observations for a variable that has only two possible values: 1 or 0, representing the occurrence or non-occurrence of an event (heads or tails, present or absent, injured or injured, selected or rejected, and so on). As with the very first spreadsheet, we have a probability that something will happen (e.g., tossing a weird coin that has a 0.30 chance of heads), and we draw a value of probability using RAND; if RAND is <0.3, we score 1; otherwise we score 0.

The difference from the first spreadsheet is in the way we model the probability that something will happen. We generate events in two groups, so obviously we want different probabilities of the event in each group, but we make it more interesting and realistic by introducing a covariate representing a subject characteristic: subjects with different values of the covariate have different chances of occurrence of the event. How do we develop a linear model for a dependent variable that ranges between 0 and 1 (the probability of the event)? Not easily! We can't use p = a + b*X, because X can in principle range from minus infinity to plus infinity, so no matter what the values for a and b, we can easily end up with values of p outside the 0-1 range, which are impossible. To do linear modeling we have to use a + b*X, so we need a transformation of p that ranges from minus infinity to plus infinity. Use of *odds* gets us halfway there. The odds of something happening is the probability it will happen (p) divided by the probability it won't happen (1-p): odds = p/(1-p). Because p ranges from 0 to 1, odds must range from 0 to infinity. Take logs and the result ranges from log(0) (which is minus infinity) to log(infinity) (which is plus infinity). The required transformation is therefore log of the odds: ln(p/(1-p)), also known as the *logit transformation*. We generate the p for each observation using ln(odds) = a + b*X = ln(p/(1-p)), so p/(1-p) = exp(a+b*X), and therefore p = exp(a+b*X)/(1+exp(a+b*X)). This explanation will lose most of you. I've done it in two stages in the spreadsheet, to make it simpler to follow.

It seems incredible that so much of our modeling of events is based on a mathematical convenience. Does nature really model the effect of subject characteristics on the log of the odds of an event in a linear fashion? I doubt it, but apparently the approximation to reality is good enough in many cases.

So much for generating the data with a linear model. Now we need an analysis that will recover the linear model so that we can adjust for the effect of the covariate in our estimation of the difference in risk of an event in each group. The analysis is called *binomial regression* or *logistic regression*. *Binomial* refers to the sampling distribution of values of a binary variable, and *logistic* refers to the transformation required. Unfortunately it's not easy to set up such analyses in Excel. You can buy add-ins for the purpose, and the US Defense Technical Information Center has provided one free (see article for instructions and link), but I have not evaluated any as yet. Meantime you will have to use more sophisticated stats packages such as SAS and SPSS.

The default outcome statistics from logistic regression are not easy to understand. Recall that the outcome when we compared the means of two groups is simply the difference in the means. With logistic regression we model the log of the odds, so the difference between two groups is expressed as the difference in the log of the odds. As we have learned. a difference in logs is a log of a ratio, here the log of the *odds ratio*. We can back-transform this outcome simply to an odds ratio, and that is the way the effects are usually reported. The effect of the covariate is also an odds ratio, per unit of the covariate. (I usually evaluate the effect of a covariate per two of its between-subject SD, because the resulting effect on a continuous dependent variable can be evaluated as the magnitude of the effect of the covariate by referring to the standardized magnitude thresholds for the dependent.)

Odds ratios are unsatisfactory outcome measures, because they can be hard to interpret. When the probabilities of the event in each group are small (<0.10, or chances of <10%), the odds ratio is approximately equal to the *risk ratio*: the ratio of the probabilities or risks in the two groups. A risk ratio of, say, 2 always means twice as likely, but an odds ratio of 2 means a risk ratio of less than 2 when one or both risks are >10%. But even risk ratios can be hard to interpret when it comes to evaluating risk in real terms, so it is always a good idea to calculate the adjusted risks in each group from the modeled odds, with confidence limits. The effect of a covariate should also be shown as actual risks for typical values of the covariate for subjects in each group. Unfortunately I can't demonstrate these assertions via an analysis in the spreadsheet.

In this introduction to event-type outcomes I have omitted modeling of a dependent variable representing a count of events, such as players' or teams' numbers of injuries. One approach is to apply the usual linear model to the root transform of the count, which is included in most of my analysis spreadsheets. For counts greater than 10 or so, the root transform has an error independent of the count (the error is actually

0.5), so in principle there is no problem with non-uniformity of error. But Nature does not necessarily produce linear effects on the square root of counts, root transformation doesn't work for low counts, and back-transforming root effects is awkward. An approach that works with any counts is called *Poisson regression*, named after the sampling distribution of counts of random events. Poisson regression works via linear modeling of the log transformation of the count, so after back transformation the effects are ratios of counts or *rate ratios*. Whether or not Nature models rates in this way, it is easy to understand effects that can be expressed using terms like "twice the rate". Poisson and binomial regression are members of the family of *generalized linear models*, which includes the usual linear models. For more on modeling of events download an article (Hopkins et al., 2007) from the Clinical Journal of Sport Medicine.

Concepts: odds, logit transformation, binomial or logistic regression, odds ratio, risk ratio, rate ratio, Poisson regression, generalized linear modeling.

Other Designs

You now have the expertise to create data simulating any study you are likely to undertake in the biomedical disciplines. If the design is complex or has a complex outcome statistic, create some data with a spreadsheet and analyze them, either with another spreadsheet or with a stats package. In the process you will learn more about the effect you are studying, the confidence limits will help you decide on the required sample size, and you will feel more confident that you have mastered the analysis.

References

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

Cohen J (1988). Statistical power analysis for the behavioral sciences, 2nd Ed. Hillsdale, NJ: Lawrence Erlbaum

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

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

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

Hopkins WG (2007b). A spreadsheet to compare means in two groups. Sportscience 11, 22-23

Hopkins WG, Marshall SW, Quarrie KL, Hume PA (2007). Risk factors and risk statistics for sports injuries. Clinical Journal of Sport Medicine 17, 208-210

Published Dec 2007